UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES & ENGINEERING

School of Electronics & Computer Science

Volume 1 of 1

The DNA of Web Observatories

by

Ian C. Brown

Thesis for the degree of Doctor of Philosophy

March 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES & ENGINEERING

Web Science

Thesis for the degree of Doctor of Philosophy

THE DNA OF WEB OBSERVATORIES

Ian Christopher Brown

This thesis investigates the proposed Web Observatory (WO) which will offer access to globally shared data and apps, delivering insights into the nature of the Web and also society-on-the-Web. Understanding how different groups conceptualise and engage with WO concepts is vital to understanding the drivers for adoption and the requirements for adoption between groups.

Observations from the field and analysis of work relating to WOs are combined and compared with established theories of innovation and adoption. I argue that a purely technological definition of WO is necessary-but-not-sufficient to capture the set of complex interactions and interests that a network of Observatories at Web scale would need to reflect.

A new socio-technical 'DNA' model of Web Observatories is developed combining technical and architectural definitions (D factors) with socially-embedded narratives (N factors) and group perspectives and motivations (A factors). Visual model of D's, N's and A's and a new perspective on parallel modelling for technically- vs socially constructed models is introduced.

An inductive approach, which combines case studies, content analysis and extensive interviews/ observations blends data from a broad range of sources across academia, business and government. A new WO taxonomy is established and iterative analysis refines a multi-perspective model of WOs employing a constructivist grounded theory (CGT) lens. An approach combining Interpretative Phenomenological Analysis (IPA) and visual mapping techniques using a hybrid concept mapping/TRIZ approach is developed to model the findings.

Social theories are considered around individual/shared meanings to enable a definition of WO to be embedded (framed) within the social context of the individuals and groups who seek to use WO to address specific problems and outcomes.

This thesis has implications for how new Observatories may be designed and built and also for how existing systems and sources may be recruited into a global Observatory eco-system through a better understanding not only of *how* participants may join but also, critically, *why* they would choose to do so. The models/techniques developed here may find a wider application for the study of socio-technical systems and social machines.

Table of Contents

List of Tables

List of Figures

Copyright Permissions

Figure 1-1 reproduced with kind permission of BlackRock Inc. date 14/03/17. All rights reserved Blackrock Inc.

Figure 3-1 reproduced with kind permission of B. Whitworth date 14/03/17. All rights reserved B.Whitworth.

Figure 3-3/Appendix A reproduced with kind permission of Southbeach date 15/03/17. All rights reserved Southbeach Inc.

List of Accompanying Materials

None.

DECLARATION OF AUTHORSHIP

I, **IAN CHRISTOPHER BROWN** .....[please print name]

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

THE DNA OF WEB OBSERVATORIES .....

.....

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. [Delete as appropriate] None of this work has been published before submission [or] Parts of this work have been published as: [please list references below]:

Signed: .....

Date: .....

Related Publications:

1. Brown, IC., Hall, W. and Harris, L., 2013. “From Search to Observation” in Proceedings of the 22nd International Conference on World Wide Web Companion 2013, pp. 1317–1320. Rio de Janeiro, Brazil. May 2013.
2. Brown, IC., Hall, W. and Harris, L., 2014. “Towards a taxonomy for Web Observatories”. At WOW2014 Web Observatory Workshop, Seoul, S. Korea. Apr 2014.

3. Brown, IC., Hall, W. and Harris, L., 2015. “DNA: Towards a method for analysing Social Machines & Web Observatories”. At WOW2015 Web Observatory Workshop, Oxford, UK. Jun 2015.
4. Brown, IC., Hall, W. and Harris, L., 2015. “DNA: From Search to Observation revisited” in WebSci15: Proceedings of the ACM Web Science conference. Oxford, UK. Jun 2015
5. Brown, IC., Harris, L. and Hall, W., 2013. “Enabling Web 3.0 adoption in the digital economy” at Digital Economy (DE2013) Conference, Salford, UK. Nov 2013.
6. O’Hara, K. and Brown, IC., 2014. “Responsible use of data” in HC245 written submissions to Commons Committee on Science & Technology. The Stationary Office, Nov 2014.
7. Wang, X., Tinati, R., Mayer, W., Rowland-Campbell, A., Tiropanis, T., Brown, IC., Hall, W., O'Hara, K., Stumptner, M., and Koronios, A., 2015. “Building a Web Observatory for the South Australian Government: Supporting an Age Friendly Population”. WebSci15: Proceedings of the ACM Web Science conference. Oxford, UK. Jun 2015
8. Tinati, R., Wang, X., Brown, IC., Tiropanis, T. and Hall, W., 2015. “A streaming Real-Time Web Observatory Architecture for Monitoring the health of Social machines” in Proceedings of the 24th International Conference on World Wide Web pages 1149-1154, Florence, Italy. May 2015.
9. O'Hara, K., Sackley, A., Brown, IC., Tinati, R., Tiropanis, T. and Wang, X., 2014. “Security and Legitimacy in a Web Observatory.” 2nd International workshop on Building Web Observatories (B-WOW). Bloomington, USA. Jun 2014.
10. Brown, IC., Hall, W., 2015. “Web Observatories: Research Briefing 2015”. Published by Web Science Trust for US Air Force Office of Scientific research (AFOSR): Award No. FA9550-15-1-0020). Southampton, UK. Dec 2015.
11. Brown, IC., Hall, W., 2016. “Web Observatories: Research Briefing 2016”. Published by Web Science Trust for US Air Force Office of Scientific research (AFOSR): Award No. FA9550-15-1-0020). Southampton, UK. Dec 2016
12. Open Data Institute (2016) *Open enterprise: how three big businesses create value with open innovation*. London, UK. Cited as co-researcher. London, UK.
13. Brown, IC., 2013. “W.O.L.F: The Web Observatory Linkage Framework” in Digital Economy Web Science Doctoral Training Centre (DTC) research exhibition. Southampton, UK. Sept 2013.
14. Brown, IC., 2016. “Going Open: [DataNames]: a narrative analysis”. Internal consultancy report commissioned by [DataCo]. London, UK. Jan 2016.

Acknowledgements

I'd first like to acknowledge the support, energy and particularly the tolerance and good humour of my supervisors Wendy Hall and Lisa Harris. To Wendy for her innate sense of what works and to Lisa for keeping the whole thing (including me) grounded. Its doubtless a daunting task to take a person with thirty years' experience of telling other people how to do things and to make him stop talking and listen. They not only succeeded in 'teaching an old dog some new tricks' but did so with great patience and empathy. I shall remember a great deal of laughter in our time together but I think I also managed to make space in my head for some new ideas.

The staff and students at Southampton are an exceptional group and I particularly enjoyed being part of the Doctoral Training Centre from the very first moment. It is an outstanding way to structure the doctoral research process which can otherwise seem fragmented and isolating: I cannot imagine any other approach could be more welcoming and engaging than this. Whilst space prevents mentioning everyone in the group I would particularly like to thank Les Carr and Susan Halford: Susan for gently encouraging me towards an appreciation of the power of social theory and Les for being a literal "fountain" of intriguing ideas about almost everything. The events, trips and opportunities organised/enabled by the DTC are amongst the most rewarding I have ever attended.

During my studies I was privileged to work closely with the Web Science Trust which brought access to a host of new colleagues from Web Science groups in WSTnet and other partners globally and I am truly grateful for the opportunity to work with all of the trustees who are not only genuine thought leaders but also a pleasure to spend time with. I must also thank the WST and DTC admin team most notably Susan Davies, Craig Gallen as well as Claire Wyatt, Nicola Need and Jane Morgan.

I am indebted to RCUK for funding my research and to all those who allowed me access to their projects, their ideas and, above all, their advice/feedback. I am particularly grateful to Nigel Shadbolt, Thanassis Tiropanis and Anni Rowland-Campbell for allowing me to observe their Observatories and Social Machines and for their important insights on Observatories and observation. Also to Dave de Roure for championing the 'social' in the Social Machine and for his deeply insightful views on Web Science. I am also grateful to Ramine Tinati, Xin Wang, Aastha Maaden, Chris Phethean, Dominic DiFranzo and the "Wolfpack" for their feedback and for contributing so much to the enjoyment of the PhD process.

I must also thank all my interview participants for sharing their thoughts and donating their time and in particular the commercial case study partners who enabled access to key participants and invested their valuable time and resources into broadening the understanding of WOs beyond the purely academic view. I would particularly like to thank the Digital Catapult and the ODI for their input and anonymous thanks to J., B. and A. from [DataCo].

Thanks to those who reviewed and fed back on different stages of the document including Ann Brown, Ramine Tinati, Lyn Dunk, David Blundell and Simone Davis.

My family supported me though all the ups and downs of this rewarding but challenging period so special thanks to Ann, Harry and Emily.

Finally, to Joanna Lewis who introduced me to Web science and to everyone who didn't run away when I said "Web Observatory", *I thank you.*

Dedication

For my mother, Nina

Who would have been the proudest of all.

1929 - 2016

Definitions and Abbreviations

CGT	Constructivist Grounded Theory - qualitative research method
DNA	Definitions, Narratives & Agents/Actors
DSS	Decision Support System
GT	Grounded Theory - approach to building theory inductively from data
IoT	Internet of Things (Toasters[1])
IRL	In real life – a physical rather than virtually mediated experience
IPA	Interpretative Phenomenological Analysis - qualitative analytical method
nVivo	Qualitative analysis software package
VO	Virtual Observatory
W^3O	World Wide Web Observatory (network), Observatory of Observatories
WO	Individual Web Observatory(ies)
WST	Web Science Trust
WSTNet	Network of WST members
WWW	The (World Wide) Web

Conventions

Capitalisation/Spelling

- The acronym "DNA" (typically deoxyribonucleic acid) is used as short-hand to refer to an analytical presentation of features of WOs which capture the Definition, Narrative and Agents involved. The idea of D, N and A "genes" is used metaphorically only.
- Focal terms such as Observatory/Observation, Web and Social Machine are capitalised throughout to indicate a specialised usage of the terms beyond the usual meaning.
- WO may imply a physical system or the idea of Web Observatories
- W^3O is used to mean "the set of all interacting Web Observatories" forming the global world-wide Web Observatory
- WO→W^3O is a short-hand used when thinking about how WO scales up to W^3O and is intended to imply there may be differences at scale and may be read as "W^3O emerging from WO"
- The term VO implies the general concept of a Virtual Observatory hosting data (of any kind about any subject) on the Web. WOs are a subset of VOs
- Some phrases are run together as a block e.g. Data-about-the-Web vs. Data-on-the-Web implying a meme/idea
- US/UK spelling: Direct quotations/titles in US English using spellings such a "color", "center" , "specialization" and "program" are rendered to UK English equivalents: "colour", "centre", "specialisation" and "programme".

[1] Earlier humorous references to IoT were made to an 'Internet of Toasters' 1989-90 http://www.livinginternet.com/i/ia_myths_toast.htm

Pronouns

- "She" and "He" as in "The researcher can do X and "she"/"he" will see that X follows Y" is intended to include "he/she/they will see".
- "One" is used as an impersonal pronoun alternative to avoid passive forms as in "one may find several drivers".
- Use of "We".. To avoid the excessive and often clumsy use of the passive voice "it was done" the form "we" will be used in specific instances to aid readability:

    - Where "we" is intended to infer "people in general" as in "we can see that X follows Y" in the sense of "one can see".
    - Where "we" is intended to refer to the author + reader in a joint venture as in "we will consider the following topics in the next section".
    - Where "we" is intended to refer to the author and colleagues in Web Science as in "we are seeking to study data at scale".
    - It will not be used to imply a team of researchers

Other Typographic conventions

1. Edits/Changes/Additions to the text are shown [like this]
2. Where verbatim quotes are reproduced then, they are "quoted like this", whereas material, having been paraphrased, is [quoted like this]
3. Pseudonymous names are shown like this [Elvis]. In the rendering of transcriptions 'noise words' have been removed, minor grammatical errors corrected and linking words inserted e.g. "added .. [between] .. non-contiguous quotes" to aid readability
4. The order and grouping of quotations from interviews have been changed in places to highlight key points, group themes and aid readability. Care has been taken not to change the intended sentiment/meaning of the speaker. Full original transcripts are on file and reviews were offered to all quoted participants.
5. Underlined terms/citations signify hyperlinks to the appendix and/or embedded 'hover-over' definitions
6. *Italicised* terms indicate vocal or narrative stress.

Disambiguation

The term 'Actor' is used in the sense of Goffman's social actors and not in the sense of Latour's Actor-Network unless otherwise specified.

Chapter 1: Introduction

In Short ..

The backdrop to this research is the challenge of observing and understanding the social world through the medium of the World Wide Web (aka WWW, 'the Web')) using robust, repeatable techniques and tools to discover, analyse and share data – i.e., a Web Science context. This drives a second, more specific context - the need for open and reliable data sources and innovative, collaborative approaches to adapt to the challenges of managing the growing volume and complexity of Web data and growing trust/provenance issues in Web-sourced data - i.e., the proposed Web Observatory (WO) context.

1.1 Overview of the Research

In this section, the change in the complexity and pervasiveness of the Web is considered and how this leads to a need for innovative tools to support a science of the Web in academia, government and industry. Yet are all these groups seeking to innovate in the same way and are their goals compatible? Whilst there are proposals and even working examples of WOs there is (so far) little classificatory work and limited engagement from third parties. An integrative model is required to combine the social and technical elements of WOs to support the understanding of how WOs function and why groups collaborate in different social contexts to better understand WO adoption.

This project therefore seeks to widen the focus on WOs beyond technical structures to include context/framing, motivations and notional exchanges and to leverage this multi-perspective model to inform insights on engagement, adoption and collective action for a World Wide WO eco-system (or W^3O).

The Background

Within the emerging discipline of Web Science, we (practitioners) are seeking to understand the changing structure/nature of the Web itself along with the considerable impacts (both risks and opportunities) that this vast, largely unregulated system has on users, markets and society as a whole. With the adoption of Web technologies across all sectors of society, such insights are also of interest to commercial groups, public sector and third sector groups where each may have differing/overlapping objectives for the Web around improving models, enriching knowledge of their markets and informing policy.

Key challenges emerge since Web data can be:

- Complex, massive and dynamic (difficult to process)
- Multi-national, multi-stakeholder, multi-source (difficult to share/license)
- Incomplete, undocumented, unsigned (difficult to validate/trust).

On the WWW's 28th anniversary (12/3/17) Tim Berners-Lee published[2] an open letter highlighting three such challenges for the Web: unethical capture/use of personal data, unethical use of (mis)information (incl. fake news) generally and unethical social (political) control through the manipulation of social media. These challenges are rooted in the trustworthiness and accountability of data – where data is from, what is done with it and the costs of achieving suitable levels trust/accountability.

Keeping trusted records over time (i.e., longitudinal, curated data with provenance) for even a single source or topic may be both technically and financially challenging. The sheer scale of capturing and storing data from many potential sources and topics invites a division of labour and cooperation between communities of trusted data holders and curators. A technical solution to discover/curate individual sources of data has been proposed by the Web Science community and has been termed the Web Observatory (WO). The action of sharing/collaboration between these individual WOs potentially creates a World Wide Web of Observatories (W^3O) whose aggregate potential exceeds that of any individual WO via opportunities for data synergies, collective intelligence and collaborative (interdisciplinary) research.

Consider an analogy: we may collect weather data locally, both historically (through maritime logs), and currently (through local sensors/measurements) in order to combine data into meaningful weather patterns globally and, more significantly, characterise complex models of climate change. No single discipline nor system has been able to address the challenges of climate change fully, and it is only through international cooperation and collaborative interdisciplinary thinking that robust models have begun to emerge. So it is with understanding complex and shifting patterns of data on the Web which may affect us as profoundly (in a social sense) as extreme weather does in the physical world.

The importance of this endeavour (and this research) spans the interests of research done by academia, business and government since the Web exerts a significant and growing influence in all these areas.

[2] http://webfoundation.org/2017/03/web-turns-28-letter/

Despite a huge potential increase of data through social media, smart devices and the coming Internet of Things (IoT), access to much of the most valuable data is increasingly via so-called "walled gardens" i.e., the commercial, proprietary systems of large corporates such as Google, Twitter and Facebook.

In contrast to the stated ambitions of open and linked data repositories such as the Internet Archive, DBpedia and several 'OpenGov' initiatives, relying exclusively on proprietary/commercial services raises long-term strategic and ethical questions about the potential for control/filtering of research data without public scrutiny. How much willingness/openness can be expected from commercial entities to share (license) data, to allow data to be combined/remixed/re-used or to disclose (proprietary) algorithms? The potential for unfair practices and monopoly pricing for data access in the future for researchers in academia, government and smaller businesses seems clear. Such an uneven playing field also has international implications for infrastructure-poor or data-poor countries in a knowledge-based global competitive market. We have seen examples of political arguments over net neutrality and fairness/control erupt such as that between India and Facebook over the proposed provision of a 'free' (but tailored/edited) internet service[3].

In contrast, an open collaborative system might, for example, enable current/domestic research to be re-used and extended with historical/international data freely shared from multiple sources, leveraging the resources and expertise of multiple Observatory contributors. Such combinations are technically and even legally challenging with current isolated or proprietary systems while the results may be significantly more valuable than smaller, manually collected samples which are not designed/stored/curated for reuse and extensibility.

This project is therefore centred around important issues of accessibility, trust and potentially conflicting interests between users of data in what is increasingly becoming a data-led global economy and a web-mediated global society.

The Project

In this project, a series of perspectives on Web Observatories (WOs) will be explored through an examination of related literature, observation of observatory projects/events and a broad range of interviews with users in this space. The output will be a characterisation of WO which goes beyond the current technical perspective to include the social and socio-technical elements which will assist in engaging/encouraging potential WO participants who may build or adapt sources and systems that contribute to the wider WO ecosystem (W^3O).

[3] https://www.theguardian.com/technology/2016/feb/08/india-facebook-free-basics-net-neutrality-row

As an approach for collaborative data-centric research, the Web Observatory invites comparison with existing approaches, tools and data sources, though so far little work has been done to disambiguate WOs or W^{3}O from other systems which may (partly) overlap in function and intent. A global eco-system of Observatories rests on a cooperative/collaborative model and have been characterised as 'Social Machines', i.e., comprising both social and technical elements. If this is accurate, it may be difficult to engage and recruit existing systems and data sources to a Web Observatory model or to scope the creation of trusted sources without understanding both what is expected from systems and from the participants (i.e., *how* to participate and *why* users would choose to participate.) The Web Science community is attempting to bring about this innovation and encourage its adoption and to do this, an understanding of adoption is fundamental.

The Internet (most often appearing to users via the Web and Web-like apps on Smartphones and Tablets) is a relatively recent innovation and yet has seen rapid adoption in mainstream social use via easy-to-use browsers and smart devices compared to the adoption of other technologies historically (Figure 1-1)

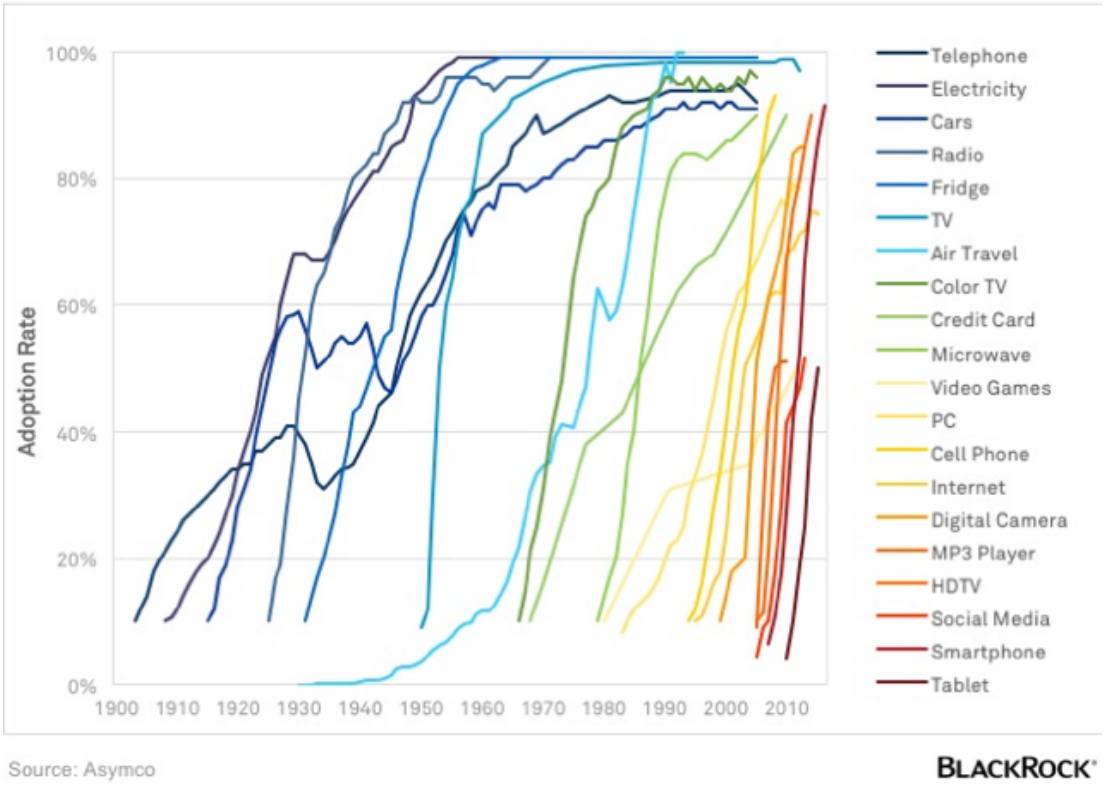


Figure 1-1 Source https://www.blackrockblog.com accessed 03/2017

Two perspectives arise: (1) The speed/extent to which technologies are adapted to fit into society and (2) the speed/extent to which society is changed by, and adapts to, the technology. There may also be ethical questions relating to the extent society may wish to moderate the speed of adoption and mediate the social effects. Web Science emerges in part because the Web is

unusual, if not unique, in that it is both an *agent* of change and a *lens* through which that change can be observed.

Invention is, however, not always synonymous with adoption. An understanding of the process of adoption (See Ch2/Ch10) and the alignment of technology to purpose by society may be thought to be instrumental both in understanding socio-technical effects and also in managing the impact of new technologies - particularly where one might wish to promote/discourage the adoption of certain approaches.

This research will examine the nature of Web Observatories from both a technical and social perspective in order to produce a better understanding of the elements underpinning the process of adoption vs. resistance, interoperation and cooperation between new and existing systems and how WOs may offer new innovative possibilities. Research groups, commercial organisations and funders may have cause to use this work to build and integrate systems into a collaborative Web Observatory eco-system (W^3O) potentially offering access to a vastly wider range of trusted data/analytics than any single system might offer. Indeed, trust in such systems may be amongst the most significant contributions of a potential Web Observatory eco-system.

The Application

Most recently the risks/opportunities of data taken out-of-context, 'alternative facts' or even deliberately misleading data and fake news on the Web, the responsibilities of data stewardship[4] and the need for trust and accuracy have come into greater public focus. Fuelled by the debate around the 2016 US presidential elections and the UK Brexit vote there has been a call for action against fake news, for open transparency and regulation of algorithms[5] in the UK and for a code of ethics/accountability[6].

Using Google Trends (Figure 1-2) as a basic proxy for public engagement/curiosity, we see little change in the trend level of interest in the more standard issues of transparency vs. privacy over the past five years.

[4] e.g., As US government climate change web sites/data sets are taken down under the new administration
[5] https://www.theguardian.com/business/2016/dec/18/labour-calls-for-regulation-of-algorithms-used-by-tech-firms
[6] https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

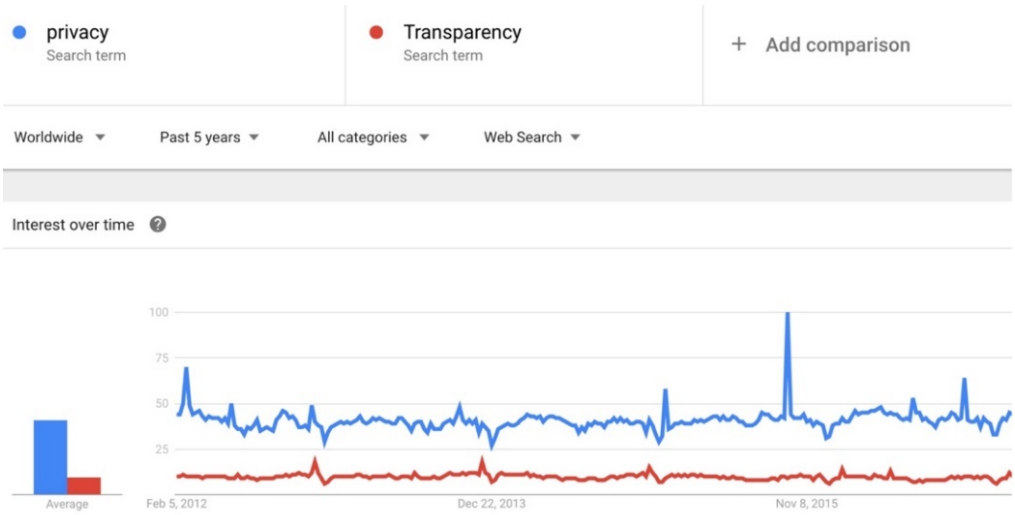


Figure 1-2 Transparency vs. Privacy: source Google Trends 01/02/17

Whereas the notions of "fake news" and "post-truth[7]" (below) appear to be attracting significantly more focus than 12 months ago.

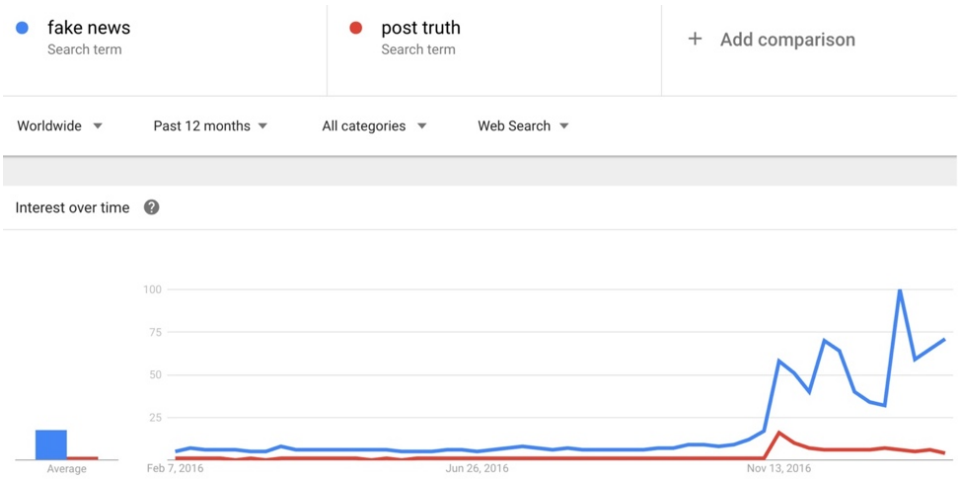


Figure 1-3 Fake news and post-truth: source Google Trends 01/02/17

This trend may suggest a switch in focus from 'data *about* politics' to 'the politics *of* data' and speaks to the implications of "mining" reality from Web data to create a Web-centric version of the truth. Equally the transmission of that truth across cultural networks such as the Web has implications for social control in what David Roberts called 'post-truth' politics[8].

As information becomes increasingly shared and reshared (remixed) via the Web and hence disconnected from the original source, the issue of trust and trusted sources becomes highly relevant. Not only do prominent organisations collect data from/about our activities on the Web

[7] Oxford dictionaries word of the year 2016.
[8] https://en.wikipedia.org/wiki/Post-truth_politics

e.g., Google, Twitter, Facebook, Amazon, eBay but also less prominent *data-brokers* including, for example, nine organisations targeted by a 2014 FTC investigation into the practices of such companies revealing data acquisition on a significant scale:

> "Just one of the data brokers studied holds information on more than 1.4 billion consumer transactions and 700 billion data elements, and another adds more than 3 billion new data points to its database each month."
>
> **source: www.ftc.gov**

Thus, in an environment with so much data being captured/emitted via the Web, we have an opportunity to gather and share research data on/about the Web on a global scale. This requires accessible, trusted services which may be deployed with public oversight for ethical purposes. There are also risks**:** of valuable data remaining underutilised publicly or being "stockpiled" privately for commercial gain with little/no visibility around data sources, validity, aggregation or application. With independent trusted sources of verification/validation such as WOs, we may better address the risks from private/hidden data sources used in economic/policy development which may be filtered, altered or subverted for unethical purposes.

1.2 Observing the evolving Web

.. but the Web of what exactly?

From its inception in the 1990's, the world-wide web (WWW) grew, first slowly, and then at the turn of the century more strongly, not only in terms of domains/websites (Figure 1-4) (www.internet_live_stats.com) but in the ways it was applied. As new technologies improved access speeds and allowed dynamic content to be queried/displayed, information exchange became:

- Increasingly 'read-write'- rather than the previous 'read-only' consumption of static information on Web pages.
- Oriented around transactions between users and the contribution of data

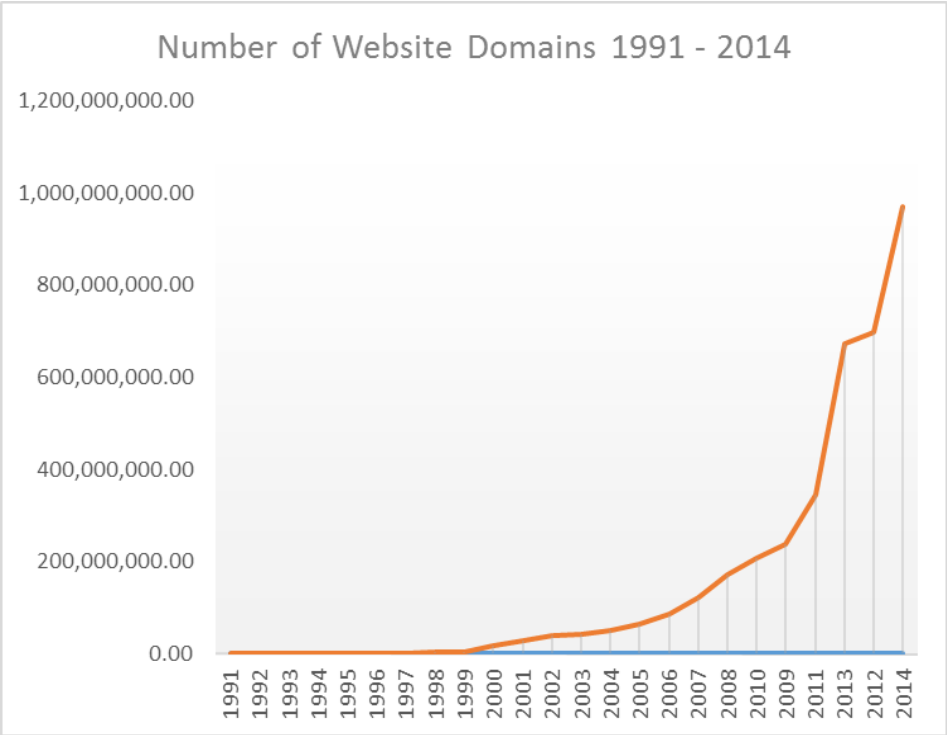


Figure 1-4 Internet growth by number of website domains estimated 1991-2014

The Web thus developed as a platform for eCommerce and collaboration (O'Reilly 2005) and with the ability for non-technical users (rather than only programmers) to add, search and modify content, the scope and complexity of the supported interactions and number of users grew. It has since blended with other technologies, coalesced around other platforms/networks and has "gone mobile", transcending desktop 'tethered' browsers to become part of the substrate for ubiquitous computing thus shaping our perception of the modern Web 25 years later.

We note from Schueler & Hall (Figure 1-5) that, in addition to raw numbers, the types of host and usage paradigms also changed through this period signalling changes in application/usage.

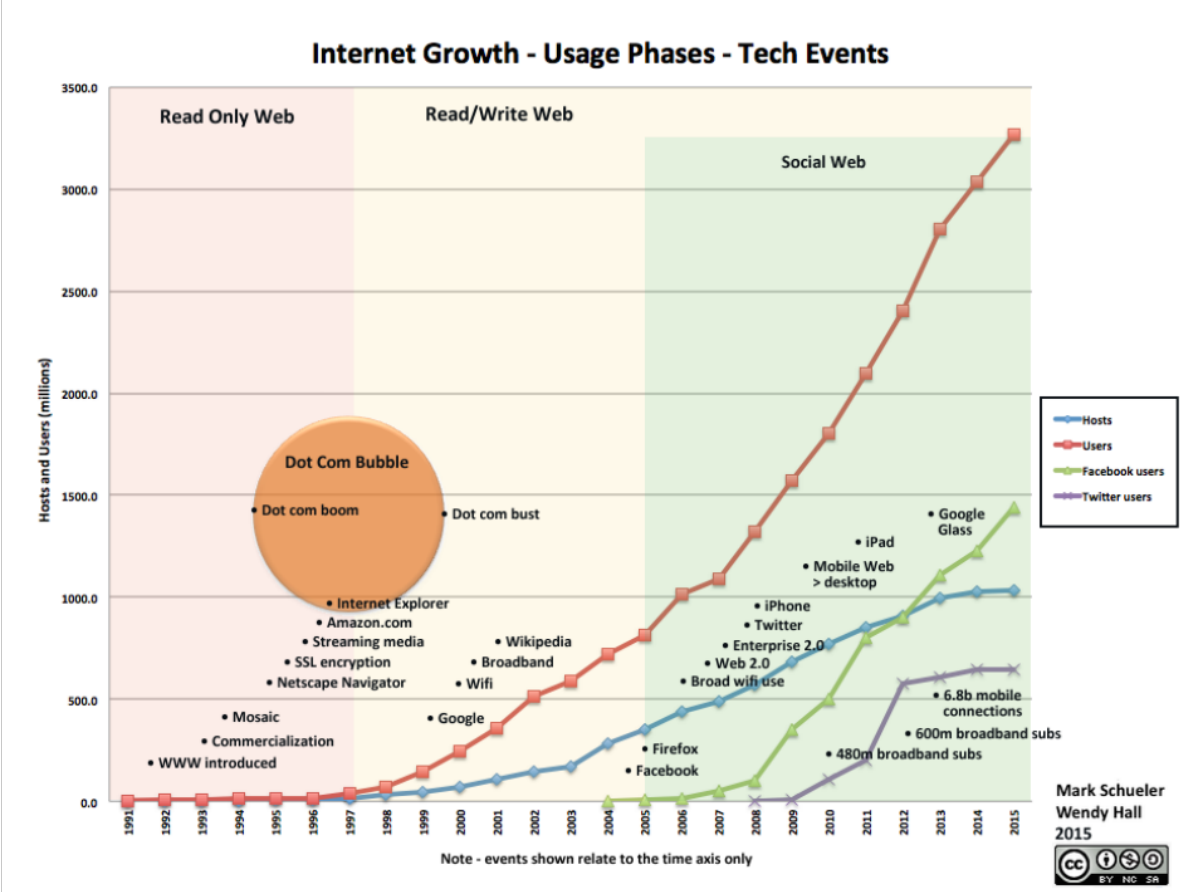


Figure 1-5 Internet growth [9] by phase/platform 1991-2015: (Schueler & Hall 2015)

This ease of access and ever wider offering of content and services has helped to drive WWW from a single server at CERN (a simple but elegant hypertext system handling small sets of static pages on a handful of academic servers) to a globally-distributed Web. This comprises millions of servers, billions of mobile devices and, increasingly, Web-enabled sensors and 'smart appliances'. This may lead to a global 'Internet of Things' (IoT) handling an, even more, vast corpus of messages, rich media and apps. According to recent estimates, there are now more than a billion websites (more precisely website addresses (www.cisco.com)) and more than 3.4 billion users (www.internet_live_stats.com) accessing the Web via more than 7 billion mobile internet devices (www.independent.com).

(Berners-Lee & Fischetti 1999) foresaw WWW as being not only a *Web-of-Pages* but also a *Web-of-People* co-creating/extending the system and also a *Web-of-Data* allowing users to come together with data/applications via Web infrastructure and standards. Despite this, decades of development separate the broad adoption of each of these models and the tools to model/monitor each type or phase of the Web are disjoint and distinct.

[9] http://growthchart.weebly.com/

Usage, content and delivery on the modern Web have changed significantly from a static HTML page model for academic documents and now mediates a wide range of complex social interactions spanning crime to government and collaboration and communication from the local to the global. What it means to 'study the Web' has therefore changed in term of volume and complexity and a new discipline of Web Science (including the use of so-called Web Observatories or WOs) has been proposed (Berners-Lee et al., 2006), (O'Hara et al., 2013). This project will seek to examine the WO concept in terms of what they are, what they do and why users may choose them as an approach.

1.3 Evolving more complex Web tools

.. in the face of an evolving Web ecosystem

While there have long been analytical tools for Web servers, the demands of the modern Web greatly exceed the original remit of simply tracking "page hits" and server performance. We now analyse the Web-of-People and their choices using social analytics *across* systems rather than for individual sites or servers, and increasingly we are engaging with a Web of semantically-rich data crossing system- and user boundaries. Thus the need for new tools for new measurements is driven by the scale/complexity of the Web's own evolution.

As the focus has broadened from tracking individual pages, apps and servers to tracking behaviour and relationships between people persisting *between* different apps/servers, the conceptualisation of Web data itself has shifted both in analytical focus and in the intended audience. Focus has moved from individual pages/apps to 'social graphs' which analyse relationships across app networks. As the Web continues to evolve the importance of linked/structured data through so-called 'knowledge graphs' grows in importance. This reflects the strong growth in volume and velocity from specific sources (so-called Big Data) and data gathered and synthesised from multiple sources: so-called 'broad data' (Hendler 2013).

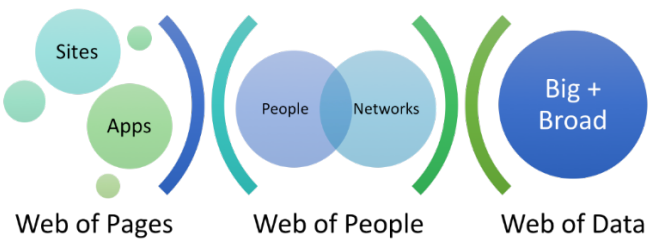


Figure 1-6 Web-of-Pages → Web-of-People → Web-of-Data

As the conceptual focus moves from left to right in Figure 1-6, the underlying dynamic of who/what is producing the data and the tools needed to analyse this data have evolved accordingly:

e.g., basic Weblog tools have been replaced by analytics for the Web, Social Media, Apps and, increasingly, IoT and Big (Broad) Data.

	Web-of-Pages *Website/Blog* *+ linking model*	*Web-of-People* *Social network +* *profiling/advertising* *model*	*Web-of-Data* *Big Data/ IoT* *+ AI model*
Data Produced chiefly by	Humans	Humans	Machines
Data Consumed chiefly by	Humans	Machines	Machines

Table 1-1 Data by whom, for whom?

I will broadly characterise (Table 1-1)

- the *Web-of-Pages* as "data created by humans for humans to consume" i.e., data that is generated by people for people to read directly
- the *Web-of-People* as "data created by humans for machines to consume" i.e., data input by people ostensibly perhaps for other people to read but often aggregated by machines for analysis/modelling
- the *Web-of-Data* as "data created by machines for other machines to consume". i.e., data both created and consumed programmatically often at such large scale that it defies human bandwidth to consume/inspect it personally.

With such a shift in scale and complexity, this suggests the possibility that tools/approaches designed for older, less complex ecosystems might be rendered obsolete or at least ill-equipped to deal with newer paradigms, prompting an opportunity/requirement for new tools to 'exploit' this new ecosystem.

Cybernetics (Ashby 1956) offers us the "law of requisite variety" which links the resolution/capability of tools to the systems which they seek to measure:

> e.g., we might be ill-equipped to document the nature of rainbows armed only with a *black-and-white* camera.

Thus as the Web gets 'bigger' in the sense of 'big' data (more volume, variety and velocity), there is an increasing requirement for more powerful tools and techniques to collect, curate and analyse that data. Approaches need both greater capacity and resolution/differentiation: e.g., Mapping content vs. user choices/behaviours, allowing us to build and analyse social graphs to understand the increased complexity of what is being transacted socially as well as technically.

The complexity of the tools develops as the nature of measurement grows more conceptual based on the underlying physical measurements, i.e., from clicks to purchases, from 'Likes' to popularity and from tweets to sentiment and even threat assessment[10]. The business models of a growing group of purely data-oriented companies have seen the creation of some of the largest corporates in history.

These corporates invest heavily in specialised analytical systems since simpler tools developed to monitor earlier, less complex versions of the Web presumably fail to capture new interactions and nuances for which they were not designed, leaving a gap for new tools/approaches to support an understanding of their own markets via the Web.

1.4 A Need for Web Observatories?

To address the gap between older approaches and new broader requirements, a new class of tool called Web Observatories has been proposed by the Web Science academic community inspired by the success of the Virtual Astronomical Observatory[11] (VAO) programme. Given, however, that Web Science has access to existing Web data repositories, archives and analytics, is there, in fact, a need for a new class of tool as described by (Tiropanis et al., 2013)?

It might be argued that existing repositories and analytical tools are sufficient for the purposes that the Web Science community are suggesting with minor technical adaptations and hence new Web Observatories might be unnecessary. This may, in part, be true and indeed part of the WO vision is to recruit *existing* sources/systems to a broader eco-system as the astronomers did with the VAO. Thus, not only the technology but also the *collective action* is relevant here, and we will return (Ch8) to the VAO to see what can be learned from the astronomers' experience with this approach.

[10] Companies such Recorded Future, iSight and Palantir offer event-oriented Web analytics measuring threats against individuals, companies and markets.
[11] http://www.ivoa.net

Individual instances of WO have been characterised by the community (e.g., Tiropanis, Hall, Chua, de Roure - see Ch2) to be (trusted) assemblies of information. Data is combined with analytics/apps to focus on particular specified phenomena (i.e., not monolithic stores centralising 'all knowledge'). Such individual systems can reduce the technical hurdles to access local data for researchers/observers by providing a centralised source of historical/current/modelled data.

In essence the WO in (Figure 1-7) gathers and links data, metadata and comments/annotations over time in order to assemble historical views, current/snapshot views and simulated/modelled future views of the data and share/provide these views as services to users.

An opportunity exists, however, to enhance local insights with aggregated datasets, tools with broader research data through linkage to other distributed WOs. The coverage and synthesis across different domains are achieved through discovery and sharing between WOs forming a proposed *World-Wide-Web of Observatories* (W^3O). Thus, it is essential to view individual WO in the same light as individual Web nodes/servers, i.e., with the intention that multiple instances will 'interoperate' forming larger, more powerful structures based on multiple data sources.

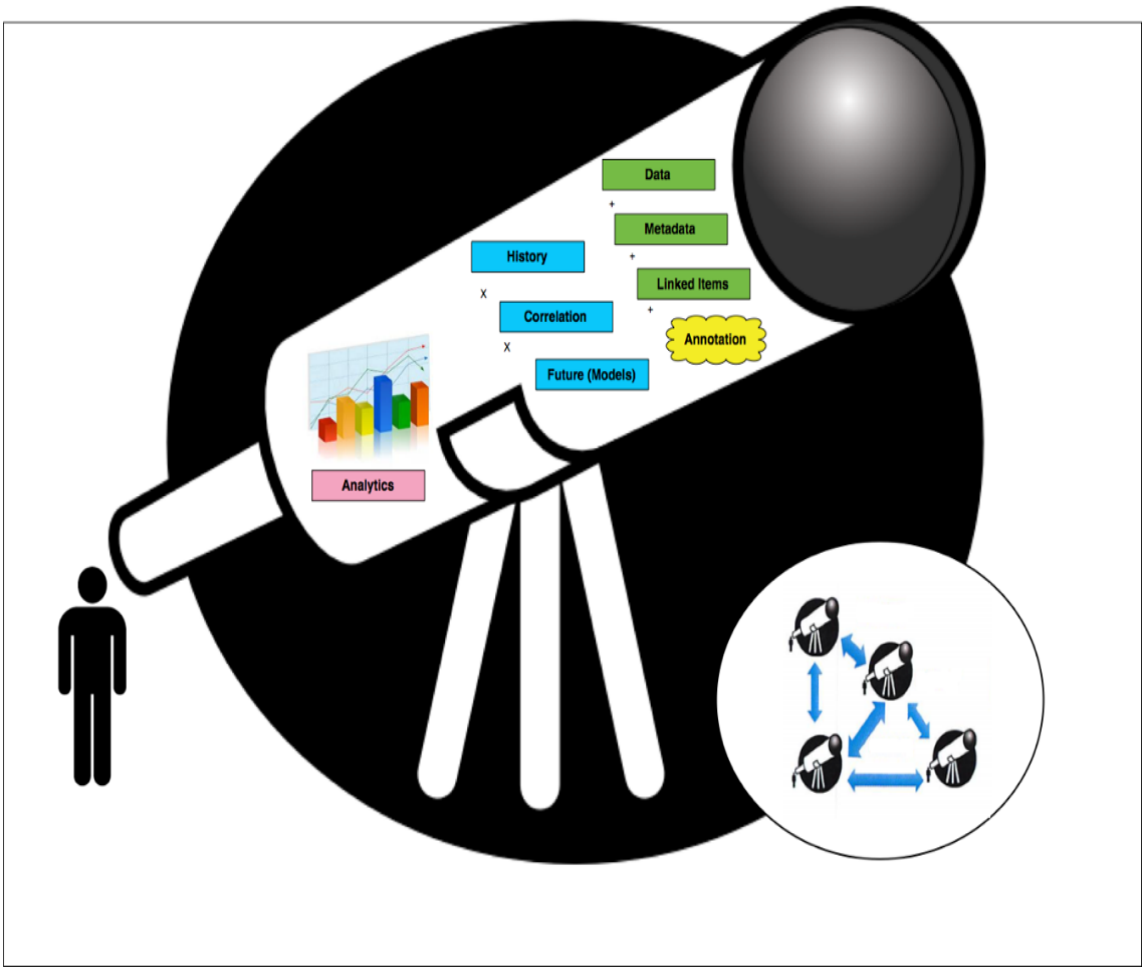


Figure 1-7 Generic WO and W3O concept. Adapted from (Brown 2013)

WO vs. W^{3}O are related but distinct ideas (akin to Web Servers vs. "the Web") and for the remainder of this document a clear distinction will be maintained between individual physical WO nodes and the effect of interconnecting WOs to allow the emergence of W^{3}O which is denoted hereafter as WO→W^{3}O.

To clarify the implications of this distinction a summary of the broad conceptual differences from (Hall & Brown 2015, 2016) are shown in Table 1-2 and whilst not intended as a strict normative definition of WOs, this is a useful position from which the research will proceed.

	A WO	THE WO (W³O)
Is a physical system	Yes	No
Is owned/operated by an authority	Yes	No
Draws data from more than one source	Maybe	Yes[2]
Contains Open Data	Maybe	Yes[3]
Contains private data	Maybe	No[4]

Table 1-2 WO vs. W^{3}O: reproduced from Hall & Brown (2015)

- Note 2: If only one source were available this would effectively be a single WO.
- Note 3/4: Unpublished/private data would, by definition, not be visible for W^{3}O consumption (whatever the selected license)
- This model supports *combining* private data from one/more selected WOs with openly shared data from W^{3}O.

1.5 Extending the Current View

In proposing a global network to support our understanding of the Web, the importance of understanding the requirements and expectations of the corresponding networks of machines, datasets and people is implied – as is the need to understand the nature/scope of what is required to participate. The Web itself grows organically and without overarching global ownership/authority (albeit based on global technical standards) and thus it may be as important to understand (socially) *why* users will participate as it is to understand (technically) *how* they do so. Given:

1. There is little published work defining or disambiguating Web Observatories from other classes of system, despite work on specific examples and general references in the literature. (Tiropanis et al., 2013), (Tiropanis et al., 2014), (Brown et al., 2014), (Difranzo et al., 2014) and (Walker et al., 2015), variously propose that WOs may be engaged with a range of research in academia, business and government.
2. It is not yet fully established to what extent existing systems/repositories may simply choose to 'act' as Observatories vs. needing to build bespoke systems/interfaces to participate or the extent to which existing services (e.g., Google) may evolve to support 'observation' as a paradigm.
3. WO has been characterised as a *Social Machine,* and yet existing Observatory work mainly describes individual WOs from a structural/material perspective. If WOs are Social Machines what are the 'social constraints' (described by Berners-Lee in 'Weaving the Web') that are being addressed by this machine?
4. Given WOs in a broader forum are more diverse than a single instance of a WO we also require a basis from which to combine systems and potentially to examine the engagement between WO systems, something which has received little focus so far.

If we use a social-cognitive model which considers that WO is an arrangement of data/functions that is given meaning/significance by individuals and social groups, and that these meanings drive users to perform/behave in certain goal-seeking ways, we may examine WO from the technical and social perspectives of:

1. Materiality (what it *is*)
2. Significance (what it *means*) and lastly
3. Whether (considering theories of performance) how WO may be *transacted* as it is used and applied to real-world exchanges of data.

Below I will argue that each individual perspective is necessary-but-not-sufficient to characterise the operation of WO and W^3O. Thus the goal of this research is to capture each perspective individually as a vocabulary of potential[12] elements. This offers a way to depict, structure and arrange them such that each piece can be considered alone, but also their complex interactions and emergent behaviours may be modelled and analysed allowing us better ways to predict and to encourage the interoperation between diverse systems.

[12] It is not proposed that WOs must exhibit *every* element

1.6 Limitations & Disambiguation

Within Web Science, WO belongs to a broad discussion of the impact and properties of the Web itself, how it relates to the underlying technical and social networks and how our co-creation of content, structure and purpose on the Web creates emergent properties and behaviours. This Web *Science* WO is, however, only one characterisation of the data-gathering and analytical approaches for the Web used in government and business which may be largely indistinguishable from WOs in other than name and which might overlap and interoperate with WOs. It is, therefore, important to *distinguish* between systems focussing on Web Science issues from systems (i.e., Virtual Observatories/VOs) which gather data on/about the Web for other purposes but not to *exclude* them.

While generic VOs may not serve as WOs in isolation, they may validly provide services/resources or interoperate with WOs enabling novel insights in an emergent W^3O (a Web-OF-Observatories). Hence this second broader class of non-academic observatories is not so easily discounted from the scope of this project since several core concepts around the acquisition and management of data 'span the tribal borders' between Academia, Business and Community/Government.

Caveat/assumptions

For this project, a focus/balance must be found between narrowly excluding any system which chooses to coin a different system name (e.g., 'Social Observatory') and characterising WO too broadly to include any/all systems which simply store/process data on the Web.

Some *ex-ante* assumption/assertions are illustrated in (Figure 1-8):

1. The existence of WO systems *in vivo*, a body of published research material and an active research community implies that no further proof of the existence of WOs per se - at least as a meme (if not as a novel approach) is required
2. Not all Observatories are ***about***-the-Web even if they are ***on***-the-Web
3. Not only data about-the-Web is of interest to Web Science and WO (W^3O) participants/providers will have interests other than Web Science
4. Not all data on W^3O will necessarily be about-the-Web or even originate on-the-Web
5. Not all data on WOs will be open – but only shared data is notionally visible via W^3O.
6. For a WO to be Social Machine both social and technical elements must be considered.
7. Not all Social machines are Observatories and *vice versa*.

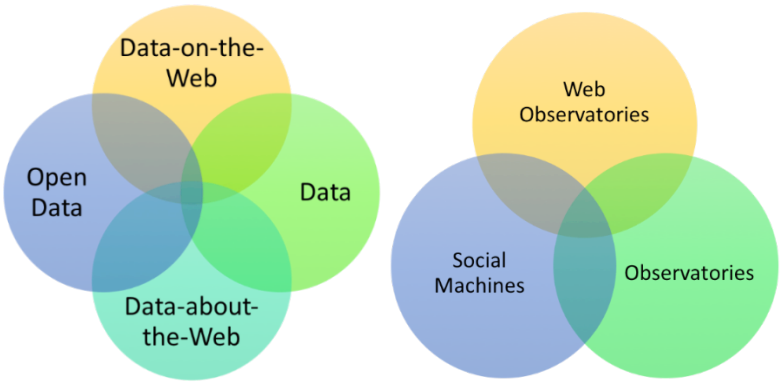


Figure 1-8 Disambiguating related but distinct concepts

1.7 Research Questions & Approach

1.7.1 Research Questions

As we attempt to engage with multiple groups/stakeholders as part of a global, collaborative WO effort we may need to consider WOs as a socio-technical (rather than a purely technical) idea to encompass the notion of collaboration, motivation and cross-cultural differences. This drives the following research questions (expanded in Ch3):

- Which perspectives can help us to clarify the structure and nature of WOs not only as purely technical artefacts but also as an assemblage of users and technologies within a social context?
- If we consider WOs as being socially-embedded in the processes and ambitions of different groups who use it, is there evidence to suggest that WOs might be perceived, structured or operated differently across social groups?
- What benefits can a socio-technical model of WOs offer in terms of insight into the creation of new observatories, innovative applications and the encouragement of participation by existing systems and data sources?

Beyond the *intrinsic* technical elements and (social) user elements of WOs may be a number of *extrinsic* factors that must be considered as eco-system factors (inputs) or emergent properties (outputs).

> e.g., Trust may be a critical input/output of the WO yet trust is neither strictly speaking a technical "component" nor is Trust a user.

1.7.2 Research Approach

The first part of the project (Ch4-5) identifies a set of features/facets from the WO discourse to develop a supporting taxonomy for WOs, comprising different clusters of blocks/facets from which WOs may be constructed/construed. The second part (Ch6-8) examines the conceptualisation of WOs across three sectors or 'tribes' (Academia, Business and Community) through the personal experiences and viewpoints of multiple practitioners – including my own observations/experiences. The earlier content analysis is thus cross-validated while examining key elements of what existing WOs and similar/related systems are doing within particular social frames/groups. The emerging model is presented (Ch9) and observations from the research and on the model are presented (Ch10-11).

Care has been taken to avoid offering broad tautological observations as novel findings (i.e., that governments tend to act more like governments and less like universities) or that individuals/groups tend to act in their own self-interest). Instead, the focus is on how groups/features/behaviours may differ within a WO context and to what extent this is reflected in a model that supports/incentivises a desirable interoperation between WOs. Indeed, the idea of 'performativity' in networks (Healy 2015) and data more generally (Gloria et al., 2013) based on earlier work by Butler, Derrida and Callon allows for the reversal of the idea that '*roles produce particular behaviours'* focussing instead on how '*behaviours may instantiate new roles'*. In this case identifying specific behaviours/interests may allow for the identification of emergent roles/groups that are of interest to WO communication/adoption.

Such emergent roles may give improved perspectives on how users aim to innovate with WO and what factors will encourage/dissuade their participation.

1.8 Structure of the Report

Ch1 – **Introduction** has introduced the idea of the Web Observatory as a tool for Web Science in response to an evolving Web. The idea of examining both the social and technical elements of exemplars to define a meaningful vocabulary for Web Observatories is presented.

Ch2 – Literature review presents selected literature relating to Web Science, Observatories and the development of analytical tools.

Ch3 – Research Framework introduces the research framework based on a review of research literature and comparable research projects.

Ch4 – Conceptualising WO reports on experiments comparing WO with other classes of system and exploring varying WO perspectives and how they may be embedded in roles.

Ch5 - Seeding the WO Model presents the development of the initial WO taxonomy from source materials.

Ch6 - Testing/Refining the WO Model presents views of the WO vs. W^3O and considers the supply of vs. demand for data.

Ch7 - Pilot projects presents results from observations and analysis of pilot projects.

Ch8 – Participant Interviews presents interviews and themes from three different tribes and new roles which emerge.

Ch9 - The DNA of Web Observatories presents the final grounded theory and associated models.

Ch10 - Observations and discussions summarises the overarching themes and recommendations arising from the individual experiments.

Ch11 - Conclusions summarises the findings of the research, outlines caveats and limitations of the approach, possible implications of the findings and suggests future work.

1.9 Conclusion

In this Chapter we have explored the growth in size, complexity and importance of the Web and the opportunity to study the effects of the Web on society (and *vice versa*). More powerful tools are required to achieve a study of a more complex Web and this leads to the proposal for a shared virtual observatory for Web data based on an eco-system of individual sources and tools. As part of a collaborative exercise to share data, understanding the objectives and requirements of the diverse collaborators is an important factor in ensuring adoption and participation. Thus, in addition to the important technical work on WOs which has been achieved so far, this project seeks to extend a technical view of WOs to a wider socio-technical view.

In the next chapter, we will consider a selection of literature relating to data innovations, the development of the Web, Web Observatories and the study of socio-technical systems.

Chapter 2: Literature Review

In Short ..

The ability to deliver infinitely replicable digital data via low-cost, high-speed global networks has created *hypermobility* in information. This has enabled innovative ways to access media, information, analysis and other electronic assets which can be bought/sold, stored and shared globally and almost instantaneously at near zero marginal cost. Models of how such information innovations are successfully adopted are introduced, and the social-vs-technical nature of innovation per se and the corresponding Social Machine perspective for Web Science are discussed. Examples of open systems known as Web Observatories (WOs) based on earlier scientific virtual observatories (VOs) are introduced and whilst originally grounded in a purely academic context (i.e., Web *Science* Observatories), systems/sources innovating outside of academic research are also included since these may also contribute to a global shared Observatory (W^3O) eco-system and hence to a broader understanding of the Web and society as a whole.

2.1 Introduction

In the first half of this section, the historical path of information innovations and the idea of technical innovation and adoption per se will be reviewed. Accessibility to data and technology will be traced from ancient physical libraries and curated collections to what has become the modern Web-of-Pages and Web-of-People with its 'data deluge' and the emergence of searching tools. The importance of trusted data and its relationship to power/authority is noted. The relevance of studying the impact on society of digital data and the Web will be considered in addition to collecting data from/about the Web as a proxy for social structure and behaviour.

In the second half, we will consider, with examples, virtual repositories of data-***on***-the-Web including those capturing data-***about***-the-Web over time by locating, harvesting, linking and sharing targeted data sets. The goal of better understanding society for academia, business and communities via a science of the Web is introduced. Current work in Web Observatories is examined and by bringing together underdeveloped themes in the current literature, the groundwork for the selection of research methods in Ch3 is laid.

A note on grounded theory and prior explanatory theory

In Ch3, a *grounded theory* approach will be proposed – a method in which new substantive theories emerge from observations and are thus inductively "grounded" in the experimental data and not in prior theories. It is notable that grounded theories should not rely *ex-ante* on other structural or explanatory models, i.e., until *after* the emergence of the grounded theory (that is *ex-post*). Ch2, therefore, includes a review of background/focal theory relating to innovation, digital data and Web Observatories but *excludes* references to pre-existing explanatory theories of how the research findings may be interpreted until these theories are examined in relation to experimental results. Thus, in Ch10 we will revisit broader theoretical perspectives to consider how these relate to the substantive experimental results which underpin the grounded DNA theory.

2.2 From 'Parchment & Pages' to 'Podcasts & Pokes'

The power to distribute and determine the truth

Philosophers/academics at least as far back as Mesopotamia and the ancient Greeks have collected information in physical records/repositories in an attempt to describe and preserve knowledge in a form which is safe and trustworthy. (Krasner-Khait 2001) gives a brief but informative overview of the transition from private collections (such as Aristotle's) which were selected, curated, preserved and often hidden/guarded from the public through to the creation of the first libraries in a modern sense (such as the great library of Alexandria 300 B.C.) where wider access and sharing knowledge publicly became more fashionable. Knowledge was often stored as physically fragile parchments, which were difficult/expensive to copy (via specialist *scriptoria*) and impractical for many people to access. Innovative storage technology was introduced as more robust flat-bound (stackable) wooden books (codex) with pages arrived in the 2nd century A.D.. The library concept was further developed by the church in the early centuries A.D. with both the production (copying) of material and lending (distribution) of books between monasteries.

The development of Gutenberg's movable type technology in the 15th century broke the stranglehold of a handful of specialist groups (including the church) with regard to the manual production (copying) distribution/curation of books and increasingly brought printed material to society in a mass-produced format. Whilst the central idea of early libraries and books was accuracy and trustworthiness of the information contained in the 'book' - the responsibility for defining/ensuring accuracy led to two further consequences:

- The crystallisation of specific narratives and interpretations of what was true/accurate (i.e., what is *produced* (published) gains the status of 'truth')
- The focus on hegemony and scientific-, religious- and political orthodoxy. Those who produce it, endorse 'the truth' and are apparently endorsed *by* 'the truth.'

Thus, we observe that the ability to emit/distribute information is related to the ability to propose the truth of what is being distributed and the credibility of the distributor.

Technology plays a vital role in the social impacts of controlling the sharing of data (as pictures, words and music). In (Kittler 1999) *Gramophone, Film, Typewriter* a clear link is established between controlling the publication/performance (i.e., the availability) of information/media and the social influence exerted by 'gatekeepers' who regulate supply, costs, locations, skill/knowledge and orthodoxy. The introduction of key media technologies affected the balance of control/power through the ability to duplicate (encode) information as (initially analogue) recordings which could be more easily copied and distributed/broadcast to a wider audience. People were no longer subject to monopoly prices for unique/timed live performances, nor (via the Typewriter) did they need to convince/pay publishers to distribute their thoughts, histories, viewpoints or scientific findings. More modern electronic tools also allow the freedom to edit, reuse and recombine existing materials in what (Lessig 2004, Lessig 2008) calls "Free Culture" and (Diakopoulos et al., 2007) calls the "Remix Society". Both free access and the ability to remix have direct relevance for the value underpinning an open WO eco-system and the shift in authority that this might entail. Thus, a tension arises between technologies which restrict, license/meter usage of media/knowledge vs. those technologies which enable distribution, sharing and remixing; a tension which results in commercial/ideological conflict. Kirkpatrick generalises this link between technology and social power stating that:

"There is no experience of technology that is not at the same time an experience of a kind of social power."

(Kirkpatrick 2008)

He is, however, careful to distinguish between different classes of power ranging from influence/authority (which are consensual and based on an acknowledgement of some *right* to obedience) up to domination/coercion (which are imposed and based on a fear of the consequences of disobedience). Society, he argues, and those who "live together" share infrastructure which mediates, shapes and gives shared meaning to their experiences and this extends to technology which confers choices/abilities on their users ultimately shaping their behaviour. Each tool removes/circumvents some limitation/blockage faced by the user until the

next blockage/constraint is encountered: a process reminiscent of the theory of constraints described by (Goldratt 1984). Each instance of the technology implies "a social construct/group from which it is used and in which it is embedded" (Fichman 1993 in Kirkpatrick 2008) and the corresponding "politics of design" (Feenberg 2002) shapes/guides the underlying technology towards the solution of a particular problem. The classic example of design tension between problems is the goal of *transparency* which requires more data/distinctiveness versus the goal of *privacy* which requires less data/distinctiveness. Thus, the deep integration of technology in society and the social impacts of information technologies on society are established, and this theme is closely allied with the dematerialisation of physical assets and the development of digital encoding.

Digital data and networks

Over the last 60 years, a very recent development from a historical perspective, information science has developed mathematical theories of encoding information digitally (Shannon 1948, 1949), (Nyquist 1928) and distributing via electronic information networks (Licklider 1968), (Cerf & Kahn 1974). This has enabled real-world implementations of globally interlinked systems that mediate vast collections of information in digital form.

Vannevar Bush's vision of the "Memex" had previously provided a vision for 'browsing' through huge stores of linked information using a central viewer in "As we may think" (Bush 1945). This inspired later hypertext systems (Nelson[13], (Engelbart 1962)and alternatives to the Web such as Gopher and Microcosm) and eventually to global hypertext-based systems such as WWW (Berners-Lee 1989). Linking systems and sources together effectively, however, required digital networks and global, resilient inter-networking protocols that started life as part of the ARPANET project proposed by Licklider in the early 1960s as "an electronic common for all". This led to what we now know as the Internet. Thus, combining the means to encode, store and transmit/receive data digitally, the core technologies for surfacing this data via a 'Memex-like' paradigm such as Berners-Lee's WWW were in place. Several vital technologies converged to enable the delivery of products and services in ways that were not practical/affordable using older, less performant analogue technologies. While the early pioneers may not have envisioned the World Wide Web, their contributions enabled it:

[13] Xanadu project

- Multi-source linked information browsing (Bush)
- ARPANET (Licklider, Taylor, Sutherland)
- TCP/IP (Kahn, Cerf)
- Hypertext (Engelbart, Nelson)
- GML (Goldfarb)
- Wireless telephony (Fessenden) and Cellular networks (William Rae Young).

The ultimate result was the creation of the Internet - a cheap (subsidised), reliable, high-speed, ubiquitous, global communications network (both wired and wireless) and WWW as a universal interface amongst (and between) social groups for real-time, multi-media processes and interactions between groups of humans and also between human and machines.

As computing, media and telephony devices have merged in form and function, networks of powerful, low-cost mobile devices have become ubiquitous (Weiser 1993) and increasingly pervasive from a technical and social perspective, enabling communicating, sharing and interaction (Rousch 2005). This underpins an array of social processes ranging from entertainment to business (O'Reilly 2005) and from government to crime (Yip & Webber 2012).

The early focus was chiefly on models of dematerialised electronic "content" (e.g., music, podcasts, pictures and pages of the written word) which transformed the nature of publishing and the economics of physical distribution and gave rise to new models of what could be easily shared and consumed. This has expanded to include models of exchanges in the form of dialogues, choices, transactions and relationships comprising

1. What people say (post)
2. What they think (comment)
3. What they 'like' (upvote)
4. What they buy/sell (even *abandoned* transactions)
5. Whom they know (who they 'poke' or friend)
6. How they conceptualise (tag).

This represents the surge in *datafication* [14] delivering a much richer model of transactions and communication between networks of users across systems globally. Moving from reading static pages to reading/writing via user-editable 'wikis' and messaging/blogging platforms opened the

[14] The rendering of something into an item of referenceable data

way for users to interact with others via a 'Web 2.0' (O'Reilly 2005). Cheap access to global computing and network resources have made digital networks a convenient and cost-effective platform to self-organise nearly all types of business and leisure activity (Shirkey 2008) making the study of 'digital society' an effective proxy for studying society itself.

Studying digital data as it appears on the Web and what happens as 'digital footprints' of our behaviour are created are a fundamental approach of Web Science (Berners-Lee et al., 2006) This delivers a unique opportunity to understand in detail how we live, what we think/want and how society interacts/changes over time in a way that has never previously been practical/accessible.

The accuracy and context of the data gathered is, however, fundamental to the accuracy/truth of any research findings. Returning to the historical credibility of publisher and published data discussed above we have seen the creation/aggregation/distribution of content in a modern digital context. Increasingly this is separated from the clarity/reputation of who has produced/claimed it (and how?) leading to *unvalidated* data, unsubstantiated claims and lower confidence in the unregulated/uncurated Web as a source.

To understand the nature and context of the data that is gathered, we have seen a growing focus on data-**about**-the-data (so-called *metadata*) which looks at classifications/aggregations, types of activity and profiles/locations of behaviour. Large bodies of digital (meta)data are being harvested about the Web and its users by community systems such as the Internet Archive (Kahle et al., 2001), by hardware/infrastructure providers for mobile networks and handsets, by commercial search engines such as Google (Brin & Page 1998) and even contributed by the crowd in systems such as Wikipedia. This forms a unique historical record of social discourse, political thought, business/markets and mapping the movements and actions of the human race.

The growth of social networks such as Facebook and search engines on the Web (providing services without an explicit service fee) are founded on studying/modelling this type of data and extracting insight/revenue from it. We see some of the largest and most profitable companies in the world using this approach. Comprehensive data gathered by these few companies is however typically considered proprietary, making it difficult for researchers at other companies, governments or universities to obtain access - even where their research does not impact commercial interests. This creates potential tension between free/open data systems and commercial offerings.

Commercial groups may not automatically share valuable data assets (about us) without some inducement to do so. This has created a significant asymmetry/divide between a small number of companies (Berners-Lee calls them “walled gardens”) with vast information resources and researchers in government, academia and the remainder of smaller companies who would also benefit from insights into social data but have limited access.

Such data is unique and valuable but without careful stewardship may over time become increasingly private, proprietary and difficult to access thus underscoring the importance of ‘digital data’ (in terms of technologies, policies and access/ownership) to be highly pervasive across sectors of society.

In the next section, we will move from the digital data innovations described above to innovations *per se* and consider how such changes/innovation are modelled providing a contextualised view of that-which-is-adopted versus that which is not.

2.3 Innovation/Adoption

This section reviews contributions from key authors including Rogers, Kline, Ram, Abell, Markides, Davis, Adner and Christensen in order to reflect on technical versus organisational/social factors for adoption. This prepares the following section on the *impacts* of adoption due to different types of innovation: sustaining (evolutionary) innovation versus disruptive (revolutionary) innovation.

This review is relevant since if effective/realistic models for innovation and technology adoption can be established then patterns/behaviour within the WO eco-system might be observable and/or predictable. Outcomes may be better understood, ultimately leading to the development of better theories to manage the WO adoption process and the impact of WOs. In Ch10 we will revisit these models in the light of experimental results and the grounded theory which emerges from them.

The idea of Innovation

In modern usage to innovate (lit. "to make new") has become infused with the idea of something inherently good though logically as (Ram 1987) argues this may not always be the case.

> e.g., Burning witches instead of drowning them may have been *innovative* at the time - supplying as it did a single method to both punish and deal with a dead body afterwards. The users of the ‘new service’ were undoubtedly no more enthusiastic about this approach than the method it replaced.

Whilst this may seem to be a flippant example, the point of differentiating between *supplying* innovation and *demanding* innovation is a serious one and distinguishing between different types of innovation will be a key theme below.

Despite the evidence from Markides' London Business School research to support the idea that only 15% of new entrants will survive the first 5 years of entering a new market, it has been difficult to predict where dominant technologies will fail to defend against challengers and be supplanted. Christensen's widely-quoted model of disruptive innovation (Bower & Christensen 1995) has proven to be more *explanatory* (in a retrospective sense) of what has happened and less *predictive* of what will happen. (Markides 1998) and (Markides & Crainer 2010) have argued that successful challenge is often predicated on "breaking the rules of the game": new entrants with little chance of challenging in terms of market share, performance or customer experience in a status quo, may experience considerably more success where the market decides (or can be *persuaded*) that the rules/requirements have changed owing to the (now) desirable features of newer entrants.

To understand the likelihood of WO adoption we might then consider what problem WO is addressing, for whom it is being solved, with whom a WO ecosystem is competing and whether WO represents a completely new/empty space or extends an existing approach.

The innovation push/pull

In the classic models presented by Schumpeter, Bass, Rogers developed in 1930-1960's, innovation was considered to be a linear process consisting of Invention (ideation), Innovation (development) and diffusion (deployment) driven primarily by research. This simple linear model is probably over-simplified, and in reality, there can be numerous feedback/feed-forward loops, which dampen or accelerate the changes made to the original idea after being fed to the market.

- For a *proactive* or push model research groups may be considered to be paramount and hence drivers of the process - supplying new ideas and designs and passing them through a development + improvement process before being released to the market.
- Alternatively, the model can be considered to be *reactive* to demands from the market in which demand filters back to the improvement + development process for changes to existing services and, in turn, calling on new ideas from research if incremental improvement is insufficient.

(Abell 1980) argues that all companies are driven by the answer to three questions:

- Which customers (who) should be targeted,
- Which products and services (what) add value to the chosen market and
- Which methods/processes (how) deliver these in a cost-efficient way.

Innovation, he argues, happens when gaps are created or discovered: i.e. a new WHO, a new WHAT or a new HOW, hence these three perspectives must first be established for WO to create a baseline understanding from an innovation perspective. This model maps neatly onto the *pull* concept (new customers/WHO's) or *push* concept (new products or methods/WHAT's and HOW's) and also to (Markides' 1998) perspective of driving change internally vs. waiting for the market to ask for it. Yet despite this seemingly simple model of innovation, push combined with market pull, there is evidence of technically sound innovations failing to be successful (Betamax video, Sony MiniDisk, Phillips LaserDisk in the consumer audio-visual market alone). (Adner 2012) points here to failures of considering benefits to the wider eco-system vs. local/individual benefits as underlying several well-known failures to adopt.

Within the context of this project, we might consider to what extent these criteria (new players, new services, new methods) are perceived to be in play and relevant to the potential users of WOs and the extent to which the technology is being pushed or pulled?

Much has been written concerning each of the concepts of ideation, development/refinement and diffusion but for the purposes of this research the focus will not be on *how* the Web Observatory was developed from the idea that inspired it. Instead, the focus is on factors affecting the level of adoption (especially within a wider eco-system of observatories) and to a lesser extent the development/refinement process insofar as this affects adoption. Are there features or applications of WOs that speak to particular patterns/models of adoption?

(Rogers 1995) model relies on a series of subjective decisions/perceptions leading to varying speeds of conversion by the various classes of adopter listed above following a 5-stage process/journey to adoption called the innovation-decision process namely:

1. **Knowledge** - the user learns of the existence and function of the innovation
2. **Persuasion** - the user forms a positive view of the benefits to themselves
3. **Decision** - the user resolved to take on the innovation
4. **Implementation** - the user commits the time/resources to adopt
5. **Confirmation** - the user is (un)able to realise the benefits

Rogers says diffusion can gain/lose momentum through influence factors with an interpersonal information network of recommenders/detractors within the social group. This may affect the perception of the idea and its innovative nature according to five evaluation criteria:

1. **Relative advantage** - how much improvement is offered by the new idea
2. **Compatibility** - the difficulty/ease of absorbing the innovation into current practices/systems
3. **Complexity/Simplicity** - the effort required to operate the innovation
4. **"Trialability"** - the level of cost/effort/commitment required to test an innovation
5. **Observability** - the level of visible effect to others in the group (positive/negative) - a communication factor

When 10-20% of the population have adopted the innovation (the early adopter group), strong mainstream adoption is thought to start which Rogers describes as the "heart of the diffusion process".

A further key concept in (Rogers 1995) is that of the *Technology Cluster* which is a group (from the users' perspective) of related technologies such that a halo effect (either positive or negative) relating to other technologies in the same cluster exists. The effects on WO adoption will be considered below in relation to other technologies with which it might be considered to be clustered. The author also states that delineation is important to determine the boundaries of the innovation being considered to avoid the effect of mixing the adoption of one technology with that of a related improvement.

This is particularly pertinent for WO as we must consider the innovation due to the Observatory itself and not that due to *any* system using the Web (i.e., the innovative nature of the Web itself)

(Ram & Jung 1994) refer to additional characteristics of early adopters in particular in terms of "use innovativeness" as a missing factor in Rogers. This speaks to the extent to which a user seeks to apply an innovation to a novel problem (the possible link here to factors contributing to disruption should be noted) hence the willingness to apply the new technology distinguishes the early adopter from the laggard. Understanding how different users apply WO to their problem space is thus important to map the likelihood of adoption and co-operative engagement.

In contrast to viewing adoption exclusively as a system of *positive* drivers (i.e., with resistance viewed as an imperfection) we may also consider innovation resistance to be the key process in adoption.

Innovation Resistance and User Behaviour

(Ram 1987) argues that innovation processes are not smooth with natural momentum which is occasionally impeded by imperfections, but rather they are *always* characterised by varying levels of resistance and hence specifically presents a model of innovation resistance. He criticises the "Rogerian" model in which (he feels) late adopters (laggards) are presented as a "bad thing" and that it does so with a narrow focus - looking only at successful innovations - something which implies that all innovations are a positive force and offer clear improvement. This, Ram argues, is not supported by the evidence of the high rate of new product failures. Resistance to change (particularly for its own sake) is not only a normal phenomenon (a homeostatic desire for equilibrium) but may be more relevant to the innovation process than the positive drive to adopt. He argues that innovation resistance should not be considered the opposite of innovation adoption since resistance/adoption co-exist in every diffusion process such that the resistance is either overcome over time (successful adoption) or not (product failure). Rogers later (1995) accepted this pro- innovation bias criticism of the diffusion model.

Ram expands on Rogers' original five stages of adoption and presents a more complex/differentiated model of resistance based on 26 parameters in three groupings.

- **Innovation** Characteristics - the nature of the idea/innovation
- **Consumer** Characteristics - the nature/perceptions of the adopter/market
- Characteristics of the **propagation** mechanism - the nature of how the perceptions are transmitted

This, he argues, will either lead to adoption, feedback for modification or rejection whilst (Klein & Sorra 1996) seek to differentiate additional outcomes beyond adoption versus resistance offering a spectrum model covering:

Resistance ←→Avoidance←→ Compliance←→ Commitment

Adoption, they argue, is often decided at the senior management level whereas the actual value of using the innovation is determined by the nature of the actions of ground-level users. Adoption in this model is thus not simply buying or even implementing an innovation; it is defined as the process of gaining the targeted, appropriate and committed use of an innovation. In this case, it is more often the failure of the implementation (Rogers Stage 4) rather than the failure of the underlying innovation that causes an innovation to be tried and rejected at the Confirmation Stage (Rogers Stage 5).

Success here lies in the success of the implementation and is based on fit along two axes:

- Innovation values fit (Poor, Neutral, Good)
- Implementation climate (Strong, Weak)

We may, therefore, consider not only perceived opportunities but also problems reported by WO users as highly significant and the social/political climate which exists for implementation and potential adoption.

Human Factors

Many of these approaches appear to find common roots in psychological theory such as the theory of reasoned action (TRA) by (Ajzen & Fishbein 1977)) and the simplified adoption theory that sprang from it known as the Technology Acceptance Model (TAM[15]) (Davis et al., 1989). This offers a model measuring perceived usefulness (PU) and perceived ease of use (PEOU) Figure 2-1 which then drives behaviour - in this case, adoption behaviour.

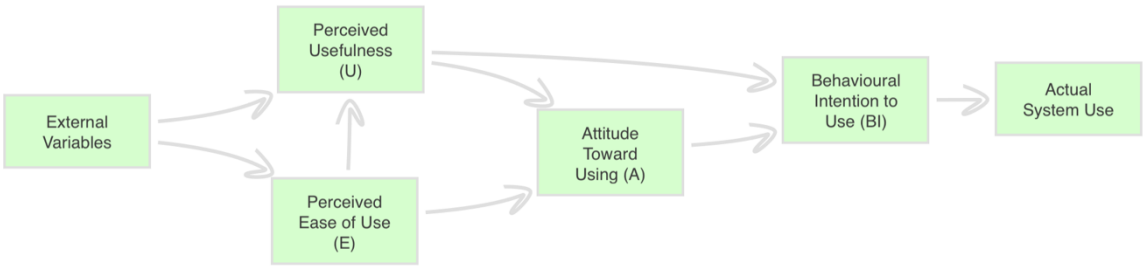


Figure 2-1Technology Adoption Model[16].

Hence human factors become apparent beyond an analysis of the nature of the innovation itself and are instead cognitive judgements become relevant. We will, therefore, consider cognitive perceptions (framing) of WO by users in order to inform the natures of drivers/blocks in the adoption of WO.

A review of the models above suggests that the process might be summarised as a continuum between perceived benefit and perceived costs (difficulties/risks). The adopters ability to understand and value the relative upsides/downsides and the ability and opportunity to communicate these capture the differences between innovations which are adopted and those which aren't. More accurately, it models the competing forces, the human behavioural/cognitive

[15] Updated in (Venkash et al. 2003) as UTAUT
[16] http://en.wikipedia.org/wiki/Technology_acceptance_model

aspects and hence the net direction/speed of the journey towards adoption or continued resistance.

Potential weakness of these models lies in that

- The output of the process (use/no use) should perhaps be linked to the input stage (external variables) since visibility of choices by others is often thought to be influential
- Based on the weighting of factors and the inclusion of what are considered to be relevant external factors complex models may be no more accurate/predictive than simpler (parsimonious) models such as LUM (Lazy User Model[17]) which simply predict users will select the lowest effort (cost) option for a given outcome.
- Innovation is seen here as a homogenous concept without reference to different *types* of innovation or differences in the way the innovation may impact the adopters and the rest of the market.

In the next section, the impacts of adoption in terms of evolution/extension vs. disruption will be discussed as a context from which to understand the eco-system that WO may be thought to be perceived as an enabler or a threat.

2.4 Disruption

With the creation of digital products and services and their supporting networks comes the ability to share data instantly, globally and non-consumptively (i.e., an effectively unlimited supply of perfect digital copies at virtually zero marginal cost of production, storage and distribution). This bypasses not only the original limitations/economics of the historical hand-copied libraries but also, as the tools of production/publication are 'democratised' [18], the flow of painstakingly copied and curated facts may become a torrent of unfiltered opinion. This hypermobility of data may remove some of the positive checks and balances of review, quality and provenance since this data can also be shared automatically, anonymously and reported vicariously with little distinction between official/unofficial sources.

Search engines have evolved to pre-filter/pre-select what we see from a vast range of possible search 'hits' and this process shapes and reinforces our understanding and perceptions through what has been called the 'Filter Bubble' (Pariser 2012). Thus the appearance of fake news,

[17] (Collan & Tetard 2007, 2009)
[18] i.e., become widely/cheaply available (Anderson 2010)

alternative facts[19] and the filtering/control of sources of information (selective blocking of internet sites) remains a concern globally. Tesich coined the term 'post-truth' with regard to political reporting of the Gulf war and the rise of 'post-factual politics' is currently in sharp focus following 'Brexit' and the 2016 US presidential election.

Certainly, the introduction of digital has had profound impacts on traditional economic/legal models, often disrupting concepts such as 'location', 'jurisdiction', 'ownership' and 'inventory' and deeply transforming long-established notions/limitations of physical exchanges.

The disruption of Blainer's [20] "tyranny of distance" – typically seen as a positive thing - may, therefore, come with several other related outcomes in a Web market such as:

- Open access vs. theft?: for consumers/owners of intellectual property
    - (i.e., disrupting the 'tyranny' of ownership)
- Free speech vs. hate speech?: for media/news operators
    - (i.e., disrupting the 'tyranny' of oversight)

Each will be greeted or resisted depending on the perspective of the affected person. The Web grants equal access/abilities to all participants regardless of social status, merit or trustworthiness noting that there may be technical, political, educational and economic barriers in practice.

The nature of disruption of- and by "Web-like systems" (Berners-Lee et al., 2006a) is a key theme for Web Science. Its goal is to understand how the constant change in content, capability, capacity and structure of these systems impacts (and is impacted by) behaviour, culture and choice. The authors comment that these systems may demonstrate unplanned (so-called *emergent*) effects when operating at Web-scale and that these present both opportunities and risks for the Web.

It has variously been argued that:

- Technological elements such as convergence/innovation are purely "pulled" (invited) by social requirements (Pinch & Bijker 1987) and (Sismondo 2011)
- Behaviours/opportunities are purely "pushed" (imposed) by the development/convergence of technology (technological determinism) (Ellul 1954) and (Schumpeter 1935) noting that (1) and (2) are incompatible explanations or
- Some 'compatibilist' hybrid of the two such as 'soft' determinism; describing the reaction of society to technology over time (invention, accumulation, diffusion and adjustment)

[19] With point of view/opinion elevated to the credibility status of an observed fact.
[20] Blainer was talking about the impact of vast distances and the resulting separation on Australian culture

expressed as a 'social lag' (Ogburn 1922) which has a similar structure to (Rogers 1995) theory of innovation diffusion.

Kranzberg characterises thus:

"Technology is neither good nor bad; nor is it neutral."

(Kranzberg 1986)

In this section, we have reviewed the importance of perspective on the production and consumption of data and how WO may be perceived. Whilst the Web itself has undoubtedly been a disruptive innovation it remains to be seen how/if WOs will affect established research methods. Given WOs' inclusive, collaborative objectives can it succeed as a *disruptive* innovation in the strict sense or must it seek to be *evolutionary* and inclusive?

In the disruption process the importance of basic human desires for survival, enhancement and improvement become relevant, and thus theories of motivation and incentive remain key to understanding the nature of a socio-technical Web. Such motivations/behaviours, however, exist not only in isolation but as 'net' decisions/behaviours in group contexts between networks of people who learn/adapt in group contexts and may act en masse.

2.5 Networks: Machines, Social Production, Sharing and Culture

Both positive and negative network externalities (the effect that one person's ownership or use of a product can have on the value of the service to another person) have been documented since the development of the early telephone networks (Vail at Bell Labs). In a Web context, the benefits of the implementation of additional network nodes typically refer to Metcalfe's Law, and indeed the benefits[21] accruing to a successive larger eco-system of WOs sharing their expertise, content and analytics would be expected to exceed the usefulness of any individual member WO.

The possibility of benefits, however, is not a guarantee of benefit.

e.g., simply because users are *technically able* to share content and services does not imply that they will do so without some *utility/incentive*. Often social connections running in parallel to technical connections may be relevant.

[21] Though the precise magnitude of the effect is debated.

Models including social elements/incentives should, therefore, offer better insight into the way in which shared platforms operate/interoperate versus the naïve assumption that "if you build it – they (sic) will come ".

This view is borne out in (Hendler & Golbeck 2007), which points out the requirements for network effects to apply to the Web. There is a need, not only for more Web users or more Web pages (content) or potential linkages, but for actual technical linkages between these users (a social graph) and links between content/pages for Metcalfe to apply.

Hendler goes on to characterise Web 2.0-type social networking systems as being "highly social but relatively link poor" whereas Web 3.0 systems (Semantic Web systems) as being relatively "link-rich but less social". The authors make the important characterisation of the social graph as the vector for content sharing rather than the nature of the content alone and attributes the successes of many Web 2.0 systems to this social vector (rather than tagging/folksonomies as has been claimed).

The implication for sharing Web data is that sociality, common interests/motivations and "WO social graphs" may be highly relevant to adoption beyond the technical enablement of sharing through discovery and standards. While we should not refrain from the use of more formal knowledge descriptions, we should also be seeking to enrich them through an understanding of social channels.

WOs may need to natively support (or at least not prevent/distort) such social communication and collaborative production. Thus, the *social design* of WOs may be as important as the technical design.

The power of social "production" has been characterised in "The Wealth of Networks", (Benkler 2006) citing open systems, social production and the transformative nature of sharing and openness on traditional market and pricing mechanisms. Benkler observes there has not been a discrete transition from industrial models to an information age but rather a stepped transition from an industrial age to an industrial-information age, whereby the means of information production are controlled and channelled through the same means as for production in the pure industrial model. This, Benkler argues, is inappropriate, inefficient and no longer necessary.

He characterises the existing model as a 4x4 matrix (Table 2-1) of centralised vs. decentralised facilities that are driven by Market vs. non-Market motives.

	Market-Based	*Non-Market*
Decentralised	Price System	Social Sharing & Exchange
Centralised	Firm Hierarchy	Government & Nonprofits

Table 2-1 Transactional framework adapted from (Benkler 2006)

We see market-based effects in the creation of large web data oligopolies such as Google, Facebook et al. yet are these *interim* effects or more permanent subsuming, dominant commercial pressures on data collection/management leaving the WO representing a potential bridge between commercial and non-commercial social sharing?

With nearly 3 billion people (2014 figures) able to engage in social production via the Web the means for information production (previously reserved for large industry and government) has been delivered to a global audience. Benkler observes that this non-Market production constitutes a threat to the industrial-information economy and that a battle for control of the ecology of the digital environment (between open and closed approaches) has resulted. This exposes a potential conflict between commercial/non-commercial WOs and the quality/utility of open vs. proprietary data/services.

In a Web-of-Data context nothing precludes individual systems/services being commercial, however, the broader notion of a shared network of Web Observatories specifically comprises "that which is openly available" and so for the development of WO, the importance of such trends and network effects is underlined. Benkler's models of information economies and transactional systems show clear parallels to the eco-system that a global WO eco-system would require. Individual WOs might operate as open source tools across both market and non-market sectors.

There are significant implications for commercial WO services and "for-profit" groups seeking to operate WO systems as they relate to the interoperation with not-for-profit or open WOs in the wider ecosystem. Considering the emergence of interactions between these alternative approaches will be important in understanding how WOs may function.

2.5.1 WWW gets more sources, more volume, more contexts

"Weaving the Web", (Berners-Lee & Fischetti 1999) recounts the original vision of the Web in terms of three key capabilities:

- Global hyperlinked documents
- User-generated content
- Actionable semantic data.

This relates to the Web-of-Pages, Web-of-People and Web-of-Data introduced above - see also (Hall & Tiropanis 2012). Even though all three of these concepts were defined and technically supportable from the outset (e.g., URL links to pages on other servers, user-editable pages in Berners-Lee's original browser and URIs supporting links to datasets rather than Web pages), the widespread adoption of each these concepts is separated by decades.

The reasons for the success of WWW versus earlier curated networks are debatable. It may simply be that WWW was free-to-use at a time when services such as CompuServe and AOL were charging fees. It has also been suggested that it stems from being unregulated, lightweight, based on open and free standards and because it has a *tolerance* for missing links and broken servers. Berners-Lee acknowledges the Web is "messy" though if messiness/ambiguity is a significant factor, this raises an important issue for the "Web of Data". The tolerance for ambiguity and "messiness" is significantly lower for datasets versus Web pages, requiring not only adherence to technical standards but also unambiguous agreement on the naming of items and the definitions of meanings. In effect, a markedly more complex (mature) technology stack is required than for Web 1.0 or Web 2.0. If users are to share and combine datasets from multiple (previously unknown) sources, the key challenges of discovery, format, trust and liability must be addressed.

2.5.2 WO and the Semantic Web

(Berners-Lee 1998b) addresses these challenges when laying out a first architectural vision of the representation and processing of Semantic Web data. This is a brief working paper with short descriptions of several layers covering the basic requirements of data representation and additional layers, which, Berners-Lee proposed, would allow for extensions, conversions of data, signatures/trust in data as well inferences about the meaning.:

In a Scientific American paper, (Berners-Lee et al., 2001) describes a model of what could be possible in a world of intelligent agents powered by Semantic Web data with the authors painting a futuristic picture that business and non-specialists can understand and frame within their own context.

In a private conversation in 2012 however, Hendler commented to me that, in retrospect, he believed that he and Berners-Lee had "significantly underestimated the complexity of implementing the Semantic Web". Additional social/human factors play out in technology adoption beyond the availability of technical standards, and these, he felt, were perhaps not as aligned for the adoption of the Semantic Web as for the Web of Pages despite significant progress having been made in areas of Semantic Web.

The Semantic Web literature acknowledges the lack of a *killer app* and given the true potential of a global WO may require many of the protocol and metadata enhancements which underpin semantic data it is interesting to consider if WO in its developed form may itself be considered the missing killer app.

In a 2014 lecture in Montreal Hendler explores this point claiming it has recently become clear that, while not directly accessible to Web Users, search engine technology (such as Bing and Google with KnowledgeGraph and Hummingbird) are increasingly using semantic and linked data representations to enhance and contextualise search. In fact, Hendler claims, the Semantic Web is to some extent already upon us though perhaps not yet in the form portrayed in the 2001 Scientific American cameo.

As the Web has grown in size/complexity in terms of participants (human and machines), locations (fixed location, mobile/apps) and content (sources, formats, volume and interfaces) three notable trends have emerged:

- A trend towards "bigger" Data: more volume, velocity, and variability in both the sources and content/structure (both data and metadata)
- A trend towards "broader" data - (Hendler 2013) combining data from multiple diverse sources
- A trend towards unintended (emergent) properties and behaviours indicating "sociotechnical effects" on the Web (the interaction between the content and the systems/structure) (Shadbolt et al., 2013) and (O'Hara et al., 2013).

This underscores the importance of understanding and integrating datasets in a Web-of-Data while providing no clear solution to the easy semantic tools problem suggested by Hendler and Berners-Lee. The extent to which these growing sources of data support linkage, curation and automated actions is relevant for Web Science and Web Observatories given:

- The tolerance for gaps, errors and ambiguity is far lower for an automatically processed Web-of-Data than for a human-decoded Web-of-Pages

- The current message "Error 401 - Page not found" is cited by Berners-Lee as being instrumental in the success of the Web and has different implications for recovery from a failed Semantic Web equivalent query (e.g., "Error 4xx – Context not understood")
- The lack of accessible user level tools to wrangle semantic linked data
- The lack of a compelling killer app for the broad base of users.

Whatever the standard/approach adopted, for a Web-of-Data to be interpretable and even actionable, the meaning of the data must either be included/explicit, referenced (obtainable) or deducible with significant repercussions for the design of shared data repositories. This may be less troublesome for internal datasets where the provenance/meaning is likely to be clearer but may be particularly difficult for externally-sourced shared datasets where there may be multiple formats, topics, licenses. Combining these would represent a significant technical challenge without a suitable guide.

Capturing the technical nature of the WO allows us to consider individual WO instances but in order to define an architecture or framework in which multiple WOs might interact requires an understanding of the drivers and motivations for their existence and operation. Theoretical social frameworks and the concept of Social Machines will, therefore, be considered below both from a general viewpoint and as a theoretical lens.

2.6 Social Machines

2.6.1 People and machines working on the Web

> "In the working of every system, there is a wheel within a wheel, which, according to its position, aids or counteract the ends proposed to be accomplished. Thus, the genius of modern invention has applied simple contrivances of self-adjustment to the most complicated machine .. (should we not seek) a nobler object? .. the self-adjustment and perfect working of the Social Machine."
>
> **From "Mutual Improvement, Or, A Scheme for the Self-adjustment of the Social Machine" (Allen 1846)**

Writing 150 years before the Web, there is no question that Allen can be referencing Berners-Lee's combination of human input, actionable data and computation to solve problems in society. Rather, Allen is talking about the complex machinery of society itself and yet the idea of processes that can be engineered and regulated, and that would induce the crowd to solve societal problems, run through both the historical and modern versions of the term "Social Machine".

The term itself is ambiguous in the wider literature, and a choice of a definitions/terms is critical here. The discussion on what constitutes a Social Machine is broad, on-going and, at times, heated. In the sense intended here, these are not simply the purely technological Social Machines (socially embedded machines) of (Rousch 2005) "Computing means connecting" nor the ethereal philosophical Social Machines of (Deleuze & Guattari 1983) - they are something in-between and, as such, complex to analyse.

As an example of the lack of common understanding, (Meira et al., 2011) (citing Rousch who uses "Social Machine" in only a limited sense to mean *connected* technologies) proposes a mathematical/algebraic model to represent both the concepts and programming of Social Machines and asserts (inconsistently with the Berners-Lee definition) that the cited example (Futweet) is a:

> "real example of a Social Machine **because** (*my emphasis - Ed.*) it is designed and built to be networked with other applications".

I note that the proposed model has no mention of human input (despite citing Berners-Lee) and thus arguably misses the point of the Berners-Lee model in which humans and machines are collaborating rather than machines/agents collaborating without input/direction. Something (Shadbolt et al., 2013) describes as "large-scale interaction of humans with machines".

Berners-Lee 's Social Machine is embedded in societal rather than technical challenges:

> "Real life is, and must be, full of all kinds of social constraint – the very processes from which society arises. Computers can help if we use them to create abstract Social Machines on the Web: processes in which the people do the creative work, and the machine does the administration.. The stage is set for an evolutionary growth of new social engines. The ability to create new forms of social process would be given to the world at large, and development would be rapid."
>
> **(Berners-Lee 1999)**

Berners-Lee is not talking about random convergence/usage, but rather *engineered* systems (albeit "evolving" ones). This is tied to the debate between technological determinism (where the machines provide a basis for the solution to the social constraint: technology "pushing" behaviour) and socially-constructed technology where Berners-Lee's social constraint gives life/meaning to the technology: society "pulling" technology.

This has implications for the study of WO since it has been claimed that Web Observatories are themselves a type of Social Machine as well being a tool for *observing* Social Machines. We will return in later chapters to review this claim based on four criteria:

- That WO has discernible technical and social components, i.e., is not only socially-embedded (generic requirement)
- That this combination produces something that neither element alone produces, i.e., neither component is incidental (generic requirement)
- That it addresses social constraints through a combination of human input and machine administration on the Web (Berners-Lee & Fischetti 1999)
- That this process represents a large-scale interaction of humans with machines (Shadbolt et al., 2013)

Given substantial changes in the size/complexity of the Web itself in the intervening years, the Berners-Lee Social Machine definition may also need to evolve and hence we may need to retain some flexibility to account for this.

2.6.2 Social Machine perspectives in the literature

Whilst Social Machines and WOs are related concepts as described above, the Social Machines is a much broader concept and overlaps with concepts such as human computation, social computing, collective intelligence and crowdsourcing (Shadbolt et al., 2013). The literature on Social Machines is divided into three main areas: technical aspects, identification/classification and social dimensions.

The focus on Social Machines (in the Berners-Lee sense) starts in 2010 with a paper which bridges the debates in a series of papers on semantic web to the example of Social Machines. Hendler and Berners-Lee (Hendler & Berners-Lee 2010) argue that the key to the evolution of Social Machines capable of dealing with issues of privacy, provenance and policy is the maturation of semantic web technologies which otherwise limits the development of Social Machines to primitive examples incapable of reasoning logically over the data they store. Whilst the authors fully admit (and even stress) the importance of human factors, this is primarily a paper about technical solutions and standards.

(Tinati & Carr 2012) focuses on the need for a socio-technical balance in the methodology to understand Social Machines and in some ways is attempting to redress a balance in which much of the focus in Web Science has been on the machine. They argue methods have not kept up with the development of the Web. A mixed methods approach based on Actor Network Theory (ANT)

is sketched out here but not described in detail until the follow-up paper (Tinati et al., 2013) which presents a refined version of the approach (called HTP) and presents a worked example analysis of Wikipedia. This is a valuable contribution though it is unclear to what extent practitioners need expertise in Actor Network Theory to utilise the approach effectively.

In terms of identifying and classifying Social Machines, many uses and examples can be found in the recent literature including (O'Hara 2013) (politics), (Van Kleek et al., 2013) (health), (Evans et al., 2013) (crime), (Martin & Pease 2013) (scholarship), (Van Kleek et al., 2014) (personal data) and (Tiropanis et al., 2014) (government). These themes are reflected in the structure of the investigation by the recent project: SOCIAM - theory and practise of Social Machines. In this case, the SOCIAM project itself is employing a WO approach (The *Macroscope*) to study Social Machines.

Two further papers on structure/taxonomy for Social Machines (Shadbolt et al., 2013) and (Smart et al., 2014) are methodologically relevant for this project in terms of the constructs/dimensions chosen to model the Social Machines. Their initial groupings are attributes concerning "contributions" (the purpose/output), attributes of the "participants", and "motivations". As a critique of this analysis, I note that the definition offered of "what the participants do" is instead a less differentiated mixture of "what", "how" and "why" relating to both the task and the participant. The motivational model is sourced only from literature on social networks and potentially risks missing insights from more general models of motivation.

The authors offer a revised definition of Social Machines from that of Berners-Lee:

> "Social Machines are Web-based socio-technical systems in which the human and technological elements play the role of participant machinery with respect to the mechanistic realisation of system-level processes."
>
> **(Shadbolt et al., 2013)**

To some extent, this definition potentially omits a number of critical aspects of Social Machines including the self-determining (non-Turing) aspect as well as the importance of the sociality "at scale". (Smart et al., 2014) concedes this but does not offer a revised definition. Based on this, the characterisation I offer here of what constitutes a Social Machine is a pragmatic one and in part fuelled from the observation that overly inclusive definitions are, at least as unhelpful as overly restrictive ones.

e.g., "*Could* a web page be a Social Machine?" is perhaps more useful than "Is *every* web page a Social Machine?". The distinction I offer is simple (though hopefully not simplistic) and is as follows:

Actors	*Constitutes*
Person + (any) Technology → generic Solution	Use of a Tool
Person + Tool via HTTP → Web Solution	Use of a Web App (where "Web" may also be non-traditional browser/mobile app)
Person(s) + Web App 'at scale' → Shared Solution	Use of a (collaborative) Web Platform/Service
Person(s) + Web Platform → Solution co-determined by people and the Platform	Use of a Social Machine

Table 2-2 Machines vs. Social Machines

Thus I submit that solutions:

- Missing human computation and/or computer-supported collaboration
- Lacking distributed 'web-like' apps / access
- Not exhibiting network effects from sociality 'at scale.'

whilst potentially significant or valuable are unlikely to be Social Machines.

Non-Web Social Machines (such as clocks or shared agricultural tools) also appear in the literature; they are however not the focus of this research. Whilst an overly restrictive definition for Social Machines may be unhelpful an extended definition is implied here:

> "Social Machines are Web-based socio-technical systems in which the human and technological elements play the role of participant machinery with respect to the mechanistic realisation of system-level processes and benefit from network effects from

> use at Web scale to address socially-constructed challenges in terms of these processes".
>
> **Brown adapted from (Shadbolt et al., 2013)**

The final theme is oriented more towards an understanding of the "social" in the Social Machine and comprises an arc of papers resting on the premise that *purpose* is key for the definition of Social Machines i.e. that a Social Machine must be *for* something (or many things) and tracking these purposes/goals (akin to Berners-Lee's definition) the requirement is asserted to be:

> "a purposefully designed sociotechnical system comprising machines and people"
>
> **(De Roure et al., 2013)**

thus enabling a useful insight into the machine itself. I find this to be more flexible/nuanced than theories of individual motivational theory or broader socially-shaped behaviours alone and influences later work on narratives (N) in the DNA model. This approach steps away from considering people and technology as layers in the Social Machine (Bowker says one cannot simply put people "on top of" technology) but rather the Social Machines span various technical and social components and are *involved* in Social Machines. The analogy that "music" is not the same thing as either the instrument or the player may apply here - the Social Machine is not simply the addition of users to the machine but rather emerges from the interplay between the elements towards an end or purpose.

The element that emerges as intention/purpose is refined into Trajectories in (Page & De Roure 2013) which collects a series of objectives and paths potentially through different physical platforms (as individual trajectories) to form an understanding of the overall Social Machine interaction. The lifecycle is also flagged as being of interest - thus implying that objectives and behaviour changes over time are both expected and relevant.

The key insights from this and (Tarte et al., 2014) is that within a single interoperating Social Machine the individual actors may have their own differing agendas and stories leading to much more complex behaviours than might be explained by a single (apparently shared) or rigidly enforced single objective. The narrative approach suggested uses lifecycles and plot points and, it is argued, this may not only help researchers to understand the individual stories of the participants but also characterise and understand the health and functioning of the Social Machine from a combined perspective.

This group of papers appears aligned with (Smart et al., 2014) in pointing out that to say Twitter (or Facebook) "is a Social Machine" may simply be shorthand which should more correctly be

expressed as "Social Machines are created" when Facebook/Twitter users interact via the platform for some purpose. The point made here is that different Social Machines may emerge at different sizes and timescales and so we take from this that these elements are *part* of a larger Social Machine thus avoiding the potential danger to drop to such a low level of granularity that each/every hashtag, re-post or like is thought to *intrinsically* be a Social Machine its own right – thus requiring a new definition of hierarchical Social Machine collections through which to group the smaller machines. Hence it should be considered that the simpler shorthand (though perhaps inaccurate) may be more pragmatically useful. The authors highlight key features to identify around the Social Machine eco-system including:

- The constituent machines (as defined by intent)
- The actors (human and technological)
- The design (signification) process
- The ground rules as defined or evolved

And this approach has informed the narrative elements of the model developed Ch4-8.

WO as a Social Machine

As mentioned above the WO is not only used to study Social Machines but is, itself alleged to be a (scholarly/research) Social Machine through which researchers may collaborate, exchanging access to data, tools, methods and papers. It is therefore of interest to consider how the study of WO as a Social Machine may be of value.

It should be noted (after (Smart et al., 2014) that we may, instead, be claiming that Social Machines are/can be created "on WOs" rather than claiming that WOs are *inherently* Social Machines.

In terms of collaborative models, work at MIT (Woolley et al., 2010) demonstrated not only that collective intelligence[182] (c) can be shown to exceed that of the general intelligence of the individual (known as g) but that the group can also comprise both human and machine actors. This appears to show potential benefits of collaborating via/with a Social Machine.

It suggests potential supplementary research questions around the type and configuration of systems that mediate this collaboration. Research indicating superior results for man/machine problem-solving than for either human or machine groups alone (using prediction markets) shows the way for examining how machine actors in Web Observatory systems might actively contribute to the performance of this class of system.

Malone's MIT group is however only working with traditional collaboration systems (e.g., Co-Lab[22]) to mediate crowd-sourced solutions for complex social problems such as climate change and are focussed on incentive engineering for participation. No research is currently published on the use of machine actors in such collaborative systems.

Interestingly, the most significant predictor of collective intelligence in Wooley's experiments was "Social Intelligence" (the ability to perceive and react to others in the group) not only for face-to-face groups but also for groups participating on-line and therefore not interacting visually. Malone theorises that social intelligence is perhaps closely correlated with other types of general empathic/interpersonal skills.

The relevance for Web Observatory is that this work implies that the design of systems that enable/mediate human/human collaboration or human/machine collaboration may have a direct effect on the collaborative intelligence of that system. Hence, we should consider the design of Web Observatory systems from a participative/collaborative point of view if we are to enjoy the enhancements that this research suggests may be available for hybrid group collective intelligence.

The challenge remains for the WO to reach a W^3O community at sufficient scale to qualify as a Social Machine. Berners-Lee's definition does not appear to describe small, local groups but rather the way in which societies can interact with technology at scale to address "societal problems". Whilst there are no hard/fast limits on the number of participants required, I submit that many thousands of users collaborating using/re-using millions of entries in Wikipedia (or classifications on a Citizen Science platform) are more readily recognisable as Social Machine interactions than, say, a small research group developing/sharing a dataset for a single purpose defined (Table 2-2) as being a collaboration interaction via a web app.

Having considered the evolution of technologies for data and the growth of the Web as a medium for social interaction, in the next section we will consider Web Science as it aims to understand the Web itself and Society-through-the-Web. This requires finding/creating datasets that can be shared/understood across disciplines/locations to gain insight from diverse global sources.

[22] https://climatecolab.org/

2.6.3 Social Machines and Socio-technical effects

Socio-technical effects are observable in the interplay between people and the technology they use. The nature of the interaction can be described through a variety of different models including Actor Network Theory (ANT) (Latour, Callon, Law), Social shaping (Vygotsky), Normalisation Process Theory (May), and Technological determinism (attributed to Veblen).

These models argue variously:

- That all behaviours and knowledge are direct expressions of the technologies developed/adopted in society (Smith & Marx 1994)
- That all knowledge is constructed and shared through social interaction with others (Vygotsky 1978)
- That new behaviours and technologies become embedded in social contexts through individual and collective agency (May et al., 2009)
- That all technologies are entirely socially constructed through behaviours and social factors (Latour 2005)

Since the WO is not engaged in the surveillance of individual users on the Web and direct contact with them to gather qualitative data to cross-check motivations can only be on a sampled basis (for reasons of scale) we are faced with the challenge of how to capture, analyse, describe and ultimately predict these socio-technical effects. If we are to capture 'footprints in the sand' from which to learn something about the actors who left these footprints, then this must break down into three component parts:

- A technical (machine) analysis - considering the Web as a piece of engineering and how its mathematical properties and network features enable or encourage certain behaviours
- A social analysis - considering how behaviours emerges as a consequence of human perceptions and motivations
- A socio-technical analysis in which the interplay between human and software agents results in some set of behaviours that are distinct from those that might be expected without the interplay.

These elements will inform perspectives in both research questions and research methods (Ch3)

The persistence of opposing views on the primacy of social or technical effects in the literature suggests that our analysis of WO should not consider one element to the exclusion of the other in the process of understanding how WOs and their users may interact.

2.7 A Science of the Web

First proposed in a short piece in Science (Berners-Lee et al., 2006b) it was suggested that a science-**of**-the-Web (distinct from doing science-**on**-the-web) was needed to explain the rich way in which society interacts with the content and structure of the Web. It suggests that Web Science must go beyond describing the current structures and processes and should be:

> ".. about engineering new infrastructure protocols and understanding the society that uses them, and it is about the creation of beneficial new systems".
>
> **(Berners-Lee et al., 2006b)**

In a longer piece (Berners-Lee et al,. 2006) cover the technical and, to a lesser extent, the social elements of the new proposed discipline of Web Science. The authors mix a discussion of graph theory, semantic technology, governance and social theories in a novel way and call for interdisciplinary collaboration to understand both elements of the social-plus-technical system that is the Web.

(Schneiderman 2007) re-iterates how Web Science views the Web as being socially embedded, therefore requiring an understanding of issues like trust, privacy and provenance. He proposes an even wider brief for Web Science calling for explanatory and predictive theories across a wide range of on-line processes. (Halford et al., 2010) echoes this call for a social scientific theory offering a "manifesto for Web Science" based on the four key elements of:

6. The co-constitution (mutual shaping) of technology
7. The heterogeneous (both human and non-human) nature of the actors (after Latour, Contractor)
8. The significance of *performativity* (the enactment/process) of the Web vs. the structure or physicality alone
9. Immutable Mobiles (after Latour) which are temporary stabilisations of actors, technologies and practices giving rise to the perception of a stable entity, e.g., the Web

This model provides a rich basis for an analysis of WO and influences elements of the DNA model which emerges from the research.

In a point of key interest to understanding cooperation/sharing Contractor emphasises the need to understand human factors:

> "..as developments in information and communication technologies continue to reduce or eliminate the potential logistic barriers to our communication and knowledge

> networks, it becomes increasingly important to identify the various social factors that enable or constrain the development of these network linkages."
>
> **(Contractor 2009)**

Contractor includes an array of non-human agents (documents, data sets, analytic tools and concepts/keywords) that may play a part in a social network interaction (citing the work of Monge, Castells and Latour). The critical nature of sociotechnical interactions are discussed in (Hendler et al., 2008) and the need for approaches to study a system which is changing "faster than it can be observed" is stressed. Emergent properties and social responsibility, are flagged as essential elements while in (Shadbolt & Berners-Lee 2008) the need to understand not only *engineering* structures (such as scale-free networks) but equally structures of *incentives* is stressed. Relevant for this research is the claim that Web Science analysis must then surely include the role of people as well as technologies and seek to place the use of the Web within a social model/context including the social dimension of the research itself.

Again the need to spread beyond descriptions of WOs as artefacts is suggested here and to include a broader contextualised description of WOs as tools (for people) and solutions (to problems).

(Hall et al., 2009) revisits technical themes from the original 2006 Framework monograph and goes on to focus on the transition from large, data-centric, collaborative research frameworks in eScience and Grid Computing (Hey & Trefethen 2002), (De Roure et al., 2005) to the "long tail" of research content creation and sharing through Web technologies (such as open wetware in Biosciences). Hall & De Roure characterise the Web as a pervasive, collaborative research platform which empowers researchers who benefit from the ease of content creation and network effects from the scale of digital science. This is reminiscent of Berners-Lee's term "messy" (they call it "better not perfect") and is closely allied to the Web 2.0 design patterns (O'Reilly 2005). The paper goes on to consider MyExperiment (De Roure et al., 2007) as an exemplar of web-based research platforms using compound research objects and workflows with a view to curation and re-use. Just as Web 2.0 technologies have resulted in Web 2.0 business models and tools, this paper predicts a further co-evolution between (semantic) technologies that will develop and opportunities for Web-based platforms for research. While this paper focuses on research-*on*-the-Web rather than specifically research-*about*-the-Web, it does not preclude it and raises many of the questions of curation, reuse, trust, linking methods with results/data, provenance and practicality that are faced by Web Science research. The Web is critically not only an object of study but also a means of data collection and a platform to conduct studies and share the results. Each of these elements might be reflected in tools for Web Science.

(Hall & Tiropanis 2012) discusses the evolution of networks according to the Web-of-Documents, the Web-of-People and the Web-of-Data model and outlines network effects and the enhancement of these effects through add-on services such as search and recommender systems. In each case, the nodes and edges of the network are re-conceptualised to give different views of how the Web might be viewed. This insight is particularly appropriate since I will argue that Observatories are conceptualised differently by different groups of users under different models of operation and motivation. As the authors contend, these models may again be viewed and treated differently by more/less experienced users offering a further dimension/possibility that certain collaborative Web systems may require time, not only for their technology to mature but also for their audience/user base to do so. Thus we may need to consider if all groups of users are equally mature and engaged with Web Science concepts.

Hall and Tiropanis offer three broad guidelines based on the success of the Web:

1. Big is beautiful (the importance of scale)
2. 'Good enough' works (the importance of pragmatism/incrementalism)
3. Openness rules (the importance of accessibility)

This offers a perspective on the types of data that might be preferred/practical (from 2), for the level of automation and semantic reasoning that may be required - at least at initially - (from 3) and for the desire to build for sharing and wide deployment (from 1). The paper makes a valuable contribution by identifying key research insights, new fora and communities of practice such as the Web Science Trust, WSTNet and ACM Web Science Conference.

WOs and Social Machines are given prominence in a revision/extension to the original "Creating a Science of the Web" in (O'Hara et al., 2013) which offers a review of the changes in focus over the intervening period. Key themes shift from engineering and technical aspects to a study of emergence, a stronger focus on personal data and social networks and the social graph as well key attributes such as influence and trust.

There is recognition/proposal that a complex interacting system may require a research model that supports such complexity and five key perspectives are offered:

1. Computational
2. Mathematical
3. Social
4. Economic
5. Legal/regulatory.

This model borrows from earlier work by Contractor and puts forward MTML (multi-theory, multi-level) analytical framework (Contractor et al., 2006) and (Monge & Contractor 2001) as an approach and informs this research in terms of non-technical frameworks to consider the WO.

The implication for this research is the conclusion that multiple perspectives are required to construct a more realistic '3D' model of complex socially-embedded systems than can be extracted from purely technical elements.

If to understand the Web we must effectively understand agents, content, structures and social behaviour as it is played out on the Web, then suitable systems and proxies for modelling and analysis must be developed beyond the current level of tools for Web pages. New Virtual Observatories (VOs) on/about the Web (WOs) have been proposed as a solution.

2.8 Virtual Observatories

The Web has enabled a level of transparency, sharing and collaboration that would have been unthinkable for scholars of only 25 years ago and this has resulted in the creation of distributed research tools known as "Virtual Observatories":

> "A Virtual Observatory (VO) is a collection of interoperating data archives and software tools which utilise the internet to form a scientific research environment in which [astronomical] research programmes can be conducted."
>
> **Source: Wikipedia**

(Keahey 2012) focuses on the intersection of cloud and high-performance computing and describes their VO as a 'Laboratory at large' using a publish/subscribe system for streams of research data coming from ubiquitous, cheap, flexible sensor networks that can be switched in/out as required and accessed over fast networks for use in large-scale instrumentation. Increased access to high-volume, high-resolution (though not necessarily "big") data requires researchers to leverage high capacity computing power which can include on-demand analytics provided through a cloud infrastructure.

The model put forward has the feel of a national power grid in which flows and resources are switched in/out as required and also offer the ability to publish algorithms and experimental workflows for re-use such as those proposed by (Burnap et al., 2014) and (De Roure et al., 2007).

The idea of marshalling information, people and tools for enhanced research effectiveness and the elimination of duplication and waste are visible across several disciplines: Astronomy and Life Sciences (e.g. DCO, iHub and IVOA). A Web Science Observatory might assume the same types of

benefits to be available for the interdisciplinary research community to create new repositories and/or join with these other natural science Observatories and social science Observatories.

The International Virtual Observatory Alliance (IVOA) and their Astronomy VO project is a highly significant precursor/template for WO and addresses the problem of slow, difficult technical access to scarce resources and complex astronomical datasets. It seemed reasonable that, rather than waiting for access to capture data from a particular Observatory that might already exist elsewhere, the dataset could be accessed from another site to build up richer, composite sets and avoid costs/delays in gathering/sharing data. The prerequisite for this was the ability to agree on standards to describe and share the datasets and APIs to access the data physically.

The VO archive comprises a large corpus of papers produced under the auspices of the VO project internationally. Whilst a systematic review is impractical, analysis of samples reveals key themes that are strongly repeated (*saturated* in Grounded Theory terms):

- Standard data formats for processing and publishing
- Distributed queries
- Web portals to enable interaction and configuration
- Web portals/tools for visualisation of results/analytics
- Standardised Web services across platforms
- Service registry for discovery and negotiation of content
- Method registry for discovery of API/tools for specific datasets

The International Virtual Astronomy Alliance (IVOA) offers a succinct description of the goals and approach (with my analogues added in brackets) which seem parallel to the needs of the Web Science community:

> "The VO [WO] allows astronomers [Web Science] to *interrogate multiple data centres* in a *seamless and transparent way*, provides new *powerful analysis and visualisation tools* within that system, and gives data centres *a standard framework for publishing* and *delivering services using their data*. This is made possible by *standardisation of data and metadata*, by *standardisation of data exchange methods*, and by the *use of a registry, which lists available services* and *what can be done with them*."
>
> **http://www.iova.net**

Wendy Hall acknowledges the inspiration VO provided to her group's proposal for WOs. Clear parallels for key principles and practices in Web Science can be seen here, and we will consider comments/insights from the VO team where they appear relevant to WOs.

2.9 Web Observatories

Despite the founding of the Internet Archive in 1996, which was perhaps the first Web Observatory, the explicit notion of using Observatories for Web Science first appears in the literature[23] in 2012. (Spaniol et al., 2012) focuses on surfacing centralised (static) web archives such as the Internet Archive, (Gallen 2013) presents early design challenges for Observatories and (Hall & Tiropanis 2012) establishes the need for a specific type of analysis in Web Science. The key (but often blurred) distinction between data-**about**-the-Web (metadata) and data-**on**-the-Web (content) is raised here. The authors propose that both content and structural/activity information is required to build a picture of the Web as "not only a shaper but also a reflection of human activity". Although this paper is not dedicated to WO, the ultimate goal of the WO is elucidated here for the first time here, which is to provide:

> "a harmonised collection of new and existing data sources and analytic tools"
>
> **(Hall & Tiropanis 2012)**

against which the authors propose a rich variety of interdisciplinary research methods can be applied.

No technical specifics of the WO are given here, and it is not until (Tiropanis et al., 2013) that we see initial outlines of the structural principles of a system. This is an important paper offering much more detail in terms of components (data repositories, catalogues, harvesters and analytics) and also some best practice principles as outlined in the earlier 2012 paper including the focus on openness in terms of harmonised access and the use of open standards.

The representation here specifically includes interdisciplinary skills and interoperation with other Observatories and is described as a "global data resource and open analytics environment" and specifically targets the Web [Science] Observatory towards addressing societal challenges.

Additional features beyond the 2012 introduction are given here e.g., synthetic data and simulation (forecasting), live monitoring and longitudinal aspects of data archival and curation (hindcasting). The tripartite nature of WO usage (namely Academia, Industry and Government) is proposed.

[23] Though there are earlier working papers and presentations available developing the idea informally.

A final element of "social innovation" is included in the WO blueprint but it is not clearly defined in this paper what this innovation comprises and if this is intended to be a result (or as a constituent part) of the WO. i.e., an innovation input required for the WO to function, an innovation output which results from the WO operation or both.

Critically for this project, this paper talks about:

> "the relevance of this effort [the creation of the WO] to multiple stakeholders – academia, government and industry – and their sustained engagement in a partnership, is key to the Web Observatory's success." (**ibid**)

Hence, it seems clear that an understanding of *what this relevance is perceived to be by each party*, what would sustain engagement and how the partnerships would be structured are of central importance to an understanding of the WO. It seems clear that the authors propose a role for WO beyond academic research alone calling WO:

> "..a framework for the harmonisation of e-Infrastructures that go beyond Web Science. As such, even though the Web [Science] Observatory is focused on data about the Web, its standardisation efforts will enable the development of observatories about all data on the Web."
>
> **(ibid)**

Contemplating sub-types of Observatory from this the wider class of interoperating Observatories globally (including non-philanthropic and "for-profit" versions) might need to consider the integration of open source tools and open datasets with private datasets, private content/tools and other commercial elements.

(Gallen 2013) summarises the W^3O architecture, like Google's architecture, as needing to support "a symbiotic relationship between participants contributing at multiple levels". He usefully characterises the problem as bringing the relatively few skilled researchers from various disciplines together with the vast array of source data from the Web via partnerships with data owners using a system which can mediate the interaction. He does not suggest, that this might relate to research in government or business, though this is implied by some of the suggested data sets. He cites several potential sources of WO data, and this list is split between content (data) and metadata - again this highlights a grey area around using data-*about*-the-Web and data-*on*-the-Web. Ethical control and privacy around the use of data sources are touched on briefly here particularly as some of the sources which Gallen cites might require legal approval to obtain.

(Brown et al., 2013) breaks apart the requirements black box more generally and a specific line of enquiry is started in which the nature of what is done with WO systems also builds on the idea that WOs may be used differently by different groups - something not addressed in (Gallen 2013). The paper makes an initial attempt to list core operational/problem-solving processes from which a WO might be constructed. In (Brown et al., 2014) the functional aspects of WOs are decomposed from the literature into a candidate faceted taxonomy and rendered as a structured concept map. Both these papers are early work, and revised models/results are presented in Ch9.

Regarding the challenge of data discovery of previously unknown datasets and services, (Difranzo et al., 2014) describes an extension to the existing schema.org standard enabling the addition of microdata markup allowing standard search engines such as Google to crawl participating sites and to assist in discovering previously unknown content and providers. Up to this point, only lists of known datasets had been available, and these were created using tools such as the Web Observatory Semantic MediaWiki (WeST at Uni Koblenz). The authors give credit to this early system as an important initial step but also observe the limits of the centralised approach as not ideal for decentralised W^3O requirements. They point to the use of microdata as a more suitable method for Observatories/datasets/tools/services discoverable through the creation of four new proposed classes extending schema.org namely: "Web Observatory", "Web Observatory Project", "Web Observatory Tool" and "Web Observatory Dataset". This is early, but valuable work and the authors make a significant contribution here. The implication of this work is that previously unknown data sets may become apparent to the WO though there is not yet a solution offered to evaluate whether the dataset is valuable/trustworthy – only that it exists. Such determinations are left in the paper to social rather than technical processes.

In the third paper in the Hall/Tiropanis series (Hall et al., 2014) the focus is given over to the analytics and querying over distributed linked datasets. The ambition of the Observatory is not only to store and curate data sets but also to run opportunistic analytics in (near) real-time. Conversely, it is *not* the aim of the WO to seek to obtain and store all knowledge locally for a particular domain but rather to discover and link to it on an 'as-needed' basis presenting logistical and performance challenges for large, distributed datasets and distributed queries. Additional implementation details are presented, and the paper goes on to introduce the Southampton University Web Observatory (SUWO). The inference from this paper is that users consume data via a remote (potentially crowd-sourced) provider such that the knowledge of the nature and credibility of that data would again require prior human “review/approval” as the core system described here does not (yet) explicitly provide trust/provenance metadata.

In (Tiropanis et al., 2014b) the architectural challenges around WO systems that offer both local and distributed data and analytics are discussed, and the paper describes a sharing approach addressing the challenge of connected Web Observatories. The authors accept that not all datasets can be assumed to be public and therefore security/access control are described both for user logins and API access. The second aspect of WOs, data harvesting, also offers two approaches to add data to any locally staged datasets, namely API access and Web crawling which may be obtained/managed on a topic-centric basis (from many sources) or on a sources-centric basis (comprising many topics). This is an important contribution in that the recognition of open and private data pre-supposes different attitudes and requirements from WO users.

In (Tiropanis et al., 2014) the authors introduce a novel positioning of the WO as a key translation between diffuse and often uncoordinated open datasets and highly centralised and formal Big Data infrastructures. The WO is positioned as "a middle layer between the Web-of-Data and Big Data in the enterprise" potentially offering advantages from the decentralised flexibility of the former combined with elements of control/rigour of the latter. Despite the popularity of Big Data approaches with business the authors point out that decentralised architectures have more potential to generate externalities and network effects while Big Data systems can lack the advantages of a decentralised Web-like approach:

> "..many of today's big data infrastructures are limited to a centralised or distributed infrastructure that is under a single administrative domain, and the data analytics team that engages with the datasets involves a limited number of people who have been granted access to it."
>
> **(Tiropanis et al., 2014)**

This might offer a new Web-based alternative, say the authors, to the earlier eScience Grid by removing the need for centralised access to limited resources/processing. The paper stresses the need for WO standards and the ability to optimise processing and harmonise access across a wide array of data sets and platforms. It predicts more growth and bigger challenges around volume, optimisation and security with the advent of the IoT (Internet of Things).

The implications here are that:

- That data may be *assembled* from multiple sensors/systems and subsequently leveraged as data rather than simply uploading *existing* traditionally gathered research data
- That resources available from individual WOs are inherently less significant than broader resources available in a global W^3O eco-system though potentially challenging to harness.

(Tinati et al., 2014) and (Wang et al., 2015) pick up on technical challenges of Big (large, messy, fast-moving) data and exposing database credentials securely. They emphasise the need for operational efficiency and security in providing the service to end-users. They compare different approaches to capturing, caching and re-publishing high-speed feeds while retaining control of permissions/licenses and data-sharing. They outline acquisition, pre-processing and streaming phases which may be difficult to optimise due to the need for data enhancement (e.g., geo-lookups) which typically call external services and are therefore not under the direct control of the WO operator.

The authors discuss ways in which streaming data may be used to create meaningful proxies for the analysis of Social Machines. The use of such proxies and in particular the risk of unsophisticated proxies/comparisons is, however, a contentious issue. (Tinati et al., 2014b), (Pope 2014) discuss the validity of (i.e., the epistemological value of) signals which:

- May simply be reporting *correlation* rather than *causation*
- May be taken out of context

even assuming that one can be sure of the validity of the data itself.

For example,

Figure 2-2 presents an intentionally absurd example of fitting/correlating Web search data using Google Correlate (developed from the Google Flu algorithm). Without further investigation, one might suppose (or, worse yet, *act*) on some presumed causal relationship between "nuclear disarmament" and the "polish government" or "euthanasia statistics" given the ranking of statistical correlations (Both are Top 5 hits). We may also consider similar unexplored links between "Scottish independence" and "Colombian cycling" whose variations are 97% positively *correlated* for the period studied. The level of *causation* remains unexplored.

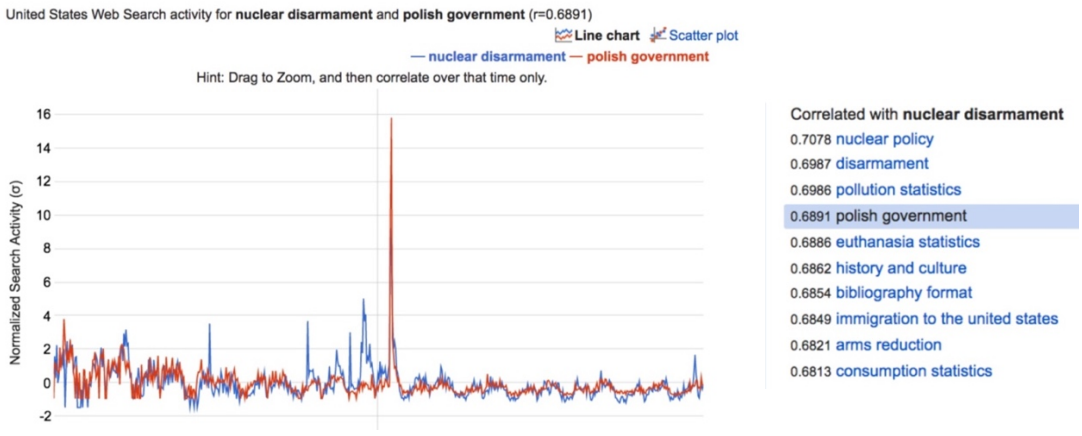


Figure 2-2 Google Correlate example of correlation vs. causation. Accessed Feb 2017

Thus we come full circle from the ancient Athenian's need for 'trustworthy copies' and the concern around working with flawed, misinterpreted or deliberately subverted material to the challenges of modern digital sets which are so vast and fast-moving that they increasingly represent what I will call '**Dark Data**'. I appropriate this term from its more typical meaning (unused or "dusty" data) to indicate instead data (or the automated actions/reports derived from them) that we are required to 'trust' as the product of a 'black box' without proof/transparency. Increasingly humans may never directly view information assets due to some combination of:

1. The lack of availability/access of the underlying data (real or simulated)
2. The lack of transparency/clarity around their manipulation of data in terms of the analysis/algorithms
3. The lack of physical capacity (time/bandwidth) to review large datasets

The generally accepted definition of dark data in the business literature is a more benign collection of inert/unseen business data (like dark *matter*) that companies do not actively use or may not be aware of. I submit that this definition is superseded by a wider discussion, given the broader data deluge extending beyond the business context, and (in the sense in which I use it) is more socially/politically charged. It is more akin to dark *warehousing* or dark *manufacturing* where machines manipulate goods/materials without human supervision, and the end-product emerges requiring human trust that some safe/acceptable process has been followed without deviation or corruption. Dark Data in this sense is that output for which there may be little/no transparency to the provenance, accuracy, algorithms or methods used to generate it. Such data may be numerical/analytical or thematic such as the claim of fake news and social media manipulation which has occupied the public following the 2016 US presidential election. WOs then must be considered in the light of the trustworthiness of the data they provide.

Functional automated 'black boxes' per se have existed for decades in the form of algorithms which use causally-linked data in applications, reports and spreadsheets and for which the designer's explicit instructions form the point of trust for us to accept the outputs. As data is increasingly harvested/mined with algorithms that search for correlations and *implied* causations, the results may be benign in intent such as filtered, personalised search results. They may also result in emergent effects such as self-reinforcing 'filter bubbles', potentially distorting the

perception of what is true and undermining the principles of objective/reproducible research. Observing and addressing such unintended effects is one objective of Web Science research.

Work on analytical services such as Truthy (McKelvey & Menczer 2013), Bot-or-Not (Davis et al., 2016) and Hoaxy (Shao et al., 2016) may help to identify fake data/news and underpin trust. Qualitative factors apply such as context, meaning and other non-technical elements such as licensing and liability. All of these factors are potentially pushed to the forefront of design for a trustworthy WO. Ironically, systems offering to ensure accuracy/truthfulness have themselves been accused of distortions or inaccurate analysis/reporting:

- Truthy (produced by a WSTNet member) came under fire from news groups as being politically rather than academically motivated and "Orwellian" in its infringement of privacy with a robust defence being posted by the research team.
- Recorded Future (a commercial Web Observatory) was also drawn into a political controversy and accused of being a mouthpiece for the US Government when reporting on apparent causal links between the release of Edward Snowden datasets and the encryption methods used by Al Quaeda.
- Palantir provides analysis of metadata for US national security purposes – a relationship which has received more visibility[24] since the 2014 quote by former NSA/CIA director Hayden stating "We kill people based on metadata".

Whilst some of the original video/text material for the first two stories has now been removed from the Web, this highlights Feenberg's political perspective on technology and the delicate balance between the deeply-cherished (yet at times irreconcilable) twin notions of 'transparency' and 'privacy'. At the time of writing the U.S. Trump administration are positioned in opposition to several domestic/international news organisations [25]in disputes over the post-truth ('truthiness' [26]) status of claims/statements made by the administration.

Despite the broad social context for data sharing in WO the existing WO literature tends to focus on technical elements and techniques relating to the construction of specific WO instances and the technical processes of data sharing, querying. Whilst they provide a significant and valuable contribution to the discussion the social impacts are perhaps underrepresented especially as these may impact on the adoption/interoperation aspects of WOs.

[24] https://www.rt.com/usa/158460-cia-director-metadata-kill-people/
[25] http://www.telegraph.co.uk/news/2017/02/24/donald-trump-bars-new-york-times-cnn-politico-white-house-press/
[26] Humorous term coined by U.S. satirist Stephen Colbert in 2006

The result of this focus is that the distinctions between processes required for individual instances versus a broader interacting set of Observatories are often conflated and are generally covered at the standards/technical level but not at what (Halford et al., 2010) calls the "performative" level. Whilst the usage of WO by multiple parties is acknowledged, the specific parties and their required processes are not investigated. (Grudin 1990) proposes that IT systems gain complexity/abstraction as their focus moves from hardware → software → cognitive elements with (Kuutti 1996) adding 'organisational' as the final abstraction and stressing that systems be understood through a study of contextualised actions (Activity theory) vs. a purely cognitive or laboratory approach. This represents a continuum from purely technical to 'socio-technical' systems and underscores the importance of studying WO in context. Finally, whilst the importance of incentives and social models is introduced, a practical model for motivations and potential collision/complementarity between groups has not yet been developed. These open areas would complement the further study/understanding of WOs and will underpin research questions in Ch3.

2.10 Examples of Web Observatories

This section covers a selection of WOs and WO-like systems (cousins) and offers a range of applications and motivations for the use of WOs ranging from:

- Generic instances where a WO but no specific data/service is offered (e.g. DIY model) to
- WO-based systems where specific tools/services are provided but no access to the underlying WO is offered (e.g. a service provider model).

SUWO is presented in (Tiropanis et al., 2013) outlining a basic premise for dividing data harvesting into source-centric (covering all topics from a single source), and topic-centric (covering one topic from several sources) approaches. SUWO is framed as a topic-agnostic testbed to investigate different approaches/formats for long and short-term storage and interchange rather than as a domain-specific Observatory (e.g., health or environment). The desire to support hybrid (heterogeneous) datasets has led to a focus on open data standards and the need to synchronise datasets as time-series using a proposed Web Observatory time identifier - thought of as a 'superordinate identifier' above source or format. Data sets are separated from analytics and visualisations (so-called 'apps'), and current examples on SUWO are implemented in a range of third-party tools and libraries such as D3.js and Tableau. It is envisaged that interoperability (discoverability) between SUWO and other WSTNet Observatories be a key focus for future development, and this has been a key focus of the early SUWO releases.

(McKelvey & Menczer 2013) introduces the Truthy social media Observatory intended to distinguish between genuine "grass roots" activity and fake (orchestrated) activities - so-called *AstroTurf*. The authors look at Observatories as a broad solution to Web Science and computational social science and offer a social media Observatory as an example on the basis that social network data offers a valuable proxy for the Big Data sources that model the behaviours/activities that may be studied in this context. The paper stresses the need for interoperability across Observatories and does this through a call for API access to both underlying data and statistics (analytics), which removes the need to agree on a single set of standards for all repositories.

(Pongpaichet et al., 2013) presents the Eventshop Observatory, which considers the Web to offer localised series of events from sensors and systems over time, which can be used to describe situations (so-called "spatiotemporal thematic streams"). This paper builds on earlier Eventshop work by (Gao et al., 2012) on combining Complex Event Processing (CEP) systems and Geographical Information Systems (GIS) as a basis for Social Life Networks. Such systems can recognise real situations based on the aggregation of sensor, mobile and OSN data across a population and use this recognition to provide information, updates and even warnings to relevant groups.

(Chua et al., 2012) presents the NeXT Observatory, which focuses on UGC (user generated content) from the perspectives of topic, person, organisation and location. This Observatory has crawled vast collections of Tweets (Weibos and equivalents) and Flickr images for:

- Image location-oriented information and check-in venues
- Topic-oriented information, such as tweets, community Q&A and discussion forums
- App-oriented information
- Structured/linked information, for a selected location.

(Gloria & McGuinness 2014) introduces Health Web Science and expands on the legal/ethical questions from an earlier paper offering an approach comprising some methodological elements from the RPI standard WO method including Data identification and description, Origination, Usage, Citation, Provenance and Policy. The schema.org extension for the discovery of data and tools is particularly stressed.

(Wang et al., 2015) introduces a project based on the SUWO template to implement a policy-focussed WO for the South Australian government. This paper is grounded in real-world applications of WO outside of academic research and comprises input from academic, business

and government stakeholders. Despite an existing set of open data materials, the regional government have found that more is required to mobilise the available data effectively since:

> "..the current data publishing and sharing solutions hit barriers due to the lack of data provenance and security mechanisms, as well as limited use/usability for applications. Further, building analytic applications that make use of those datasets in line with their sharing policies is another challenge as many users have little digital literacy. "
>
> **(Wang et al., 2015)**

(O'Hara et al., 2014) and (Sackley 2014) bring arguments around transparency vs. privacy in the application of WOs and Web Science respectively in the area of security /policing. (O'Hara 2014) cites the highly federated nature of UK police forces suggesting that centralising more than 40 systems is both impractical and lacks backing from the individual forces. The implementation of a Web-like platform, which it is stressed, must underpin the required levels of security and accountability, must also meet high levels of ethical and legal regulation. This potentially offers a more flexible and usable method for assembling a useful view of distributed data around offenders and crime. (Sackley 2014) draws on Social Machine examples to argue for the modelling of crimes/victims in the search for a better method to addressing re-victimisation which he argues might be supported through suitable models of similarity captured through social data and crowd-sourced contributions.

(Price et al., 2017) offers a concrete set of objectives for mobilising international data and researchers through the WST/WUN project by creating four WO demonstrators aligned to four major research objectives aiming to unite more than 90 organisations across the membership. The four demonstrators cover specific questions around climate change, education, public health and culture and have an independent existence at a project level (effectively as WOs) whilst remaining open for discovery/sharing of datasets across the membership (effectively as a W^3O)

(Tinati et al., 2017) proposes the application of crowdsourcing techniques used in citizen science platforms to address the classification/trust/integration issues around a diverse and uncoordinated ecosystem of IoT devices using an IoT Observatory. IoT Observatories may not be Web Observatories[27] in the strict sense save that they exist *on* the Web and that the IoT may be *web-like*). The paper makes an interesting contribution and the data gathered is mooted to be potentially valuable for wider (Web Science?) research. Notable here is that this paper also offers

[27] Indeed the idea of a Web of the Internet-of-things seems self-contradictory other than in the sense of an IoT dataset **staged** on the Web.

an excellent example of the "flexing" that is associated with the Web Observatory discourse in which the definition of *Web* seamlessly flows between:

1. The idea of **about the Web** i.e., in a "science-of-the-World-Wide-Web" sense
2. **About Webs** i.e., in a "science of (any) web-structured system/network" sense
3. Anything **Web-based** i.e., in an "(any) science-carried-out-on-the-Web sense"**.**

Whilst Web Science formally declares an interest in "Web-like systems", definitions (1) and (2) may be thought of as fairly standard. The inclusion of option (3) requires more consideration.

By analogy if we consider whether "climate science" is intended to indicate "science undertaken in a particular climate" or whether "rocket science" offers a similar duality we may choose to reject option (3) as being core to the Berners-Lee et al. original definition of Web Science. In some sense, we are the victims of a seductive play on the words "Web Observatory". The research will investigate this flexing in more depth given that a lack of clear objectives/intentions may impede funding, adoption and collaboration for the WO community.

In the commercial space, Recorded future (founded 2010) is a partly Google + CIA (In-Q-Tel) funded Web Observatory (they call it a Web Intelligence Engine). They claim[28] "8 billion data points across six hundred thousand sources" and focuses on the alignment of entities and events from multiple social media sources into a "Temporal Analytics" model for threat intelligence. The ability to query the sources and access trends is offered, but no direct access to the WO or custom/private sources is publicly offered.

Palantir (founded 2004) is a big data analytics company with two main offerings: Project Gotham which offers counter-terrorism analysis and Project Metropolis which focuses on Financial markets. Their data sets are not exclusively on/about the Web but constitute a framework for examining the meaning and relationships/patterns between diverse and distributed datasets which may be accessed via the Web.

Broadly we see that WOs may share features with other single focus virtual observatories but are proposed to span research in academia, health, government and business and to variously focus on one source (many topics), one topic (many sources). They will combine elements of data (both local/linked) analytics and importantly the contribution/curation/collaboration of users and participants. WO projects may suffer from the duality of moving between being *about* the Web or simply located *on* the Web.

[28] http://uk.businessinsider.com/recorded-future-can-predict-the-future-by-analyzing-everything-on-the-web-2015-5

In the WO literature, we see a focus on the *application* of specific WO's situating each exemplar at the centre of its of own problem space rather than a *contextualisation* of WOs and WO sources within a broader data ecosystem. We see the importance of access vs. privacy throughout these examples, and there is a common need for standards flagged as the foremost challenge: standards for data exchange and analytics exchange with the need to discover datasets and resources through extensible standards such as schema.org.

In the WO service literature, the platform may centralise its data and focus instead on providing innovative discovery/analytic features which require less openness or co-operation and more topic focus, efficiency/scalability and usability.

2.11 Conclusion

In this chapter, the nature of innovation/adoption, the emergence of digital technology, the development of the Web on top of digital networks and the need for a science of the Web have been discussed. The presentation of literature around a number of explanatory theories relating to WOs has been delayed until later chapters where they are discussed in context with experimental results since these comparisons form part of the *findings/analysis* of the project and not part of any *hypothesis* to be tested.

The ever-broader interactions between people (society) and the Web have created new social mechanisms for which current Web tools are not designed, and hence a Web Observatory, conceptually based on earlier astronomical VOs has been proposed. Web Observatories are intended to study different aspects of social process via the Web-of-Pages, the Web-of-People and, increasingly, the Web-of-Data. Given the diversity of the proposed applications, a definition of WO will need to respect the different ways in which users and groups may frame (contextualise) their expectations and experiences of the Observatory and offer a firm basis on which users can trust the results it offers.

This literature not only informs the creation of research questions for the project but also directly contributes example features/facets for a generalised model of WOs which is grounded in the literature of practitioners.

Chapter 3: Research Framework

In Short ..

There is limited work on the theoretical nature of Web Observatories and (as yet) limited empirical data to support research. One cannot therefore easily seek to confirm/refute theories of Observatories in a deductive manner nor leverage comparative datasets and statistical approaches to quantitatively model or predict outcomes.

A qualitative grounded theory (GT) approach using multiple sources/perspectives has therefore been selected to build substantive models from the available data which may (in time) prove to be generalisable. The data employed includes published literature, public discourse and the review, both in vivo and via documentary proxies, of examples of WOs (and similar/related systems).

The approach combines facet analysis with ideographic perspectives of participants whose subjective experience is captured. The approach blends faceted content analysis (CA) and Interpretative Phenomenological Analysis (IPA) from a constructivist grounded theory (CGT) perspective. Each of three main focus areas (Academia, Business and Community) are supported by multiple surveys/observations/investigations.

Research questions aiming to provide a flexible framework for defining and characterising WOs from these diverse sources are proposed. The approach will contribute new models/perspectives from which WOs may be compared/analysed.

3.1 Introduction

Available data sources and existing research approaches will be presented/evaluated to support/refine research questions which can be addressed within the timescales and resources available. A brief introduction to ex-ante assumptions and reflexive issues will be given to situate the research design choices within the context of personal perspectives as well as in relation to related/complementary research.

This approach will contribute additional tools/models to wider WO research efforts as well as extend the current technical vs. topic focus on WO to a broader socio-technical perspective. It extends the use of IPA into web-based systems and is the first study of ideographic perspectives in WOs.

The research comprises several parts:

- The extraction of key concepts from the discourse/narrative into a facet list
- The iterative grouping/structuring of these facets into a taxonomy of features grouped into Definitions, Narratives and Agency or 'DNA'
- The arrangement of facets into visual DNA templates/structures
- The refinement of the models using selected participant interviews and reviews with the community of practice along with visual narrative models.

The goal is to conceptualise WOs using multiple perspectives and elements from which Web Observatories are construed/constructed.

3.2 Research Questions & Objectives:

WOs are a nascent idea such that it is important to establish distinctiveness and complementarity with other related areas to manage impact, leverage existing capability and avoid duplication of effort. Given there are no existing definitions/theories of WOs to confirm/disprove, we must work inductively – building up from observations on the nature of WOs to evolve a candidate model/definition. This will be an iterative process and starts with 'straw' models seeking to determine what WOs are 'made from', how they are used and what factors drive users to participate. There is little to be gained from a quantitative study of this new developing eco-system area since insufficient numbers of WOs exist for statistical measures to be robust/meaningful – the focus is therefore on the descriptive and qualitative nature of WOs.

Through interactions with colleagues and other interested parties early in the project (see Ch4) it became clear (prior to selecting research questions) that different parties seemed to hold markedly different conceptualisations and understandings of WO regarding its application, distinctiveness and utility. Despite this, the fundamental IT conceptualisation of (Data In, Analysis, Data Out) seemed universal and uncontroversial, and therefore more abstract distinctions/contexts seemed to be at work.

From (Whitworth's 2009) work on sociotechnical systems (Figure 3-1) we see technical (hardware) layers augmented by software (including data), interfaces/applications (via HCI) and personal → collective goal-seeking applications via social context/grouping. It seems this approach offers a more expressive and nuanced approach for the definition of WOs than hardware classification alone and aligns with (Grudin 1990) on the complexity of systems emerging from cognitive and organisational factors.

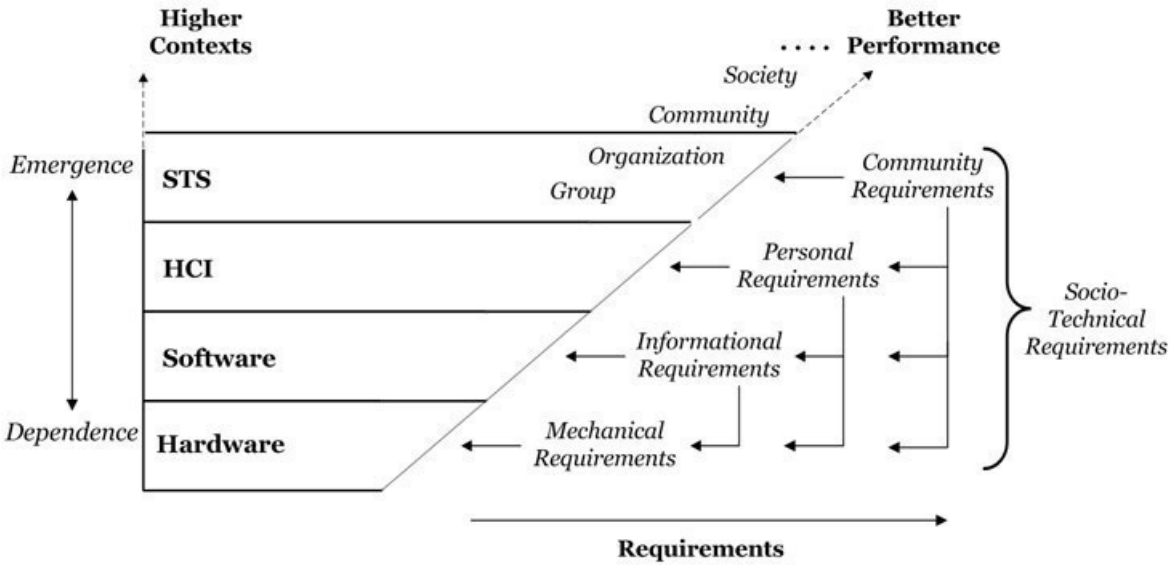


Figure 3-1 Extending technical system elements to socio-technical perspectives

Considering (Whitworth 2009) in Figure 3-1 above, (Abell 1980) and (Halford et al., 2010), key perspectives or groups of distinguishing features were extracted for a WO conceptual map in order to drive the development of research questions.

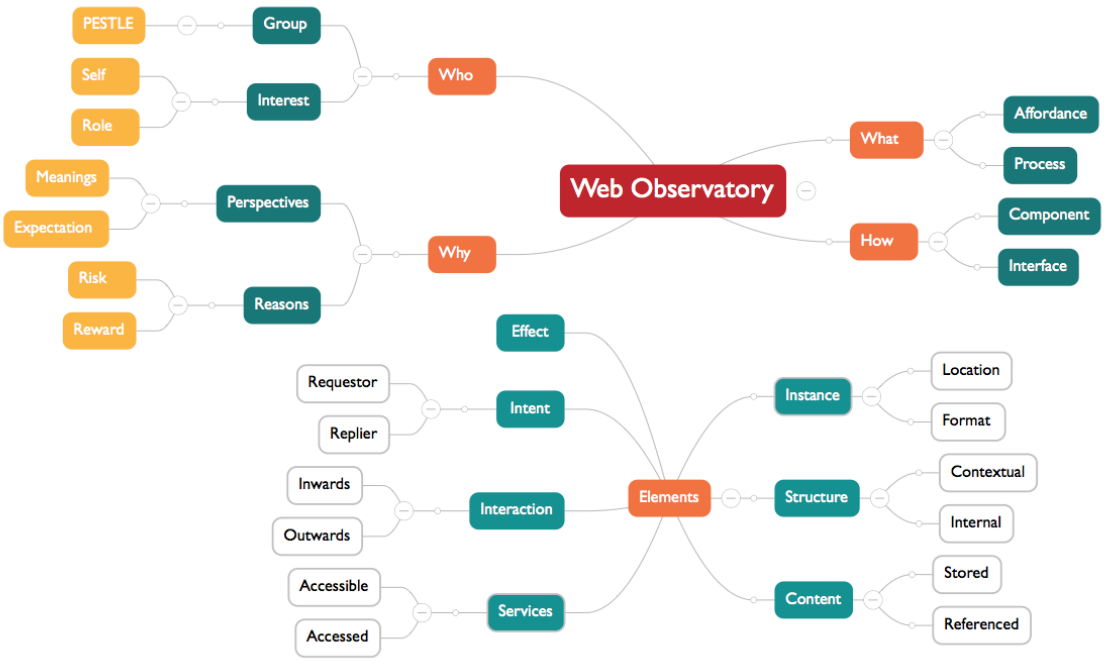


Figure 3-2 Conceptual nature of WOs

This generated candidate questions/elements by allowing an approach to be broadly seeded in key areas, asking questions to understand what a thing is, how it functions and why users engage with it. These elements/perspectives define terms of reference, mark out a broad area of investigation and also underpin/support the objectives of the research:

- To understand enough about WO structure/operation to offer a WO definition that is both robust and extensible for future changes (i.e., not only what a WO currently does but also its future capabilities)
- To understand enough about what makes WOs different from each other to offer a description of variations and sub-types
- To understand enough about WO user perspectives to understand the motivations/incentives for agents/agency both in isolation and amongst groups.

(Trochim & Donnelly 2008) assert that research questions are typically split into three types with each type supporting the development of an answer to the next in three broad categories:

1. **Descriptive** (e.g., what are the technical elements?)
2. **Relational** (e.g., how they are arranged in relation to other technical elements?)
3. **Causal** (e.g., what needs to drive the inclusion of this technical element? What behaviour is driven by the inclusion of this technical element?)

They argue these are hierarchical ("cumulative") in nature, in that to establish cause one must first establish the parameters which may relate to each other. For this to be possible one must first be able to describe the system and its components such that [3] ←[2] ← [1].

Considering the availability of data (listed above), and working from the research objectives and literature review, the proposed scope/purpose for the research questions and research elements were generated as follows:

1. The ability to identify a common set of structures/processes around Observatories
2. The ability to identify any differences in the way different groups implement/use an Observatory
3. The ability to measure which features from an idealised set of affordances a particular Observatory might have
4. The ability to identify different homogenous/complementary processes across different tribes.

3.2.1 Research Questions

As we have seen in Ch2, considering WO to be applicable by, and between, different users in different ways steers us to the features, processes and perceptions around WO and to investigate how positive/negative perceptions inform adoption. Our research questions are thus as follows:

1. Which perspectives can help us to clarify the structure and nature of WO, not only as a purely technical artefact but also as an assemblage of users and technologies within a social context?
   - *What are the social and technical elements of WO? Are there other types of element?*
   - *Is the WO predominantly a context-free tool or is it attempting to address 'social constraints' at scale in the form of what Berners-Lee called a 'Social Machine'?*
2. If we consider WO as being socially-embedded in the processes and ambitions of different groups who use it, is there evidence to suggest that WOs might be perceived/operated differently across social groups?
   - *What are the implications of any differences for engagement/interaction with WOs between groups?*
   - *Is WO innovative (novel) technically and/or socially with respect to other technologies and approaches? How does this affect adoption?*
3. What benefits can a socio-technical model of WOs offer in terms of insight for the creation of new observatories, innovative applications and the encouragement of participation by existing systems and data sources?
   - *What would a substantive model of WO look like and how might it be leveraged?*

(Bryman 2012) offers six criteria to evaluate research questions which are considered below:

	Qu 1	***Qu 2***	***Qu 3***
Clarity (intelligible)	☑	☑	☑
Researchable	☑	☑	☑
Connected to established research	☑	☑	☑
Linked with each other	☑	☑	☑
Possibility of original contribution	☑	☑	☑
Neither too broad/narrow	☑	☑	☑

Table 3-1 Bryman criteria evaluating research questions

Situating this project within a broader range of work we can see that a definition of Web Observatories fits specifically into existing WO research. Consider (Hall et al., 2014), (Tiropanis et al., 2013), (Shadbolt et al., 2013) and (Difranzo et al., 2014) who are concerned with the structure and standards of specific WOs and data exchange between WOs.

More broadly this project aligns with (Goffman 1974), (Kittler 1999) and (Star & Griesemer 1989) who have studied how individual and contextual perspectives on (technical) artefacts give rise to different meanings. There are similarities with (Contractor & Monge 2003) who propose multi-level models and drivers of group behaviour and (Malone et al., 2009) who uses the metaphor of genes to describe collective intelligence/collaborative software systems.

Finally, this project joins a growing number of projects employing IPA outside of the health/medical domain including (Vanscoy & Evenstad 2015) studying Library Information Systems (LIS) using IPA.

3.3 Data Collection

With few WOs available to study "in the wild", seed data for the research was potentially available from documentary sources, observation of participants and *related* observatory-like systems and also by interviewing those with experience/knowledge of virtual observatories.

To ensure as wide a focus as possible material was gathered from a broad selection of 800 academic/commercial/government documents and 50 video presentations to bootstrap a model. To validate/iterate this model 100 survey respondents and 100 additional interview participants across nine focus groups, 77 interview sessions and nine in-depth IPA interviews were used. More than 200 reflexive journal entries were created during the research process. Care was taken to draw from multiple sources from each category of source to avoid basing conclusions on only a single viewpoint.

3.3.1 Primary & Secondary Sources

Primary Sources	*Description*	*Impact*
Cases/Vignettes (3+3)	Primary and secondary cases were developed in Academia, Business and Community sectors looking for common/contrasting themes in relation to Web Observatory usage	Contrasting cases provide real-world feedback around WO usage (RQ1), help identify socially-embedded themes informing the study of sub-types of WO (RQ2)
- Interviews (n=75) - Transcripts (27/75) - Focus groups (9/75) - IPA (9/75)	- A large body of interviews (expert-to-neophyte) were conducted to avoid selection bias and observe a range of personal conceptualisations of WO avoiding propaganda/hegemonic views - Interviews were filtered based on relevance to create 3 groups of 9 interviews that were transcribed for more detailed analysis - Focus groups were used to expose the actual challenges of using WO tools during two international WO workshops - IPA subjects were "purposefully sampled" for deeper knowledge around specific WO projects/topics (not given in the broader interviews) allowing personal conceptualisation across socially-embedded views on the same topic to be compared/contrasted.	- Broad base (approx. 25 per sector) gives rigour to conceptualisation across groups (RQ1) and validates/eliminates the seed concepts (RQ1). Perceived problems, challenges also inform an understanding of adoption (RQ3). - 3 groups of 9 socially-embedded interviews (academic, business, community) address the ability to situate divergent and convergent WO definitions within an occupational setting (RQ2 and RQ3) - Focus groups informed academic experience of early tools and validated research methods and analysis. - IPA gives a voice to the users' narrative (RQ1, RQ3) beyond the identification of codes/tags generated through the seed model

- 3 Questionnaires (n=6) - On-line surveys (n=100)	- Targeted questionnaires were (unsuccessfully) used to validate the codes/concepts extracted via content analysis - On-line Surveys were used to gauge the diversity that the service naming had on non-expert groups	- Users were unwilling/unable to complete a direct validation of the initial DNA 3-part model - Adoption of WO will be affected in part by the understanding and (mis) conceptions of non-experts
Review/analysis of existing projects/tools (n=75)	Systems/tools sourced from the WSTNet group were compared with commercial systems to determine architectural/functional similarities	Supports analysis of the distinctiveness of Web Observatory (RQ2) in terms of structure and application
Participant Obs of: Ac./Industry Workshops Project meetings (SOCIAM, WO, WST) Commercial meetings Seminars	Meeting notes were taken in relation to Problems/Objectives/Solutions surfaced in Academic/Business/Community groups noting and distinctive features for WO.	Publicly-stated (official) reasons for acting are not always those revealed in private interview contrasting expectations against motivations. (RQ1, RQ3)
Reflexive journals (n>200)	Written/voice journals were used to reflect on classifications, problems and objectives both internally and on discussions with peers/colleagues	Reflective analysis forms part of the research method and directly impacts (RQ1-3)

Secondary Sources (n=800)	Description	Impact
WO focussed: Papers/journals/Poster Workshop submissions	Key themes/concepts were extracted from academic sources discussing WO, VO and related analytics	Analysis seeded models (RQ3) and supported analysis of sub-types (RQ1) and differentials analysis of WO vs related systems (RQ2)
Grant Submissions Academic and Product Presentations/Talks Webinars	Such submissions/talks are typically focussed on problem > impact vs. detail/calculations in academic papers	Understanding the perceived problems/solutions around WO/VO informs the motivations for participation (RQ3) and the conceptualisation being put forward (RQ1)
Web/Press Articles White papers Product demos Video presentations/lectures	Product documentation covering fundamental architecture, functionality and proposed problem space	Distinctiveness (RQ2) and motivations for interoperation (RQ3) are supported through this analysis
Published datasets and analytics.	Open data sets were examined to look for explicit reasons for requesting/using such data	Understanding the various groups motivations for providing and consuming data underpins a model of encouraging participation in a wider WO Eco-system (RQ3)

Table 3-2 Primary/Secondary data sources

Secondary sources were searched and selected based on:

- Known communities of practice such as WST, Web Science and WWW/Web Science track conference/workshop materials
- Orientation towards "virtual observatories", "Web science research", "Web-based research", "Web-based analytics" and both generic and specific search tools were employed to build a corpus. Materials were reviewed and analysed using manual confirmatory classification and automated entity extraction tools to seed a model with concepts and issues from the secondary materials and published literature - albeit refereed/reviewed sources were given more weight when contradictory concepts were found.

3.3.2 A note on automated analysis & entity extraction

Automated textual analysis (including auto coding, word frequency, dendrograms/clustering, word trees/clouds, co-occurrence matrices) enables the processing of larger volumes of data than would be possible manually but at the risk of unvalidated (spurious) results. These approaches were checked for accuracy (see Ch4/5) in relation to manual classification of the same material and found to offer poor results in terms of frequency as the sole proxy for importance. Also, semantic errors in 'fuzzy' searches were noted:

e.g. "*might*" (as an expression of possibility/doubt) vs. "**might**" (as a synonym for power).

Accordingly, these tools were used as a convenience to *seed* initial 'bottom-up' concept lists which were then refined/eliminated in combination with a manual review to enable "*gisting*" of longer documents to support the decision of where to review in greater depth. Ranked word frequency, fuzzy search of terms being "near" others and keys terms co-occurring were used to establish pre-filtered candidate concept lists, but *manual* validation/filtering of the context and importance of concepts through multiple sources and participant confirmation was required for inclusion in the final models. Final top-down cross-checking of refined facets against the source material was performed to check the fit of the final theory against the data in which it is theoretically 'grounded'.

Some counts and quantitative measures (e.g., Jaccard Distance) are used to assess similarity based on raw transcript data. However, no inferential statistics are employed on data retrieved using fuzzy search criteria due to the risk of poor semantic recognition and also, more fundamentally, due to the questionable nature of frequency as the sole criteria for importance/relevance.

A note on ethics

In terms of the ethics process for this project, all interview participants were engaged in line with the ERGO ethics process (ERGO/FoPSE/9487) which comprises a formal ethics plan and internal plan review with the University of Southampton. They were informed of the opportunity to withdraw and/or review their contribution in line with Southampton ethics policy.

- Online material, questionnaires and ethics forms are available via the Southampton iSurvey On-line Service
- Interview subject names are either withheld or are referred to pseudonymously
- iSurvey participation reference numbers are held against participant (pseudonym) names in a password protected system. An unreferenced list is given in the appendix.
- IPA participants (pseudonymously named) were given an additional opportunity to review/revise their input/quotes – 1/9 participants requested some minor edits 0/9 participants withdrew.

There remain wider issues around the ethical use of Web Observatory data *per se* which are not explored by this project but may be significant in the (re)use of shared data by groups with different objectives and different approaches to the ethics of data use.

3.4 From Questions + Data to Research Methods

! A review of research methods in (Creswell 2003), (Bryman 2012) and (Bazeley 2013) suggested a qualitative approach for this project (for example using case study analysis or participant observation) as in (Yin 2008). Creswell argues that qualitative research is more appropriate in fields where concepts are still evolving and where little theory has yet been developed. Qualitative approaches typically employ coding (vs counting) techniques where the various qualities/perspectives/facets of the subject under analysis can be marked and associated with tags/codes which support later analysis and visualisation. Thus, a number of methods can be built on top of what one might define as a "facet-oriented" platform.

Data sources are coded/tagged to represent key themes. An example of thematic analysis → entity/facet is given below from the Virtual Astronomical Observatory (VAO) project. Italics are added to signify the identification of a theme, facet or construct:

> "The VO allows astronomers to interrogate *multiple data centres* in a seamless and transparent way, provides new powerful *analysis and visualisation tools* within that system, and gives data centres a *standard framework for publishing* and delivering *services* using their data. This is made possible by *standardisation of data and metadata*,

by *standardisation of data exchange methods*, and by the *use of a registry,* which lists available services and what can be done with them."

source: http://www.iova.net

Whilst coding approaches do not preclude all quantitative analysis (descriptive statistics around the incidence of tags/codes are entirely typical) care must be taken to ensure quality/robustness of any broader (inferential) statistical analysis. Once sufficient volumes of data on interoperating WOs become available the ability to test the impact or contribution of different factors more robustly may come into play using techniques such as:

- Facet Theory (Guttman & Greenbaum 1998), (Shye 1999) and (Brown 1985).
- Generalised Morphological analysis (Zwicky 1957), (Ritchey 2012)

However, within the scope of the current project, I offer only the first step of constructing the core set of contributing factors leaving further inferential/predictive analysis for future work when it can be underpinned by a more robust body of data.

Gathering data to build (rather than test) theories inductively offer two broad approaches: *theory-by-observing* approaches in the manner of Observational Analysis and Grounded Theory and *theory-by-doing* approaches in the manner of Participatory Action Research. The dominant approach selected here has been Grounded Theory and Participant Observation.

Further consideration of (Creswell 2009) and (Bryman 2012) led to the selection of grounded theory (GT) after (Glaser & Strauss 1967) particularly given the lack of a well-formed and mature theory of Web Observation. A key research objective here is the formation of new definitions and theories from the observation of the phenomena and the possible effect of the social context.

It has been argued (Thomas & James 2006) that objectivity may not be assumed and, in fact, an off-shoot of GT known as constructivist grounded theory (CGT) see (Charmaz 2014) and (Bryant 2002) particularly emphasises the co-construction of data and theory by the researcher as part of the interactions during the research processes. It is within a flexible constructivist grounded theoretical framework that I assembled a range of pre-interview and post-interview sources.

Considering the subjective "lived experience" of participants with relation to the ideas and systems underpinning Observatories invokes the notion of IPA (Interpersonal Phenomenological Analysis). IPA allows the "reality" of Observatories to be closer to a series of intersubjective agreements of what it means to "be an Observatory" or "to observe" than to a formal definition by which the community should be expected to abide.

Within the grounded theory framework, I have chosen to implement a range of techniques including facet analysis (Ranganathan 1967), (Spiteri 1998), concept maps (Novak & Canas 2008), process hierarchies, content analysis, participant observation and the framework of the WO/W^{3}O cast as a multi-dimensional model.

(Creswell 2003) considers it good practice to situate the choice of such research methods with a context of others who have successfully used such techniques and thus in a review of projects that have successfully employed the paradigms and techniques proposed here I submit the following examples:

- (Malone et al., 2009) have employed a gene/genome metaphor to defining and analysing the nature of collective intelligence systems
- (Smart et al., 2014) have employed taxonomy to categorise examples of Social Machine
- (Malone et al., 2003) employs grouping and visualisation of process flows within the MIT process handbook project
- (Contractor & Monge 2003) have attempted to integrate an understanding of social theories to the analysis of network formation under co-operation/collaboration
- (Proudfoot et al., 2011) have discussed the application of facet theory to the classification of Internet interventions while (Levy & Guttman 1985) uses a faceted approach to categorise a system of social values
- (Vanscoy & Evenstad 2015) favourably evaluate two IPA studies - (Vanscoy 2013) and (Evenstad 2011[29]) - concerning the subjective experiences of technical systems and processes in Library and Information Science (LIS).

3.5 Methods

3.5.1 Taxonomic (Faceted) Analysis

As I argue in (Brown et al., 2014):

> "in the analysis of types of entities seen "in the wild" (natural or technological) it is often helpful to group/cluster the features, behaviours, structures and other phenomena according to classification schemes which can help in generating knowledge/insight about these entities."

[29] Originally "Fortolkende fenomenologisk undersøkelse (IPA) av utbrenthet blant tre IKT - ansatte i Norge." Not available for review in English language

The linkage between classification and knowledge is asserted by Kwasnik and a number of structures for taxonomies such as Trees, Hierarchies, Facets and paradigms are contrasted in this piece. A selection process for a classification scheme I adopted is based on (Spiteri 1998). Given the definition of Web Observatories is currently in flux, Spiteri recommends avoiding the use of hierarchies/trees. The lack of automatic inheritance or transitive relationships between features also points to the use of facets, which she describes as more "hospitable" to adaptation.

(Morshead 1965) requires both a structure and a conceptual model for a Taxonomy whilst in (Brown et al., 2014) I argue that:

> "Physical implementation is less relevant according to Spiteri's test of faceted classifications in relation to other factors (though not to the implementers of Observatories themselves)."

Given that physical implementations of WOs may be only trivially different whilst applications/interaction may vary widely making the conceptual model potentially more relevant.

3.5.2 Critiques of Taxonomic Analysis

1. Selection of wrong elements through bias
2. Inflexibility with regard to extensibility or change.

The inherent problem with a taxonomic analysis is the selection of certain features/facets over others, and this may in part be influenced by previous schema, experience and influence of the researcher (the view of data may be co-constructed with the researcher). In some forms of taxonomic analysis, a model well-suited to a certain line of enquiry may have little flexibility or applicability to another and for this reason, I have selected faceted taxonomy to be the most flexible taxonomic approach particularly within an iterative grounded theory framework.

Addressing (1), the initial taxa (facets) will be chosen automatically based on lexical analysis (avoiding bias) and only later grouped/clustered manually around more prominent themes emerging from the interview data

Addressing (2), the approach here will define a vocabulary/palette of processes and affordances under the general principles of a faceted taxonomy which (after Spiteri) is best selected for extensible structures.

A lack of visualisations inherent in faceted taxonomies is addressed by the creation of enhanced concept maps (adapted from Novak).

3.5.3 Case Studies/Vignettes

The Observatory process can be examined through different ecosystem lenses. Research by (Simon 2010) on epistemological social software selects Academic + Business + Government as groups or "tribes". This model has been adapted here to a more generic idea of "Community" scaling up from a "community of one" (autonomy/personal data) through to communities of interest/governance such as Charities and finally to communities of sovereignty/government:

- Academic
- Business
- Community- Government/Public Sector, Charity/NGO, Personal

The list of case study/vignette partners is:

Academic Case	*Business Case*	*Community Case*
Post-disaster Humour (Tsinghua Pilot)	iPhone (Tsinghua Pilot)	Government Corruption (Tsinghua Pilot)
WST	[DataCo]	ANZOG/SA Government
IVOA	Digital Catapult	ODUG/ODI Open Data

Table 3-3 Cases/vignettes

3.5.4 Critiques of Case Study

1. Possible lack of rigour/reproducibility
2. Lack of a generalisable outcome
3. Too long/unreadable – a stylistic issue (!) and not inherent in the method
4. Problems justifying the establishment of causal relationships.

Addressing (1), (3) the project design is to conduct a larger number of smaller more focused cases/vignettes to look for common or distinct themes.

Addressing (2) it should be noted that one cannot generalise from a single case any more than one can generalise from a single experiment and hence the results from case studies are only generalisable to theoretical propositions and not to populations. This fits with the GT approach of building theory up from data.

Addressing (4) the proposed model is segmented into three perspectives and the chosen framework (DNA) has been explicitly designed to allow researchers to reflect different models/assumptions around causality for sociotechnical systems.

Thus the suggested design will develop a theoretical proposition/method against which to design data collection strategies which then inform the theory.

i.e., Taxonomy of function, Taxonomy of exchanges, and finally an exploration of motivation, synthesis/emergence opportunity in the form of phenomenological case studies which will then be compared/contrasted (technically 'Phenomenography') in terms of the motivations for operation in order to potentially theorise a model for interoperation.

3.5.5 Grounded Theory

Grounded theory was developed by Glaser and Strauss and stresses the creation of theories from a "bottom up" perspective: that is, the creation of theories that come from observed data/experience:

> "A grounded theory is one which is inductively derived from the study of the phenomenon it represents."
>
> **(Glaser & Strauss 1967)**

The 'theory' is generated through collection of data/incidents from multiple sources (including self-reflection/self-interview) and comprises the classification/coding of transcripts/experiences through iterative processes:

1. **Open Coding** – classifying to identify key themes and a core problem (the so-called core variable)
2. **Axial Coding** – (added by Corbin & Strauss post original GT) – to identify groups/ structure between the codes/issues
3. **Selective Coding** – restructuring of data around a candidate core theory which then guides the search for new data related to that working theory
4. **Theoretical Coding** – adds contextual/explanatory features to the model to create a hypothesis.

Memos/journals are produced documenting the researcher's thoughts and her/his insights into the relationships between substantive codes and the observed data, and memos are maintained as part of the research records.

The classic model from Strauss & Glaser is quite proscriptive ("no literature review, no talk, no taping") but has been revised in recent years following a professional split between Strauss and Glaser. (Corbin & Strauss 2007) offer a more permissive alternative in which axial coding (creating connections between open coded items) does not assume the lack of preconception/input by the researcher but rather assumes the data (meaning) are co-constructed by the researcher and the participants.

This offers flexibility as the data sources are not locally concentrated around an established corpus of (inter) operating Wos but rather scattered in the evolutionary discourse and the initial steps to create such an infrastructure and the exploratory models and systems that we expect to evolve into more recognisable Wos. Much data is in the heads of researchers and partners who are still considering what they need and what is possible with WO/W^3O.

In this project, I closely mirror the GT approach (though a distinctly constructivist – CGT – interpretation) moving from observable structural features of WO (**what** you operate) to more abstract descriptions (**how** you operate) and from then to a more complex model explain (**why** you want to operate) akin to (Abell 1980).

The absence of existing theory in Web Observatories alone makes a grounded theoretical framework a reasonable choice for this project and the flexibility to adapt and steer the project according to earlier results makes this an ideal approach to adapt to a field of study which is nascent and hence in flux.

Within my chosen CGT framework (Charmaz) key techniques/tools have been superimposed:

1. nVivo for textual analysis and coding
2. Faceted analysis (Ranganathan) for the creation of the taxonomy
3. IPA (Smith) for the qualitative interpretation of participant experience and frames
4. (Extended) Concept Mapping (after Novak & Canas 2008) using the Triz notation (Altshuller 1996) for the presentation/review of resulting models.

3.5.6 Critiques of Grounded Theory

Having discussed the agility and power of the GT method, it should be noted that there are potential weaknesses. (Jones & Alony 2013) and (Bryant 2002) outline a number of potential problems with (classical) GT including:

1. The robustness of the "theory" produced – whether it is indeed a theory in the predictive sense or simply an explanation
2. The practicality of conducting research without recourse to pre-conception (Charmaz 2014)
3. The extent to which the GT method helps to produce a theory which is generalisable (the risk of a purely substantive theory)
4. The risk that without an a priori theory which can be proven/disproven (neither of which constitutes a 'result'), the GT researcher could be left without any significant theory or result from the research effort
5. More generic issues (not specific to GT) around the influence both the research and the researcher have on the project (double hermeneutic) and the desire of both researchers and participants to act in certain ways – particularly when observed (Hawthorne effect).

Addressing (1), (4) the output of the GT process, in this case, does not require a theory per se and would still be a useful contribution if output were restricted to a delineation of the WO and WO types. Addressing (2) the project will adopt a constructivist GT approach (after Charmaz) to not only avoid criticisms of pre-conception but include these in the evaluation of inputs.

Addressing (3) the approach here will define extensible taxonomic structures which may be adapted/generalised to sub-types of WO.

Addressing (5) the position of acting as a researcher/developer with a motivation to "succeed" with a WO has been anticipated and avoided. To avoid social grooming responses (Goffman's front stage behaviour) specific assurances are given to participants around the anonymity of their responses. The idea of the research neither particularly looking for support/critique of Wos is stressed, and the desire for them to evaluate the experience as they feel is most appropriate for them.

3.5.7 Interpretative Phenomenological Analysis (IPA)

IPA (Smith 1996) is a modern qualitative approach to analysing the lived experiences of participants founded on three ideas: Phenomenology, Hermeneutics and Ideography. It borrows from the phenomenological philosophy of Husserl, whose primary objective was to understand the 'nature of experience' itself. The model was further developed by Heidegger and Merleau-Ponty to include a Hermeneutic perspective to both describe and interpret participants' experiences influencing later existentialist thinkers including Descartes. IPA argues for the value of individual ideographic accounts which are substantive (based on real/actual experiences) rather than aggregated, notional or statistical findings and specifically focuses on deeper analysis of fewer participants. IPA has been most notably used in health/social psychology and often in the study of illness or trauma where participants are included based on a shared diagnosis. More recent applications include less inherently negative experiences and non-health issues including the use and impact of information systems – (Vanscoy 2013) and (Evansted 2011).

The key strength of IPA for this project is what (Smith 2009) calls the "hierarchy of experience". Single/discrete events give us data:

e.g., I experience a rainy day in Southampton

but occur within a (irregular/punctuated) flow of contextually-linked events which share/form meaning:

e.g., My experience of British weather.

This creates a structure within which experience is couched as *an experience of something*, contextualised (subjectively) to each person and can be compared with the experiences of others (IPA calls this *intersubjectivity*). Smith defines this as:

"the shared, overlapping and relational nature of our engagement in the world."

And, as I will argue, capturing such intersubjective meanings may be fundamental to an understanding of how/why diverse individuals and groups may choose to engage with (or abstain from) the WO ecosystem even where they may choose to operate a WO in isolation. This intersubjective space, says Smith,

"account(s) for our ability to communicate with, and make sense of, each other.. ".

IPA employs a rigorous analytical process based on seven key principles (Smith 2009):

1. Line-by-line analysis of the claims, concerns and understandings participants reveal
2. Identification of emerging patterns or themes and how they converge/diverge
3. Dialogue/discussion between researchers in multi-researcher teams
4. Development of a Gestalt/structure to illustrate the relationship between themes
5. An organisational structure to trace the development of interviews to themes to models
6. Narratives around the themes supported by visual guides/summaries
7. Reflection on researchers' individual perceptions.

3.5.8 Critiques of a Phenomenological/IPA approach

1. The double-hermeneutic which, in effect, presents the researcher's interpretation of the participant's interpretation may be hard to reconcile with an "objective truth".
2. Phenomenological approaches (including IPA) have "a key commitment .. that analysis should be developed around substantial verbatim excerpts from the data" (Reid et al., 2005) but quoting sources alone more closely equates to describing (classic phenomenology) rather than interpreting (IPA) the experience. Descriptive accounts may offer only substantive rather than generalisable results.
3. The experience of individual participants may be far from representative of the population's experience and lack external validity. Group studies are required to be based on some homogenous experience/process which may be more challenging outside the medical context where homogeneity is typically equated with a medical diagnosis.
4. Phenomenological studies often employ small population sizes suggesting (sic) a lack of rigour/expressiveness. Those which employ larger sizes, however, may do so at the risk of compromising quality for quantity - "Less is more" says (Reid et al., 2005)

Addressing (1), the double hermeneutic is inherent in most forms of qualitative and interpretative analysis and in this case it is the interpreted experience that is specifically of interest. Structured methods for describing and then analysing IPA texts ensure both the participant and the researcher maintain a voice.

Addressing (2), the Interpretative element/extension in IPA extends classical description studies to work through a hermeneutic approach to interpret the texts. Participant material is analysed using an arduous process of passage-by-passage extraction, description, interpretation and commentary.

Addressing (3), the IPA method eschews theoretical sampling (see GT) in favour of *purposeful* sampling in order to represent specific voices or aspects of experience. The participant is thus thought to be representing herself and her own lived experience rather than only her wider group (tribe). In this study, reasonable homogeneity of participants within each group has been ensured using a narrow project focus rather than needing to ascertain that participants from a random group are broadly homogenous.

Addressing (4), IPA studies typically scale from n=1 to n=10 and given the intensive, iterative/exhaustive nature of the IPA analytical process even a single participant may be highly expressive, revealing many theoretical constructs which do not need to be validated through repetition across larger homogenous populations to be considered valuable. (Smith et al., 2009) offers "IPA challenges the traditional linear relationship between 'number of participants' and value of research. It retains an idiographic focus, with 10 participants at the higher end of most recommendations for sample sizes".

Comparative Phenomenographic analyses (across different groups) can shed further light on the similarities/differences between heterogeneous groups.

> "In comparison studies, the exploration of one phenomenon from multiple perspectives can help the IPA analyst to develop a more detailed and multifaceted account of that phenomenon. This is one kind of *triangulation.* "
>
> **(Reid et al., 2005)**

3.5.9 Pilot Studies

A study of three student projects using WO for a research event in China was employed as an initial pilot study and was selected for several reasons:

Close alignment with the main research topic (Web Observatories)
Accessibility of participants for observation (Southampton DTC students)
A fortuitous thematic split of the projects provided a selection of foci (academic, business, government) representing different types of Web Observatory user.
Opportunity to see WO used in vivo and gather positive/negative feedback about WO as an approach.

Confirmatory (longitudinal) material was gathered from six groups at WO event at NUS, Singapore 12 months later.

While additional/alternative sources of WO data could have been (self) generated (as participatory Action Research) by creating a new WO or taking an active part in the activities of one of the WO groups directly it was felt that this might encourage a positive bias in describing the outputs or achievements of the Observatory. As an alternative, I have observed the activities of the SUWO (Southampton Web Observatory) team and contributed some supporting activities which were not directly related to the technology/operation.

3.5.10 Ensuring Quality

In order to cross-check and review the quality of the research output the following checklist has been adapted from (Yin 2008), (Yardley in Smith 2009).

Construct validity **Correct operational measures for concepts being studied.**	*Construct validity flows from the taxonomic structure grounded in the source documents/interviews*
Have the concepts been defined in sufficient detail and do they relate back to research objectives?	*Concepts are extensively defined and related back both to research questions and previous related research*
Have operational measures been identified to match the concepts citing published studies that use the same criteria?	*The measurement here is both structural and operational and each is defined within the taxonomy structure*
Have multiple sources been used?	*Multiple sources have been used for each tribe*
Has a "chain of evidence" been constructed?	*The evidence is the material which generates the grounded theory*
Have the participants reviewed and provided feedback on the model?	*All participants were briefed and were offered the opportunity to feedback. A panel was convened and the comments that were provided were incorporated*
Internal validity **- establishing a causal relationship**	*n/a.*

	This criteria only applies for explanatory/causal studies which is not an objective of this study.

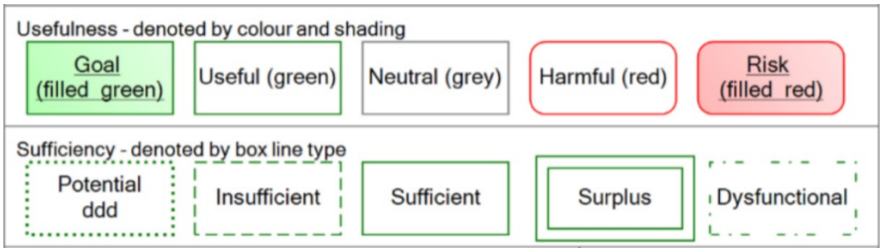
Can pattern matching/correlation be used to support the proposed relationship?	*It has been proposed here that as the number of WOs grows appropriate measures such as facet analysis for contributory factor analysis may be used which is not possible with the current small sample*
External validity **- defining the domain to which the results can be generalised**	*The project is based on observations from multiple cases and whilst produces a substantive model common factors have been identified which span cases*
Can the same results be delivered more than once? Can these results be demonstrated with other subjects?	*The method used is clearly documented and supported repeat experimentation and further research*
Reliability **- demonstrate that results/data collection can be repeated**	*Repeat observations were made at successive WO workshops to specifically test this over a 12-month period. Common themes were identified that are reported here. Also Extensive IPA Transcript worksheets detailing the IPA analysis are included to support the conclusions presented.*
Keep detailed operational notes and logs explaining what was done and how.	*The research model and analytical techniques are clearly documented and several hundred personal logs were recorded during the research process.*

Table 3-4 Evaluating research output

3.6 Tools/Notation

- The primary tool used for data coding and analysis is nVivo (http://www.qsrinternational.com)
- Southbeach (http://southbeachinc.com) based on the TRIZ notation has been used to produce the modelling diagrams which fits visualisation/modelling approaches suggested by both grounded theory and IPA.
- The TRIZ notation (Altshuller 1996) is more expressive than more generic 'boxology' notations supporting entity decomposition, sentiment/perspectives as well as arrangement/connection. For this project, a sub-set (kept as simple as possible) is used here to allow qualitative modelling with a minimum of notation.

Figure 3-3 gives the minimal notation and examples while Appendix A has a full notation guide.



e.g.,

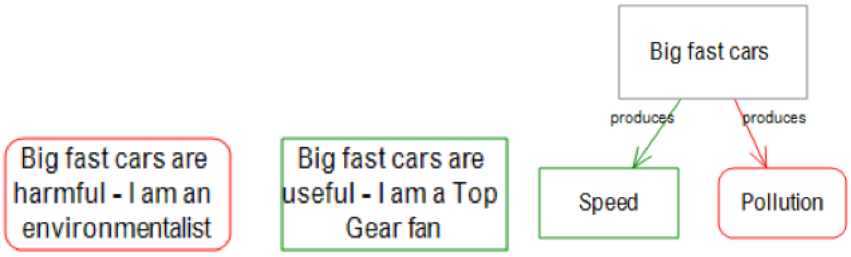


Figure 3-3 Basic notation-oriented from a perspective adapted from www.southbeach.com

3.7 Reflexive Issues

As part of the research process, it is important to consider reflexive issues and to be aware of bias, previous influence and what cognitive science calls "schema" or unchallenged assumptions of truth/value. The following addresses a number of issues from this category.

- As a member of research projects under study and co-author on some of the cited papers, the level of researcher impartiality which can be claimed might be questioned in the presence of Web Science "doctrine" or general hegemonic discourse. In essence: Is there a predisposition to a positive characterisation of WOs?
- Action Research typically encourages researchers to produce a 'better' output through practice. This was not the objective of the research and might also have skewed the analysis to focus on positive results: hence this approach was not adopted.
- Previous training in literary analysis, linguistic patterns and cognitive models leads me to look for patterns and "schemas" (values and meanings which are historically and socially constructed) and often reflected in language patterns. Thus a constructivist approach to research in which participants (co) construct subjective meanings through (shared) experience seems very natural from my own perspective but might dominate alternative methods. In this case, I submit that a qualitative constructivist approach is, in fact, an appropriate technique to investigate the social element of the Social Machine
- Classic Grounded Theory invites researchers to come to the research process unencumbered by previous thoughts or beliefs about the topic. To a large extent, I find this to be unrealistic as captured by Tufte's aphorism "there is a difference between an open mind and an empty head" and this is a known critique of the assumptions/tenets of GT. I am more aligned to a constructivist approach CGT and have adopted this adaptation accordingly.

3.8 Analytic Strategy Summary

The analytical strategy for this project is iterative and generates a grounded theory. Data are extracted/filtered/organised and candidate facets are confirmed through observation/interview/interpretive analysis (Figure 3-4). The process comprises several parts:

- The extraction of key concepts from the discourse/narrative into a facet list
- The iterative grouping/structuring of these facets into a taxonomy/models grouped into Definitions, Narratives and Agency or 'DNA'
- The arrangement of facets into visual templates to determine structure/relation

- The refinement of the models using selected participant interviews and reviews with the community of practice along with visual narrative models.

The goal is to conceptualise WOs using multiple perspectives and elements from which Web Observatories are construed/constructed.

In essence, the model is seeded through textual analysis of *all* interviews and related documents with confirmatory analysis performed on a *subset* of ABC interviews, and finally, detailed interpretation is performed on/between nine purposefully sampled interviewees in the ABC categories.

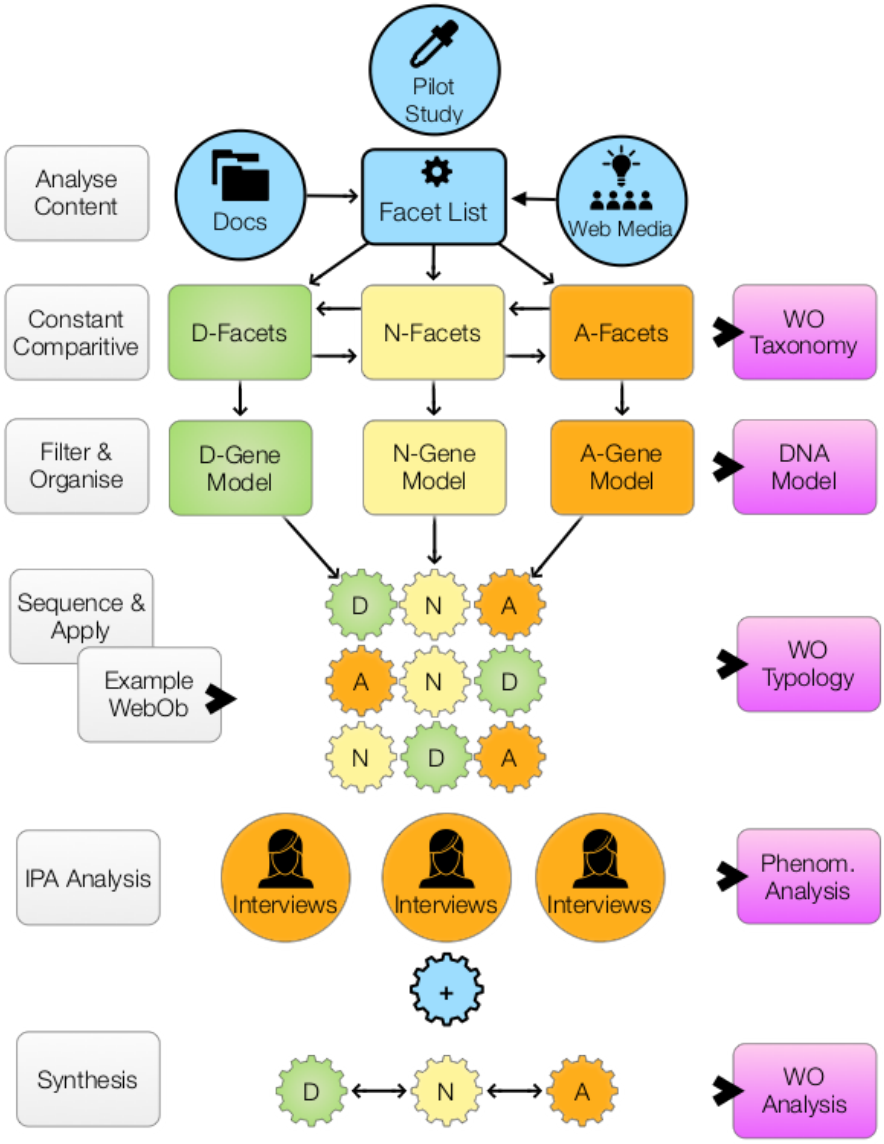


Figure 3-4 Steps and elements of the analysis.

- **Analyse Content**
    - In the Analyse content stage material from published/refereed sources, focus groups/interviews and media from the web (blogs, news, white papers, product information, presentations) are analysed for concepts (facets) that relate to the WO meme.
- **Constant Comparative**
    - Constant comparative is, as the name suggests, not a phase in the strict sense but rather an approach to guiding the development of the emerging grounded theory.
    - Concepts "fan out" in early stages and are then eliminated, merged (de-duplicated) and this process is iterated until no new relevant concepts are forthcoming - the process of saturation.
    - Sources may be consulted more than once and reviewed from differing perspectives as theoretical concepts begin to emerge and suggest themselves as important.
- **Filter & Organise**
    - In this stage, the corpus of concepts is reviewed and grouped by type - in this case by definition/demarcation (**D-facets**), negotiated exchanges (**N-facets**) and Agents/Agency (**A-facets**). This forms the basis of a faceted taxonomy of WO genes though it should be noted that no structure, causality or flow is modelled by the Taxonomy.
- **Sequence and Apply**
    - In this stage, the facets are organised according to the narratives emerging from the documents and interviews. The start of an understanding of context, intention and sequence emerges in individual models for each of the three gene groups D, N and A. The sum of these three sub-models gives us the DNA Model and the order in which they are applied in order to deliver a model, and an analysis constitutes the DNA Method. It should be noted that DNA need not be applied in a fixed order nor should it be understood that each element can be applied only once this giving rise to potentially complex and competing DNA models of the same WO or Social Machine
- **IPA Analysis**
    - In this phase, a detailed analysis of the lived experience of participants was undertaken. The objective is to look in detail at the way participants frame and give meaning to their experience both individually (phenomenologically) and in comparison to other participants interpreting the same situation (phenomenographically)

- **Synthesis**
    - Both the specific experiences of the groups under study in the light of the broader DNA model and the arrangement of factors (DNA AND NDA) can be used to classify the WO.

3.9 Conclusion

A new/emerging concept (where few theories/definitions yet exist) has been chosen for study, suggesting an inductive research approach. This is grounded not only in discourse from published material and in vivo examples, but also in the subjective treatment of the phenomenon itself as reported by users embedded in different social models (frames) of work/usage. A broad set of sources have been married together within a constructivist grounded theory framework to seed an overall model whilst allowing nuanced details of subjective experience (the "voice of the user") to emerge through targeted and rigorous analysis using a proven IPA approach.

This approach complements/extends existing WO research by providing a faceted structure for design, comparison and analysis of WO systems and also by seeking to relate the cognitive and transactional/performative elements of WO to existing WO technical/functional research. It also extends the use of IPA into the study of socio-technical systems and is the first such application in the study of WOs.

Chapter 4: Conceptualising WO

In Short ..

Studies/analyses are reported here which are used to explore the contexts from which users see WO, underpinning an understanding of how WOs are constructed and construed and linking the conceptualisations (meaning) with the process of adoption.

This work supports RQ1: in developing a socio-technical view of WO, RQ2: selecting sensitising concepts to connect to appropriate social models and RQ3: applying elements from the substantive model of WOs to wider insights on usage/adoption.

4.1 Introduction

Despite a body of papers around specific WOs, including SUWO and other emerging WOs within the WSTNet network (Ch2), there is limited work on integrative views of WOs combining both social and technical perspectives. It is not a given that individual exemplars will ultimately lead to a single consensus view or definition of what WO is (or should be). Thus we start with a general view:

- What systems/tools are already deployed in the Web Science community?
- How do different communities characterise WO?
- What impacts the perception/definition of WO?

This Chapter is structured in two parts:

- The first part covers a collection of current systems, documents and impressions from academic sources
- The second uses non-academic sources leading to an examination of the differences in perceptions around WO and how it is framed as well as non-technical factors which influence these perceptions.

Different conceptualisations of WO will be considered and a candidate model to explore the variations will be proposed leading to an initial structure/model for studying the WO phenomenon.

4.2 Conceptualisation & Consensus

The WO is a project proposed and supported by the Web Science Trust (WST) and its network of members WSTNet. It was considered that an initial review of the WSTNet lab tools and systems would provide a useful initial inventory of WO systems and sources including the application areas, data sources and tools offered as (proto) observatories. The dialogue and interactions around WO with different groups (both specialist and non-technical) were also considered for definitions and pro/contra orientation to WO when it became apparent that there was not a strong consensus around the sentiment nor around the definition of WO within the community.

This variation provides an interesting characterisation for understanding engagement and recruitment (technology adoption) for WO/W^{3}O, which will not only depend on the expert/specialist Web Science view but also on perceptions and pre-conceptions of funders, collaborators/resource owners.

4.2.1 Characterisation of WSTNet Community Systems

The WSTNet is a body convened by the Web Science Trust comprising Web Science research groups, several of whom have built individual (proto) observatories and WO tools to support Web Science research. Membership from 2013-2017 is shown Figure 4-1.

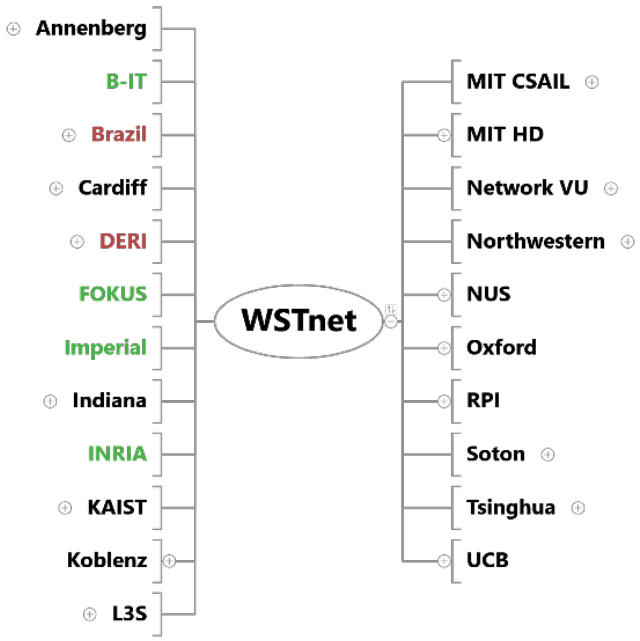


Figure 4-1 WSTNet (Left in Red, Joined in Green) (2013-2017)

The WSTNet inventory was analysed via papers, talks and project descriptions in order to characterise the range and focus of WO offerings (accessed: 4Q2013).

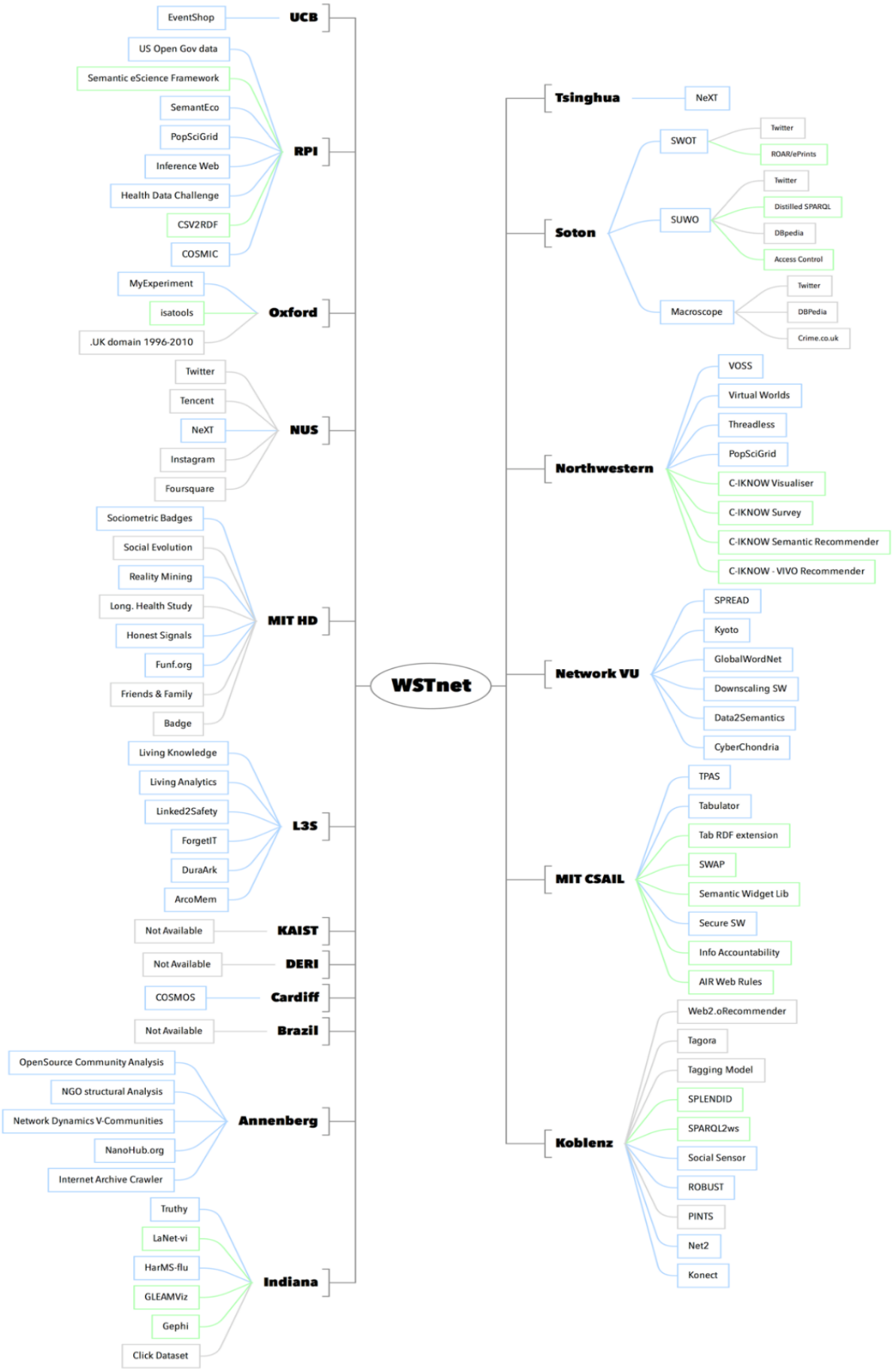


Figure 4-2 Initial 2013 WO WSTNet review - from event research journal (2013)

Figure 4-2 shows a high-level map of systems and tools colour coded to differentiate between the systems (blue), supporting tools (green) and datasets (grey). This data was gathered based on systems and resources that were cited/promoted on the Labs' own websites as of December 2013.

This gave WSTNet a potential "inventory" of 19 locations/consortia (both Oxford and MIT have two labs) of which 15 had produced or were running one or more Observatory systems or Observatory tools. A total of 50 related systems were identified with 22 supporting apps, visualisations or tools. The datasets ranged from archived web pages/websites to research repositories, social media feeds and individual social sensor projects. Given the very early stage of the WO project at which this survey was completed the assumption had been that many of these sources would become attached/integrated to the WO over time.

The most recent review of this list in 2016 (at the 10th anniversary of Web Science) shows a net growth of the Lab community and thus more potential sources and analytical tools available to participate. Relatively few of these sources have become available directly via the WO and surprisingly some of the sources that have been added are from participants outside the WSTNet group (Figure 4-3)..

Web Observatory Site	Datasets	Applications	URL
The Koblenz Network Collection	262	1	http://konect.uni-koblenz.de
Stanford SNAP - Stanford Network Analysis Project	84	0	http://snap.stanford.edu
Southampton Web Observatory	70	39	https://webobservatory.soton.ac.uk
KAIST - Korean Advanced Institute of Science and Technology	0	0	http://ir9.kaist.ac.kr
University of South Australia	0	0	https://observatory.unisa.edu.au
University of Indonesia (UI) Web Observatory	0	0	http://ui.webobservatory.me
IIIT-Bangalore Web Observatory	0	0	http://webobservatory.iiitb.ac.in
National University of Singapore NeXT Observatory	tba	tba	http://www.nextcenter.org/
COSMOS - Cardiff	tba	tba	http://www.cs.cf.ac.uk/cosmos/
SONIC - Northwestern	tba	tba	http://sonic.northwestern.edu/
RPI Web Observatory	tba	tba	https://logd.tw.rpi.edu/web_observatory
Indiana University Truthy	tba	tba	http://truthy.indiana.edu/

Figure 4-3 WO Dataset availability. Source: http://index.webobservatory.org/ (2017)

The WO portal itself however has undergone substantial improvements and greater focus on applications vs data sets has been observed Figure 4-4

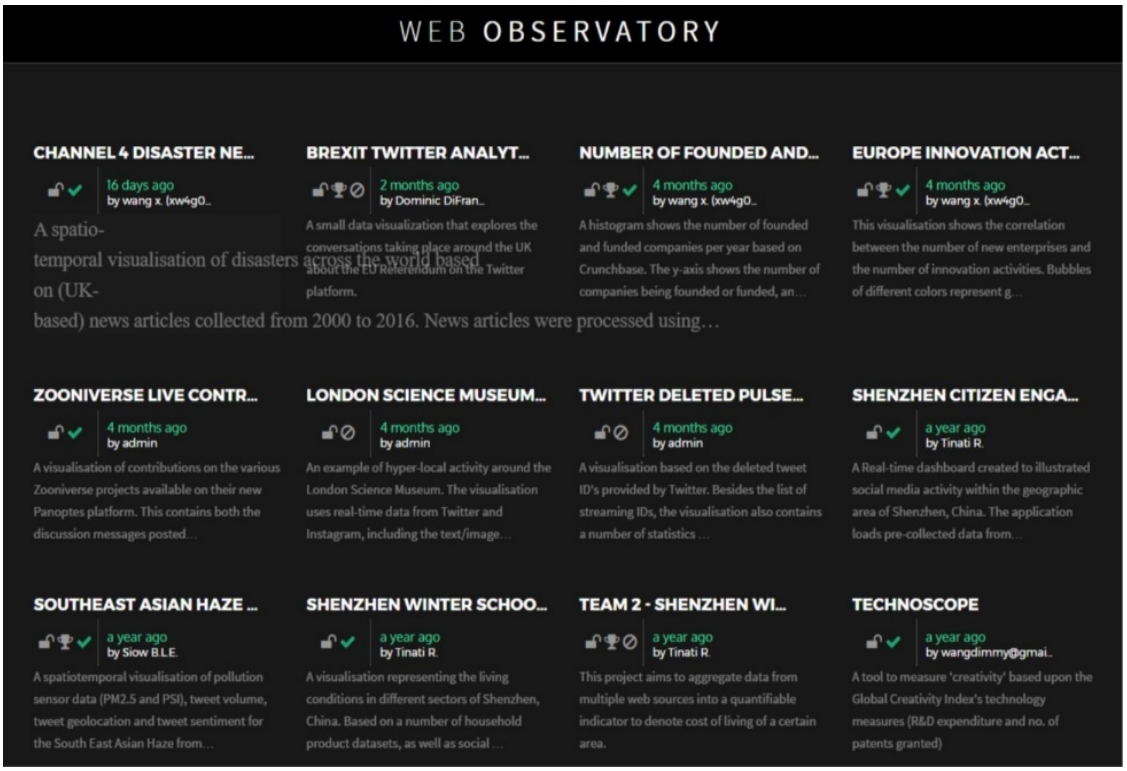


Figure 4-4 WO Application registry. Source: https://webobservatory.soton.ac.uk/new/apps (2017)

The pattern of engagement over the period 2013-2017 suggests that:

- WO owners do not automatically participate in the eco-system and that those who participate do not necessarily operate WOs.
- The participation in WO may, in fact, be drawn from one/more distinct area (data, analytics, tools) and not necessarily from fully integrated WO application offering a fully built-out solution.
- Interest/participation may be focussed on specific applications rather than data sets
- That this ecosystem of labs comprises the groups most pre-disposed and most *technically able* to integrate WO systems/sources and comprises many systems/sources that might technically have come together under the W^{3}O banner. In practice, however, only a few (primarily driven by the availability of Southampton WO templates) have, so far, found a *reason/context* to do so.

Technological skill and compliance with standards are thus necessary-but-not-sufficient for an eco-system of WO systems to flourish and as we note from (Ch2) incentive structures may be highly relevant. Considering that conceptual differences within the community might also be relevant for participation including what the WO itself was perceived to be and the expected benefits for engaging with WO, and so I next considered the extent to which the WO idea/meme was broadly understood/shared.

4.2.2 Characterisation by the broader academic community

Pursuing an explanation for a limited engagement with WO even within the dedicated WO community, I polled WSTNet colleagues and other interested parties to investigate the similarity of understanding of the WO concept itself amongst the academic community. Through these exchanges (some formal/some informal) I noted an interesting series of ad hoc 'pronouncements' on the topic of WO. After conflicting definitions (even within the same groups) had stimulated interest in this phenomenon, I logged some these comments (informally and anonymously) in a research journal. These were broadly classified as *challenges* (weak vs. strong) and *endorsements* (weak vs. strong). Many were concerned with the definition, nature and existence of WO.

Weak challenges were seen as a clarification challenge based on "definition by extension" (akin to John Locke's idea of combining/abstracting) which allows two or more known concepts to be combined in a way that extends, exposes or defines a previously unknown one.

e.g., Thinking of an eel as "like a snake that lives underwater".

Weak challenges were inquisitive in nature and *checking if* a WO was 'like-something-else' perhaps in an attempt by the speaker to perform so-called *'gisting'* - that is, to confirm the top-level ideas of WO without the knowledge (or perhaps the desire) to engage in more detailed differentiation.

- [Web Observatory? That's like analytics on the Web, right?]
- [Observatories are to do with Tweets or something, aren't they?]

Strong challenges were also related to WO being "like something" but were more aggressive (i.e., more emphatic/declarative), e.g., casting WO as "old-wine-in-new-bottles" and thus *proposing that* a WO was largely synonymous with another idea, completely subsumed by an existing concept or otherwise superfluous. This may involve a distrust of a new concept (the "Magic Bullet" characterisation) and/or an unwillingness to dilute, repeat or otherwise unpack/restructure [30] other existing concepts with which the speaker is already familiar either personally or through group competitive behaviour or tribalism.

- [Web Observatory? That's just Web analytics!]
- [Observatories are basically just Twitter harvesters!]

[30] Reminiscent of competence-destroying behaviours described by Adler

Weak support came in the form of a *social endorsement* around the idea of an Observatory (on? about?) the Web but was typically characterised by a lack of reason or evidence for supporting the concept. Such support may simply be 'social grooming' (part of Goffman's 'front stage' behaviour) and/or an overture to learn more.

- [Web Observatory - yeah, I don't really know anything about that: It sounds great though!]

Strong support most commonly as a *recruitment endorsement* - the assertion of WO being a solution for an existing (previously challenging) problem in which the WO is drawn into the pool of resources to address the speakers challenge(s)

- [Observatories are just what we need to solve <insert problem>]
- [We can combine <x> and <y> with an Observatory]

1. We've been doing this for years
2. We tried this and it doesn't work!
3. We used to call this <Grid> now its WO
4. We are really exited about WO.
5. Isn't this essentially <CKAN>?
6. Don't really understand but sounds great
7. It's just a Twitter harvester!
8. We've got an Observatory already and ..
9. Sounds a bit like surveillance.
10. We think WO is the solution for ….!
11. I don't see how this works…?
12. I don't see how this CAN work!
13. This is going to make my job easier.
14. WO will help give us access to <Data>
15. We have the datasets - what do we need WO for?
16. If we can combine <A> and <B> we can …
17. WO is going to transform the way we work

Figure 4-5 Strong and weak engagements around the nature of WO.

Figure 4-5 clusters the types of engagement and may form an interesting seed record of how operation and interoperation may be fostered and objections overcome for the adoption of Web Observatories.

Initial insights from this suggest:

1. Individuals expressed a wide range of responses based on varied interpretations or contextualisations of WO not necessarily accounted for by a difference in technical understanding.
2. The difference in magnitude between (non) engaged parties seemed to be the positive/negative association (or 'framing' as Goffman (1974) terms it) of WO with a specific solution/application.

In terms of informing the WO engagement process these findings suggest:

- Moving disinterested parties to engage may require a solution-oriented (vs. technically-oriented) approach or that different participants may orient towards either the technical solution OR the application.
- To engage those positioned negatively to WO as a solution (due to competitive reaction) an inclusive/collaborative approach (WO and competing technology as part of a wider solution) may be required
- To engage those concerned about privacy/surveillance a clear ethical policy statement would be required.

The apparent lack of significant direct experience/evidence required to make an assessment (positive or negative) of the WO was notable, and thus I considered the extent to which people were reacting instead to the assumed/inferred conceptual meanings of "Web" + "Observatory" rather than an "official" definition. The idea that detailed knowledge of the WO itself was not necessarily a factor in characterising or evaluating WO prompted a follow-up confirmatory experiment to entirely exclude the effect of knowledge about the system and to consider other social/contextual factors about respondents.

4.2.3 Characterisation by the general public

A questionnaire was deployed to 100 anonymous non-specialist participants asking if they could provide free-format definitions of several terms designed to be semantically similar to the name "Web Observatory" (i.e., using the pattern ["Web" + 'a-verb-meaning-looking']. Based on Chomsky's idea of 'deep or D-structures' (which communicates a fundamental understanding.) vs. 'surface or S-structures which sometimes offer subtle but important variations on the meaning/understanding). The intention was to investigate whether the only real term (Web Observatory) had any more public recognition than the fake terms and to note any correlating factors about the participants themselves (age group, gender, work situation) since actual knowledge of the Observatory was intended to be absent. The key objective was to investigate whether different groups had specific conceptualisations around the term. The sample comprised:

- Young 14-17 (assumed extensive digital experience, pre-work)
- Adolescent 18-24 (assumed extensive digital experience, studying/working)
- Adult 25-34 (assumed extensive digital experience) working participants
- Mature Adult 35-44 (assumed some digital experience, working)
- Late career 45-54 (assumed some digital experience) working/post-work
- Retired >55 (assumed some digital experience post/work)

The characterisations are suggested as typical rather than definitive for the age-group (e.g., 50% of the >54 group specifically identified as “retired”). Work status and classification (Business, Government, Academia) was also captured to determine if occupational framing indicated a difference in sentiment or interpretation.

Participants (n=100) were asked to confirm their age range/gender and employment status and to offer a definition of the following:

- Webservatory
- WebViewer
- Web Observatory
- WebScope
- WebWatcher

and thereafter to rank them from most positive to most negative and comment on the reason why the top/bottom was picked. Finally, some beliefs around Web Observatory were surveyed.

<5% of the responses were considered 'spoiled' as determined by a manual check of nonsensical responses.

Participants were all based in North America (to reduce potential cultural differences) and evenly split male/female with approx. 20% in each age range.

Overall Findings

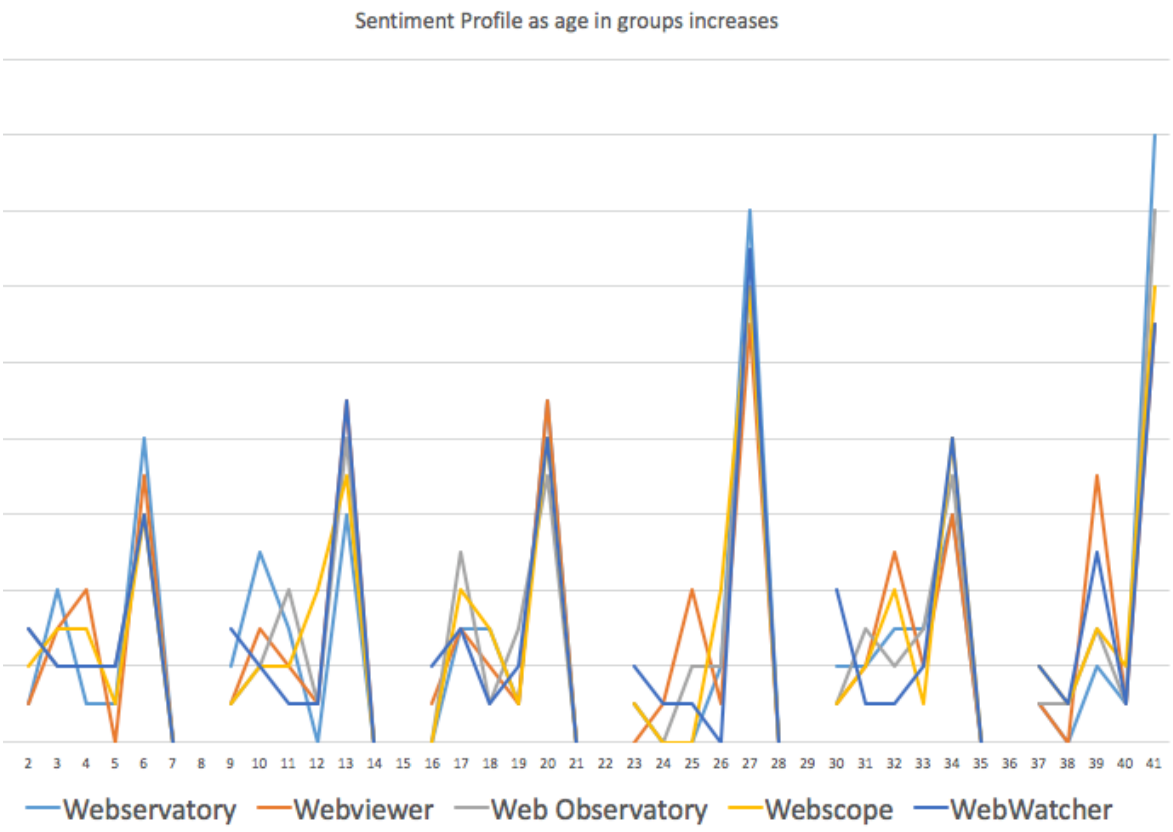


Figure 4-6 Results clustered by Age

Full details by group are given in the Appendix.

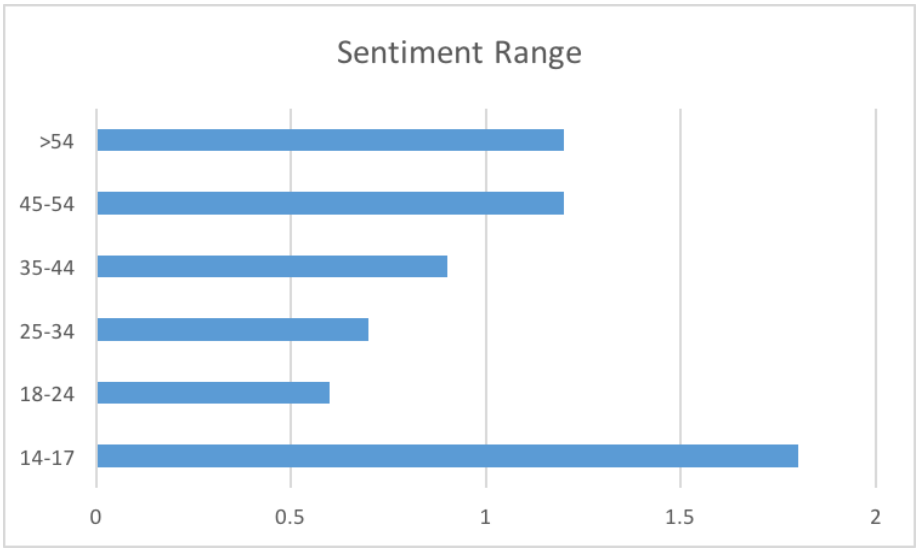


Figure 4-7 Overall Sentiment interval by group

Discussion

The survey population were asked for definitions/reactions to a number of apparently similar terms based on the pattern ["web"+<a-verb-meaning-to-look>]. Given there was a deliberate absence of information about any of the terms or about differences (4/5 terms were invented) we might reasonably expect feedback to be broadly similar across all names and all age groups -

which was *not* the case. There is no attempt to derive specific insights per se for the definition of WOs from this work, but we can consider how *non-technical / non-informational factors* may (must?) account for the observed variations in perception. Little/any of the variation can derive from technical knowledge about the terms themselves (given none was provided), and only one participant out of 100 appeared to use a recognisable definition of WO - presumably using Web search despite instructions *not* to do so.

There are two notable results:

1. Both the *ability* to answer (vs. 'Don't Knows') as well as the *sentiment* shifts as we move from younger to older groups. This may be explained by diminishing technological exposure/experience: the positive sentiment apparent in younger groups familiar with (and apparently more trusting of) Web technologies are replaced by strongly negative ratings from older/retired groups. Older (less Web experienced) groups seem less willing to assume they know (or trust) than paradoxically younger/more experienced groups *noting that there were no correct answers here* - only attempted responses or not. This may have implications for the timescales in which we may expect Web Observatories to become known and acceptable in different social groups and may steer outreach work towards specific age groups.
2. The age breakdown of the sentiment interval vs. the overall ratings is notable. The overall interval rating (0.4) masks much wider sentiment intervals at both ends of the age spectrum (1.2 -1.8) vs. the participants with experience of digital technology in the workplace. This may suggest that the working age groups (vs. pre-working and post-working age groups) are conceptualising these terms differently - indicating social, or perhaps occupational factors are being used to frame the ideas. The level of socio-economic and cultural diversity is unclear but all participants were deliberately recruited from a single country – in this case the U.S.A..

The purpose here is to explore the possible impact of non-technical factors on the perception of WO, i.e., how they are characterised by potential users. This study suggests that wide variations in characterisation can occur mediated by factors other than information about technical functionality which was effectively removed as a factor from this survey.

4.3 Implications from special and non-specialist review

This short analysis raises some interesting insights for the research:

1. That non-specialists and even stakeholders can hold divergent conceptualisations of the nature and purpose of WOs while apparently agreeing at a basic level on what the object is. This suggests that there are views or perspectives on the meme that allow for agreements and others that diverge, suggesting the possibility of a flexible cognitive basis for understanding WO. Where there is a lack of direct experience (few, if any, of the individuals had actually worked with WOs), this leads to interpretation and understanding by analogy, metaphor and extension. Our individual perceptions/meanings are potentially "framed" (in Erving Goffman's parlance) in terms of concepts that we pre-select and contexts with which we already engage. This implies the possibility of divergent or only partially-shared definitions and conceptualisations in more diverse groups and greater intersubjective agreement in more homogenous groups.
2. That the WO concept is both engaging and polarising. The lack of (any) in vivo experience does not prevent either strong support of, or attack on, the WO concept. This may be reflective of perceived status (turf) wars, privacy concerns or reactions counter to (or in line with) local cultural hegemony - noting that these are all *social*, not technical factors.
3. That this polarisation surfaced the idea of novelty/innovation vs. "old wine in new bottles" is a key theme which needs to be addressed if WO→W^{3}O adoption is to be encouraged.
4. The application/impact of WO (the perceived "so-what?") may predominantly be based on usage at work - we saw larger numbers of pre/post work participants failing to conceive of a purpose or application for many of the test terms. This informs a theory that a critical frame for W^{3}O participation may be a general occupational in combination with other layered cues/frames (Goffman calls these *Laminations*) relating to job role and goals.

4.4 Conclusion

Understanding and managing varying conceptualisations of WO and the resulting pro- or contra bias in the perception of new technologies may play a part in successful innovation adoption for WO→W^{3}O. Innovation/adoption theories (as discussed in Ch2) typically relate to attitudes to new models in terms of resistance (Ram), pursuit of benefits (Rogers), contextualisation/adaption into daily practice (May)) and reframing perceived benefits (Christensen). An understanding of these adoption processes with relation to different stakeholder perspectives may be at least as critical to WO adoption as the adherence to technical standards.

Given the ad-hoc nature of the support/challenge comments (noted down informally after the event), it should be stressed that this offers an unmoderated, subjective impression rather than a rigorous analysis of intent/context. However, since very few WOs are yet operating in vivo, few (if any) of these comments can have a basis in extensive experience with WOs. Hence, a

subjective/contextual element must play a part in the differences leading to the traditional insight that we may:

> ".. see the world (or the WO - Ed.), not as it is but as we are".
>
> **Talmud (traditional saying)**

In the next chapter we will review the construction of a candidate model for WO elements/facets through a process of content analysis and refinement.

Chapter 5: Seeding the WO Model

In Short ..

Studies/analyses are reported here which are used to inform a multi-dimensional seed model of WOs comprising functional, narrative and agentive elements. This underpins an understanding of how WOs are construed and constructed. These pieces combine to “prime the pump” for RQ1: identifying constituent elements of WOs, RQ3: organising these elements to form a substantive model of WOs and also RQ2: the extent to which social perspectives and narratives may vary across different user groups or tribes framing perceptions of WO in addition to technical factors. They create a “straw man” which is validated against later participant interviews and observations.

5.1 Introduction

This section reports on the taxonomic analysis and experimental "straw man" models/visualisations based on the taxonomy which are refined and finalised in Ch9 giving an insight into the iterative, grounded theory process and the development of the final theory.

Proxy systems/concepts related to WO were selected from a broad range of academic, news, project and media sources. The goal was to refine the definition of WO from a poorly differentiated comparative/analogy-driven model (“WO is like a ..” or “WO is better/worse than a ..”) seen in Ch4 to a more nuanced and differentiated model with specific WO features and different perspectives. These proxies were not intended to deliver a final definitive model of WO (though (Paukkeri et al., 2012) have attempted a similar task). Rather they are a "straw man" from a set of seed concepts - to be evaluated, refined, replaced, iterated and extended over time through the process of engagement with the community of practice as described in later chapters.

The raw concepts also need to be suitably visualised/arranged to aid in understanding structure, and several initial visualisations are presented below which characterise each aspect of the building blocks.

The sequence followed for elements of the model is:

Search→Select→Group→Visualise →Arrange→Review→Revise

and reflects major groups including functions, processes and people.

5.2 Sources, Searches and Scope

Conceptual "seeds" were first established via an automated lexical analysis (text mining) using the nVivo platform. By determining:

- Which concepts were presented per se (at least in the top 1000 concepts) for later manual checking rather than (naively) assuming frequency to be the final/only marker for importance.
- Given the document sources were pre-classified by Tribe no automated clustering or co-occurrence analysis was attempted as this was not felt to be meaningful. Determining that particular types of content occur in certain types of document (which were selected on the basis of that content) appeared tautological. Super-ordinated/abstract notions however were manually generated and noted for co-occurrence across tribes.

Tentative "straw man" models were produced which were manually reviewed, de-duplicated and clustered into conceptual groups/models which were later confirmed/revised through observation, questioning and manual review.

Reference papers on Virtual Observatories and Web Observatories were mined for secondary sources which, in turn, suggested further search criteria. While video/media sources could not be included directly in the lexical analysis, more than 50 video presentations were consulted as an additional guide to search terms, documents and key topics. Selected video sources were subsequently analysed (coded) using nVivo and formed part of the overall analysis in the broader research. Initial documents were reviewed and second level references to related documents were followed in several cases. All documents were desk-checked for relevance before including in the lexical analysis.

WO is a recent concept, and so the focus on documentary sources was mainly focussed on dates in/after 2012 which was the year of the first published use of the term "Web Observatory". Some exceptions were made for older academic work (e.g., on virtual astronomical observatories which inspired the WO concept) that were directly referenced in WO material.

Text mining typically uses frequency as a simple proxy for importance and in this case limited the search to the top 1000 most frequently occurring entities/concepts. Additional "gisting" is available using fuzzy searches which employ related/derived words, nearness/proximity and synonyms). Visualising query sets using word clouds/trees and cluster maps allowed a "fan out" from the automated concept list in order to determine further connected search terms and to work with more conceptual terms (e.g. From "Twiki" to "Collaboration" see Figure 5-1)

Frequency analysis using text mining as the only measure of importance is thought by (Bazeley 2013) to be too simple an analytical measure and she specifically warns against attempting interpretative analysis with such tools. It is however, she suggests, a sufficient (and efficient) filter for exploratory analysis to pre-sort and visualise concepts without more computationally expensive and labour-intensive methods - particularly where the results will be cross-validated. Visualisations are a key tool here, and while sorted frequency lists alone might not have highlighted the term 'Twiki' (a collaborative Wiki application) in the IVOA corpus query, the visualisation makes the existence of, and link to collaborative software systems in this context more apparent.

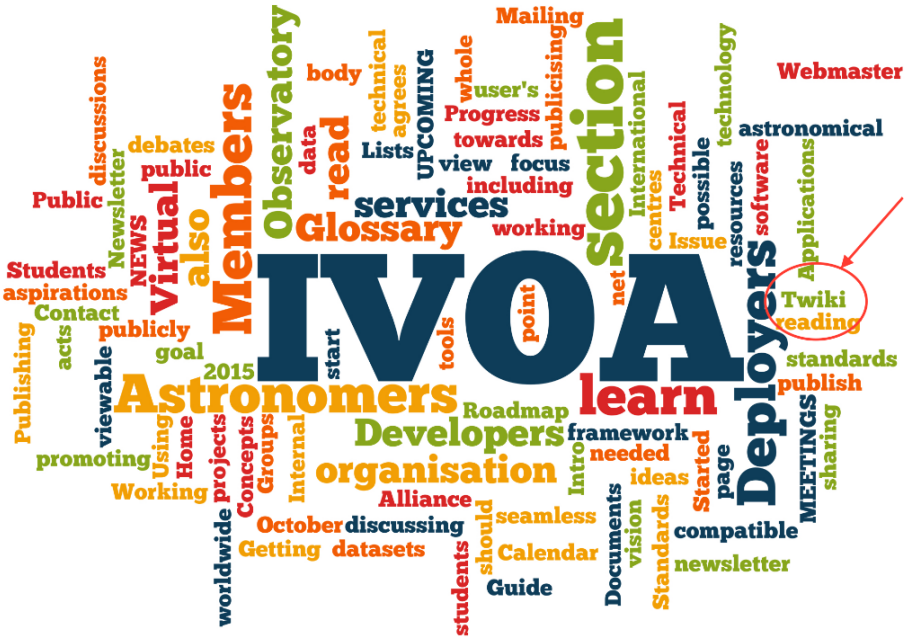


Figure 5-1 Word Cloud from IVOA.net to "gist" key concepts

Broadly speaking automated analysis has been used to reveal the existence of factors/facets with confirmatory analysis and structuring of models employing manual methods to determine importance and relationships.

Other issues apparent from text mining are that stylistically, language and expression are typically formalised and occupationally (tribally) contextualised so that problems/doubts, challenges, inconsistencies and other issues may be strongly under-represented in 'on-the-record' documents (Goffman calls this ‘front-stage’ talk). Compare this to interviews/focus groups in which participants may expect anonymity and thus speak more freely (Goffman calls this ‘backstage’ talk). Thus, the bulk of the deeper analysis has been performed on the participant interview material with the textual analysis used only to frame and seed the structures/facets.

5.3 Findings: The Taxonomy

The source data (content analysis and case study interviews) generated five top-level groups (Figure 5-2), 22 2nd-level groups and 245 foci (or features) of the Web Observatory, which are chosen up to a cut-off point. They include generic types of facets but to exclude exhaustive examples of those facets as end-node values - i.e., including "fruit" but excluding "apple", including "Social Media" but excluding "Twitter". The full taxonomy is shown in the Appendix.

e.g.,

- That data in the WO has a type [Facet>Data>Data-type] is included.
- That Data-type itself has a range of values [Facet>Data>Data>XML] or [Facet>Data>Data>JSON] is excluded from the core taxonomy

since these sorts of enumeration are thought to be more ephemeral in nature.

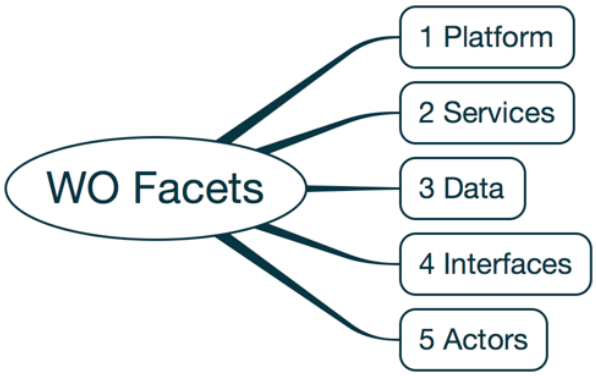


Figure 5-2 Level 1 WO Facets

This structure not only finds parallels in Whitworth's model of STS but also respects the mutual exclusivity and jointly exhaustive criteria required in the taxonomy literature.

5.4 Discussion

In terms of the design, the cut-off between broad groups and specific values in the taxonomy balances issues of permanence for data, sources and functionality for an evolving WO against reflecting how users will want to query and access data and services. From the perspective of a researcher's objectives it seems reasonable that users are less likely to query any/all data from a WO simply because it is stored in, say, a JSON format than to query data on a specific search term (Football) or from a specific system source (Twitter).

There has been consideration (Tiropanis et al., 2013) on the nature of topic-centric queries (i.e. Things-about-football-from-anywhere") vs. source-centric queries (say "anything-that-comes-from-Twitter") and the balance between the two needs to be supported by an effective taxonomy without defining WO simply in terms of its content/sources.

Thus I have chosen to model [Data] and [Services] at a high level using ideas like "online social network" [Data>Data_Source>OSN] with a series of end node values underneath or each network supported. Thus if sub- facets such as, say, Twitter_feed [Yes/No] were to be replaced by NewService_feed [Yes/No] that this might have fewer implications than modelling a specific function or data source at the top level.

The [Interfaces] facet flows from having a system and needing the data/information to get in/out via applications or APIs. Perhaps counterintuitively, the total number of facets may, therefore, be of only minor interest vs. the broader groupings/structure since the total facet space almost certainly constitutes a moving target with WOs constantly in flux iterating from one state to the next as new sources/features are added.

In addition to the sources/content they deliver, WO retains a facet for [Platform] features which do not currently include examples of implementation technology but does include platform objectives (such as cost or performance). The facet around [Actors] follows from the idea that the system has an objective although different users are likely to be addressing different projects with varying objectives, and this reaffirmed the inclusion of the need for collaboration and orchestration within an Observatory.

Overall the Taxonomy affords a useful structure to enumerate features or stakeholders but fails to be sufficiently expressive when looking for relationships (other than similarity) between facets and for arrangements other than hierarchy. Little/no visualisation beyond the structure itself is possible and, in particular, moving from a list of Actors to their motivations and how such motivations might play out in sequence quickly seems to exhaust the expressive potential of the taxonomy alone.

Ultimately the community of Observatory builders will determine how accurate this classification structure may be but in considering the criteria specified for evaluating faceted classifications per se I would offer that (Spiteri's 1998) criteria (derived from Ranganathan and the CRG's criteria) have been met here:

- **Differentiation** – Top level facets are fully differentiated
- **Relevance** – fully met. While the example end-nodes on, say, OSN sources may be irrelevant to a WO that doesn't implement them the broad grouping is nonetheless core to the classification.
- **Ascertainability** – largely met (e.g., platform objectives such as "scalability" are poorly defined in the literature)
- **Permanence** – fully met – while end-node sources/topics may change we feel the top-level facets will be stable.
- **Homogeneity** – partially. Certain classes are grouped rather than related by type/topic data. Metadata may be homogenous (or converted to such) within a particular classification, but potentially sources under the OSN classification may not be *functionally* equivalent (different content format, date/time format, (no) geo-location format, etc.)
- **Mutual Exclusivity** – partially met. Interfaces may be thought to be a subset of Services, but we have chosen to pull this out separately for the purposes of understanding WO usage.
- **Fundamental Categories** – fully met. None of the facets function as more general facet of the others

Finally, while it clear that "data" may not always imply a "service", services often do imply some underlying data which they deliver or from which they are driven: examples such as Provenance and Analytical services are cases in point. In order to avoid classifying all data as a type of service, I have elected to distinguish further between the types of data as:

- Underlying (topic) data (akin to that mapped by Dewey or LoC)
- Derived (calculated) data
- Simulated data and
- Metadata

so that we may make this distinction between the use and analysis of different data sources in the understanding that access to these data may be via services listed elsewhere in the taxonomy.

5.5 Implications

I have shown that WO may be classified via a flexible faceted approach allowing for extensibility not only within the definition of what Observatories are but also in terms of a social perspective for the Actors on what they are for. (Bowker et al., 2010) argues that we should avoid simply placing users "on top" of technical features as though one is independent or separate from the other and (Star & Ruhleder 1994) paraphrases (Bateson 1972) claiming:

> ".. that infrastructure is fundamentally and always a relation, never a thing. --

A perspective addressing *what* people are trying to achieve is likely to promote a broader understanding of how WO's are developed and extended rather than starting from a purely physical/technical perspective which describes *how* they do it. A question remains however over how well WO→W^{3}O is captured with the taxonomy particularly since while W^{3}O requires certain features to be present it is more than the existence of these features alone which allows W^{3}O to emerge. This seems to be primarily a taxonomy of WO but not necessarily of W^{3}O.

Naturally the efficiency and scalability of Observatories relies heavily on sound technical/architectural choices for storage, querying and analytics. In terms of a functional definition, however, participant interviews suggested that, like users of electricity, Observatory users may be less concerned with *how* their service is generated than with the fact that it is reliable (trustworthy), available and compatible with the devices they want to use.

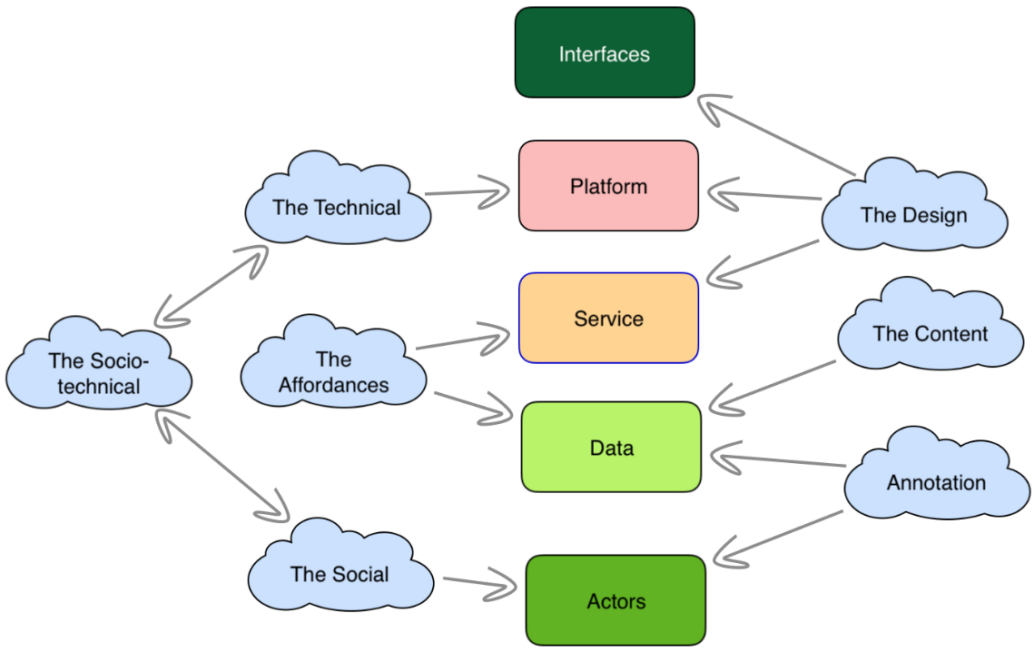


Figure 5-3 Varying perspectives associated with raw structure

Figure 5-3 shows the implied perspectives and socio-technical relationships that can be inferred from the taxonomy but are not shown by the taxonomy itself. This appers to be an inherent weakness of a purely taxonomic approach.

(Bowker et al., 2014) calls every interface a "relationship" and indeed the interface, in this case, may correspond not only to a flow of research data but also of services, communication, consensus or payment/exchange which are represented by the collection of notional processes or exchanges. These structures and exchanges are best visualised using other methods, and visualisation/arrangement is explored in the next section.

5.6 From Taxonomy to Modelling the data

As the facets were extracted from the seed data as features (captured as nVivo nodes) they were listed and arranged into groups. Noting visualisation as one of the weaknesses of faceted classifications, several graphical depictions were attempted which, while showing the relationship or organisation of components, generally fails to offer a notion of external boundaries and a distinction between function (what) and process (how).

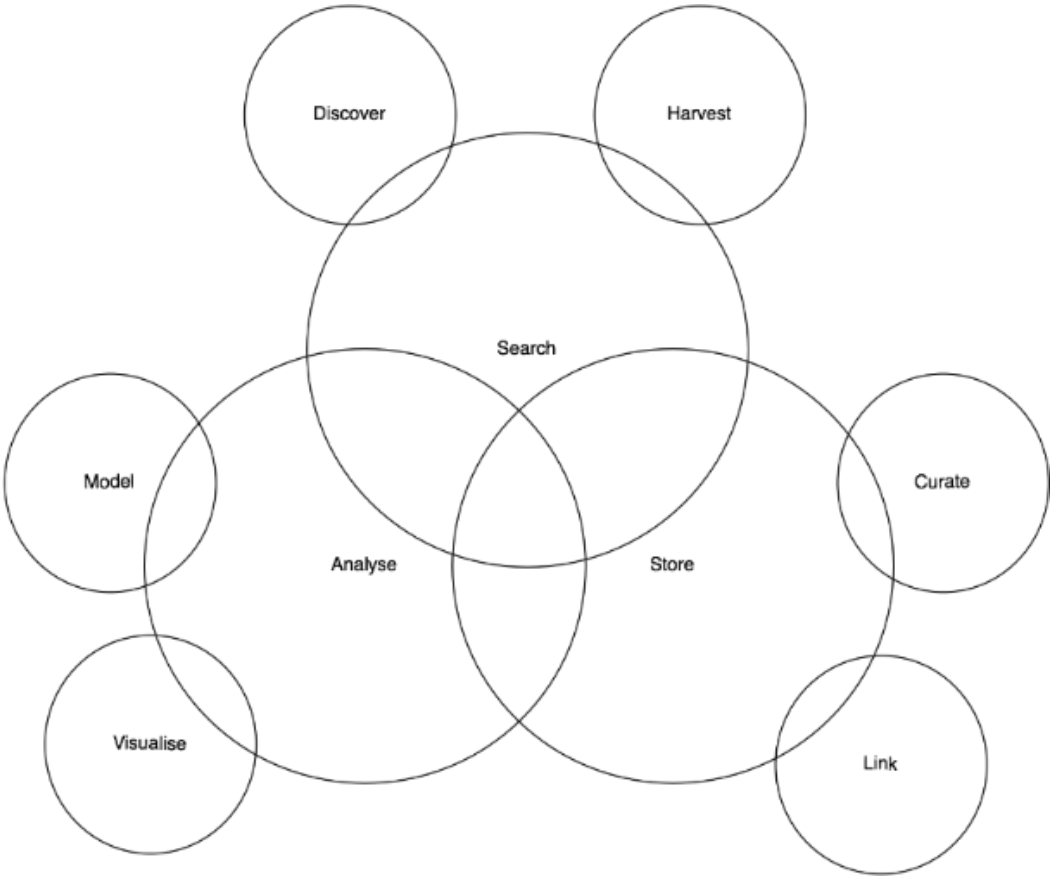


Figure 5-4 Early (unsuccessful) graphical rendering of WO facets

5.6.1 Functional View

A more detailed concept map (Figure 5-5) depicting the grouping of facets relating to the physical arrangement (Design) of the WO was sketched. It is rendered in an implementation-neutral fashion and pre-supposes no particular hardware, storage or networking approach.

This visualisation appears more useful in terms of boundaries and implied flows and was then rendered as follows. Review feedback was obtained from a selection of technical participants.

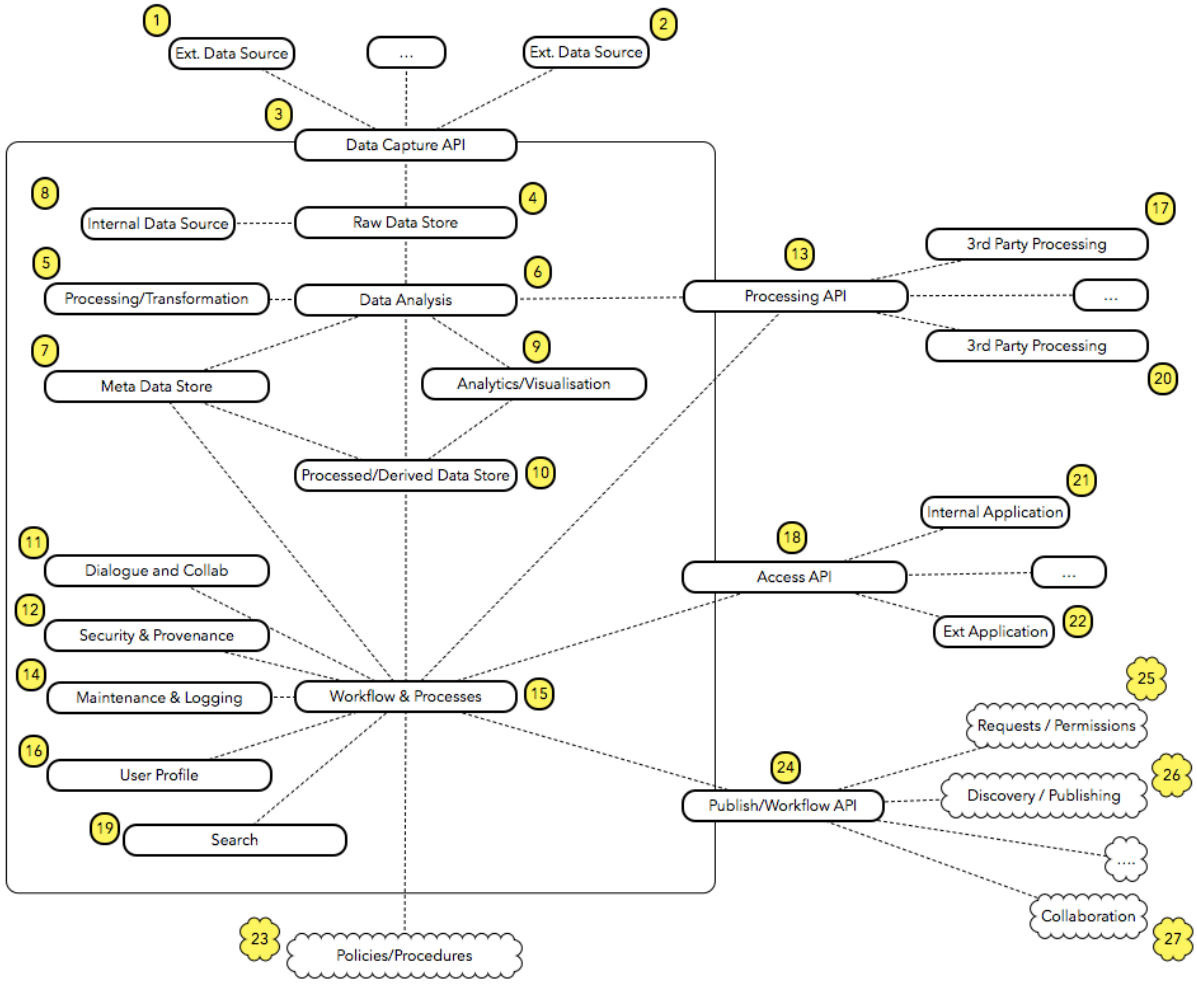


Figure 5-5 Early Concept Map of Observatory Structure from (Brown 2014)

This visualisation is similar to a concept map after (Novak & Canas 2008) but extends the notation with 'clouds' indicating manual/human-oriented processes. This approach was further extended (Ch9) into a colour coded and annotated modelling approach using Triz notation

Beyond the assertion of what is inside/outside the system, the layout is not intended to imply a physical design but is a representation of an Observatory in terms of which concepts that it could support (rather than how this would be achieved) or who would use/implement it. It can be thought of as an extension complementing the taxonomy through the ability to visualise how

certain taxonomic features may relate to one another. It should be noted that not all elements of the taxonomy are depicted here (only the notional affordances).

In each concept map, the border represents the scope of control/authority of the Observatory owner and thus elements which cross this boundary imply "flows" requiring an interface of sorts (manual or automated).

This format (with explanatory keys/descriptions and supporting videos) was used to test the model with colleagues and participants. Comments and feedback were incorporated to refine the model, and this depiction was found to be accessible for participants.

5.6.2 Process View

Facets were also extracted for dynamic (processing) rather than static (architectural) elements of the WO exchange. These seeded a process catalogue which was initially unstructured and unsequenced but came (later) to be based on the notion of interlocking/sequenced phases of discovering, assembling/processing and then executing analytical and publishing services.

The initial unstructured set of processes Figure 5-6 were identified as follows:

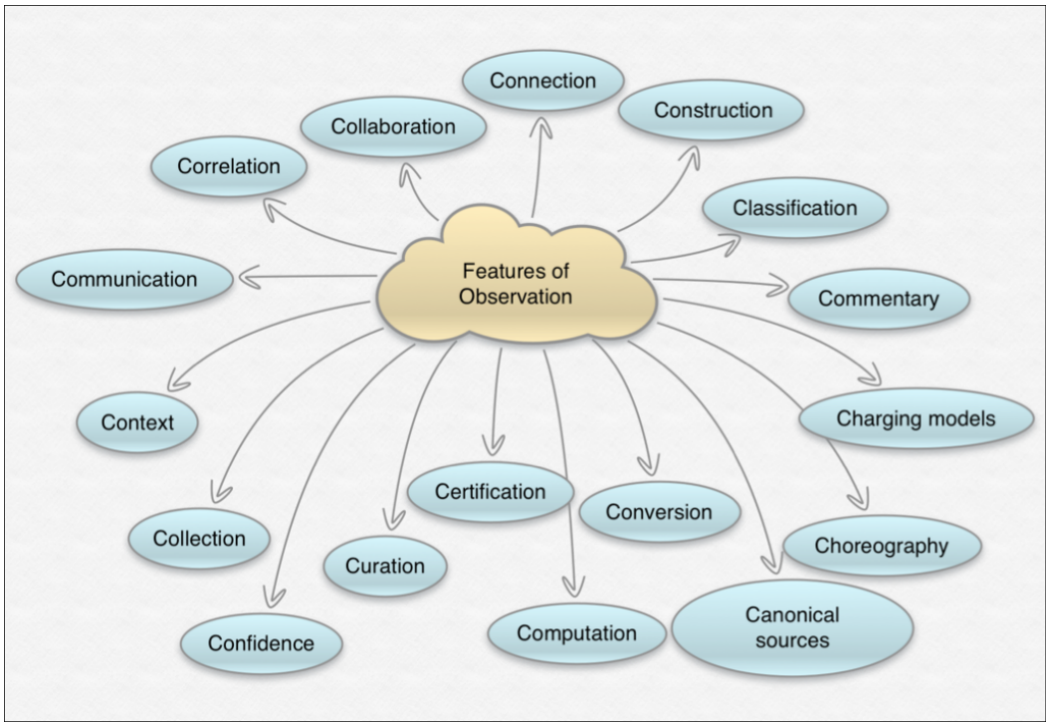
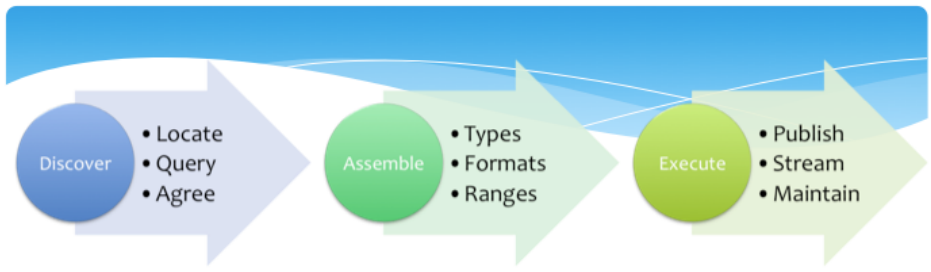


Figure 5-6 Early unstructured seed process catalogue

A detailed documentation of each process (in BPMN or web sequence format) would be straightforward (indeed some processes were trialled). However, while the processes/exchanges remain highly conceptual, it was felt that formally encoding them with a process notation might be divisive, prematurely creating the impression of normative (best practice) definitions.

An enhanced (grouped) list of seed processes and an initial grouping/sequencing was attempted which is revised and presented in Ch9.

It should be noted below (Figure 5-7) that the processes/exchanges extracted for the seed model were de-duplicated and clustered by meaning e.g., the seed ideas of 'tariff', 'cost', 'subscription' were combined to the idea of 'Charging Model'.



Communication	Collection	Context
Connection	Classification	Collaboration
Certification	Conversion	Choreography
Confidence	Construction	Curation
Charging Model	Computation	Commentary
Canonical Sources	Correlation	

In contrast to a single request/response from a known search engine the process of observation may be characterised as one or more communication processes across several repositories starting with discovery of sources, the disclosure of metadata, the negotiating/establish of technical data exchange and the grant (either manual/technical) of licenses

Figure 5-7 Interim Structured seed process catalogue: Adapted from (Brown et al., 2014)

5.6.3 Actor/Participant View

Finally, a seed list for the group of Actors/Agents facets was considered. In terms of the individuals/groups building and using the system these were identified in terms of role (context), objectives (motivation) and function (focus):

- Functional Roles
  - Policy Maker
  - Operations (including sales)
  - Project Co-ordination
  - Researcher
  - Analyst
  - Data Owner
  - Developer
  - Security/Compliance

- Objectives
    - Profit
    - Transparency
    - Insight
    - Research

A review of the agent "straw man" presented here concluded that this model was heavily focussed on internal users/roles at the WO itself, and this was potentially missing important external roles and exchanges implied by the process models and also non-human roles/agents.

This was also the most sparsely populated part of the facet list with roles/motivations poorly represented in the corpus. Other roles such as those underpinning operational requirements were partly inferred from standard IT and Ops practice. In particular, there seemed little explicit evidence in the literature on *why* data is required/used and what motivated users to engage and so an additional scaffold/model was sought to underpin an understanding of why users engage with WO.

5.6.4 16 Motivations: after Reiss

Whilst there are numerous models of personality and motivation these are sometimes difficult to apply and may be largely substantive. Some of the psychological/psychoanalytical models are based on subjective theories rather than grounded in wider empirical research (e.g. Jungian, Adlerian and Freudian interpretations are partly grounded in sources such as ancient literature and personal childhood experiences). Other more behavioural models (such as Maslow's Hierarchy of needs) are simpler to understand yet few appeared differentiated enough for the purposes of this research.

Following a review of existing research, a characterisation of motivations was adapted from the work of (Reiss (2000) 2004) in an attempt to build a vocabulary of reasons-to-engage-with-WO. Reiss' work is grounded in a survey of more than 5'000 participants and renders a manageable (though sometimes exotically described) model of 16 factors that describe human motivation. It is less well-known but more nuanced than models such as Maslow's hierarchy and is chosen as *a suitable approach* rather than the *only/best approach* for the purposes of the project. The objective is increased discrimination/classification between motivations for WO without pre-qualifying/pre-selecting more substantial psychological theories of mind and social theories of Agency/Structure at an early stage. The objective then is not to (dis)prove Reiss but rather to build on his earlier work and adapt/employ his measurement tool.

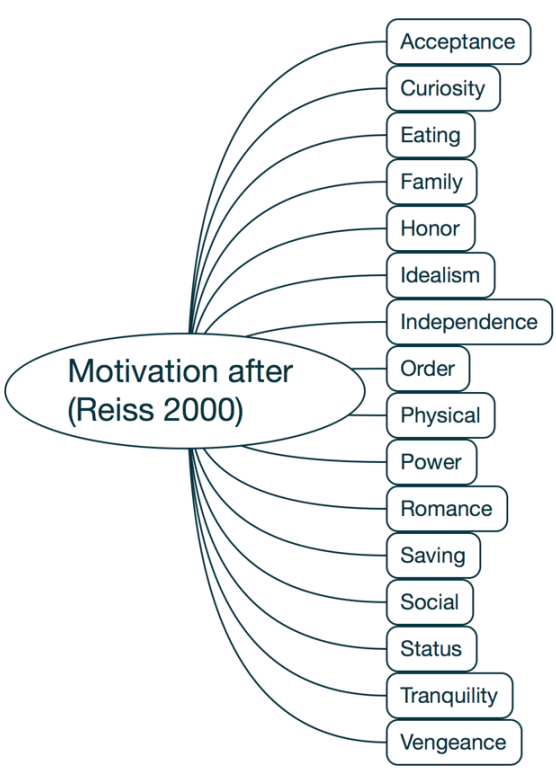


Figure 5-8 Source model from a Reiss model of motivations.

In Ch9 a revised model of motivations will be presented based on analysis of interviews and open source datasets and at the initial stage the Reiss motivations R1-16 were ordered in a sequence according to cognitive principles which first take stimulus then perceptions then meanings and then behaviours i.e., from artefact → cognition/meaning→reaction/emotion→behaviour as shown below.

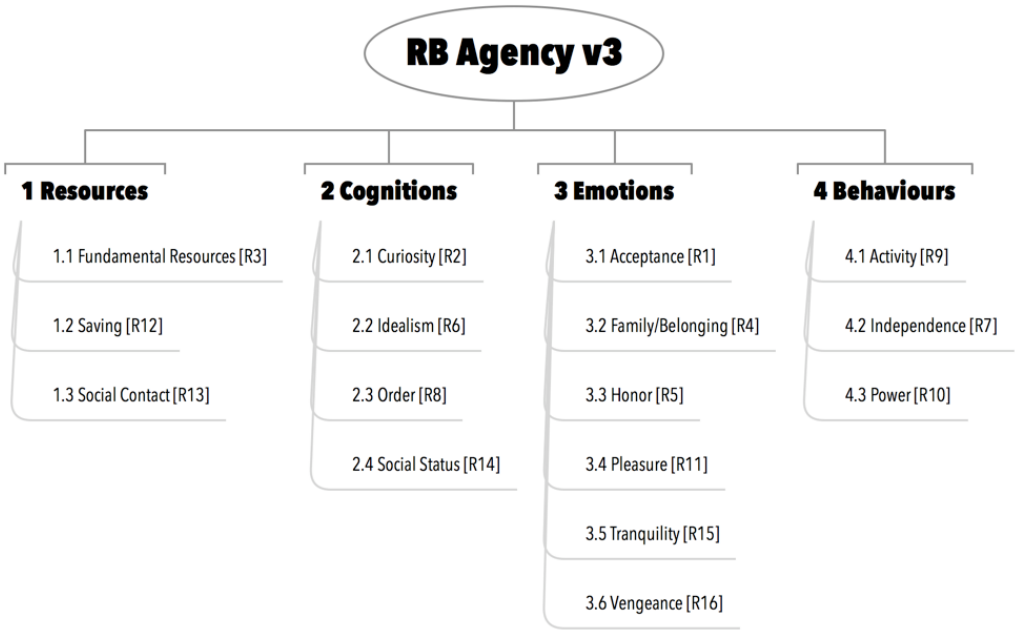


Figure 5-9 Revised model from Reiss to reflect the flow of cognition/response

There are some notable issues with the Reiss model for which adaptations have been made. I propose, for example, that Reiss' concept of "eating" seems underdeveloped and inconsistent with the other factors here in an unsuccessful attempt to introduce scale into the notion of eating by mixing the need for food (absolute/universal) with the need for variety (scalar/individual). This idea also fails to consider other basics such as shelter and has been recast here as a more general attitude to fundamental resources (food, water, shelter for individuals) and parallel concepts for business (such as capital and customers). "Vengeance" also seems oddly named when the wider Reiss narrative indicates this is more akin to "a response/reaction" - this again has been adapted.

Based on Reiss' model we are able to code/compare motivations as Reiss [R1-R16] adding new elements Brown [Bn-Bnn] if not covered by the original model.

In this section, we have created straw models for various perspectives/groupings of WO facets based on the taxonomic/discourse analysis. These models are iterated and refined based on participant interviews and experiments to deliver a final DNA model presented in Ch9.

5.7 Reflections on the experience and results

5.7.1 The Taxonomy

1. A faceted taxonomy offers a straightforward method to categorise concepts and create sub-structure. The hierarchy, however, does not allow for cross-structure or grouping between sub-structures, notional sequencing or non-hierarchical arrangement, thus driving the need to develop visualisations such as the concept map and process maps for enhanced visibility and explanatory power (see below)
2. The initial taxonomy (adapted from Whitworth) lacked a transparent narrative structure - mixing physical, operational and social elements under arbitrary clusters without any way to visually distinguish or compare these between WO systems. This led to the development of the three key perspectives of Design, Narrative and Agency under which the facets were later reorganised
3. Perspectives (personal, business, academic, charity and government), whilst part of the Taxonomy, cannot easily be reflected in the concept map of a single WO since this requires a wider "Ecology" of WOs. This led to the development of an ecosystem process phase, the W^3O concept and the Agency perspective
4. The initial process maps translate well from the extracted concepts to narratives heard from users. The three-phase model of discovery, processing and publishing, however, fails to account for ecosystem factors going INTO discovery and nor does it consider changes in

the system state resulting (EXITING) from the WOs actions. This led to an extended five-phase process model (e^5) shown in Ch9.

5.7.2 Modelling the notion of data/content

While sitting within the taxonomy, the attempt to model WO Data (content) as part of the concept map was initially unsuccessful. In essence, a WO could store or refer to any type of content (from *A*eronautics to *Z*oology), and this could change from day-to-day. Even reproducing a classification within the Taxonomy is arguably redundant given the existence of the Library-of-Congress or Dewey systems of classification.

Dis-entangling data from services is also problematic. While it clear that "data" may not always imply a service", services often *do* imply some underlying data which they deliver or from which they are driven: examples such as Provenance and Analytical services are cases in point.

Data is not ostensibly required in-of-itself but rather for a purpose (see Whitworth's definition of STS and the de Roure and Berners-Lee definitions of Social Machines) and hence this may be better reflected in a treatment of Agents and Agency. It may be possible to consider that data may itself be considered a (non-human) actor in the overall WO eco-system particularly where there may be distinctions between the existence of information as data and information as human knowledge.

I have (temporarily) introduced a topic-based taxonomy, e.g., Data>Data Topics>Society>OSN> which may co-exist with established classification schemes (such as Dewey Decimal and Library of Congress) or be replaced by them.

e.g.

006.75 Specific types of multimedia systems

006.754 Online social networks

There seems little evidence that a new mapping offers any benefits over the prior art. Searching for data (anywhere in W^3O) vs. for data on a known WO remains highly relevant and while Schema.org is currently used to identify and locate datasets within W^3O via search, the question of further refinement is left to future research.

5.8 Conclusion

I have demonstrated a flexible faceted approach for classifying Observatories which allows for extensibility, not only within the definition of what Observatories are structurally but also in terms of the social perspective of why they are used. The taxonomic breakdown has been further enhanced through the use of various conceptual visualisations that can account for the representational weakness of the faceted format by showing linkages, arrangements and sequences of concepts that are otherwise not possible within the taxonomy.

The seeding process and a comparison with facets of related systems have suggested that an innovation perspective (addressing what people are trying to achieve with specific data/services rather than simply the features/resources themselves) is a key perspective to understanding the operation and eventually the interoperation of WOs.

This does *not* imply that technology and architecture are *less* relevant than motivations - particularly given that the efficiency, scalability and reliability of Observatories rely heavily on technical/architectural choices for storage, querying and analytics. The wide availability, however, of cheap, powerful industry standard technical components and architectural approaches may tend to promote shared/common solutions over non-standard technical innovations due to the pressure for interoperability, accessibility and standardisation. By contrast, the broader application innovations, particularly at an aggregate W^3O level, may show far greater variations and specialisations.

Ultimately a differentiated model of the WO (the technical), the motivations/actions of users (the social) and the resulting exchanges (the socio-technical) are required to fully describe the nature of WO and W^3O in particular.

In the next Chapter, the "straw man" models are validated/tested in three experiments

1. The WO facets are compared with WO “cousins” (functionally similar information systems) using an established taxonomy to consider the theory that WO is simply an instance of some other existing class of system (addressing RQ1).
2. A source of data demand (requests from a government open data system) is considered to profile the reasons/motivations for which users ask for data to be shared. (RQ2/RQ3)
3. Three pilot groups are observed and interviewed to compare/validate the theoretical models obtained via content analysis with in vivo experience of WO users through group/individual interviews and participant observation.

Chapter 6: Testing/Refining the WO Model

In Short ..

In this chapter, the theoretical (*in vitro*) seed models from Ch4/5 are tested (*in vivo*) against products/projects.

- The WO functional model is compared to other IT platforms to determine if WO is functionally subsumed by them. A conceptualisation of WO vs W^3O is developed
- Seed model motivations are tested against stated motivations for downloading open government datasets

6.1 Conceptualising & Disambiguating WO vs. W^3O

The aim of this project is to understand WO as it is conceived of and used by practitioners rather than to reach an abstract philosophical definition of "observation" versus other types of information gathering. It is also relevant in an emerging field such as Web Science to be able to define and disambiguate concepts such as WO from other similar systems/entities (RQ1) and to consider what similarities/differences might contribute to a community of WOs (RQ2).

A meaningful conceptualisation of a system as it is designed, implemented and operated is more than simply a collection of features and so an analysis of WO cf. W^3O as abstract information systems and also other systems using the context of an overarching taxonomy was performed. This offers a lens through which to compare the WO features obtained in Ch4 with other systems and evaluate the challenge that WO may be 'unremarkable', i.e., that WO is identical to (or only trivially different from) existing solutions and hence not worthy of separate study.

This might be the case if, for example, WO were completely subsumed by a superclass of some system (S) that matched or exceeded all WO features. Thus, for WO to be a *type* of S, S should (at least) have all the capabilities that WO has. This does, however, ignore novel applications and novel ecosystems for similar systems versus functional/structural novelty alone.

Thus we may have to recognise the application context and ecosystem context of WO to evaluate distinctiveness at a suitable level since, at a fundamental level, one must recognise that information systems (particularly those in the same problem space) will naturally share technologies and features.

Since:

> "Put succinctly, all information technology does or can do (..) is: capture and store data, distribute data for consumption and analysis to produce information which connects people together into collaborative working environments where information is shared to produce knowledge."
>
> **(Demarest 2007)**

Demarest's definition explicitly separates interactions (Figure 6-1):

- Technical interactions between machines:
  - Machine (Purple) to/from Machine (Green)
  - Socio-technical interactions between users/machines:
  - Machine (Purple) to/from Person1 (Purple)
  - Machine (Purple) to/from Person2 (Purple)
- Social interactions between users:
  - Person1 (Purple) to/from Person2 (Purple)

It is unclear from Demarest what the potential is for knowledge transfer between distributed systems Person1 (Green) to/from Person1 (White). This further highlights that Web Observatory has two main conceptual modes (standalone and distributed) in the guises of WO and WO→W^{3}O which Demarest's model of Information interaction would depict as distinct structures:

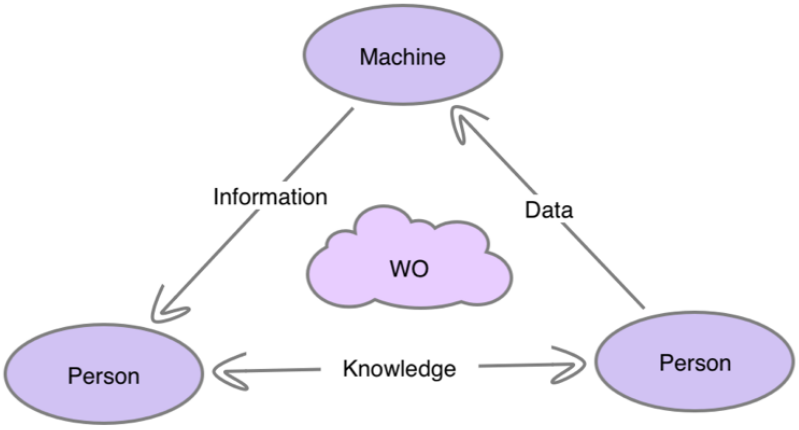


Figure 6-1 WO information interactions adapted from Demarest[31] (2007)

[31] www.DSSresources.com: accessed Aug 2014

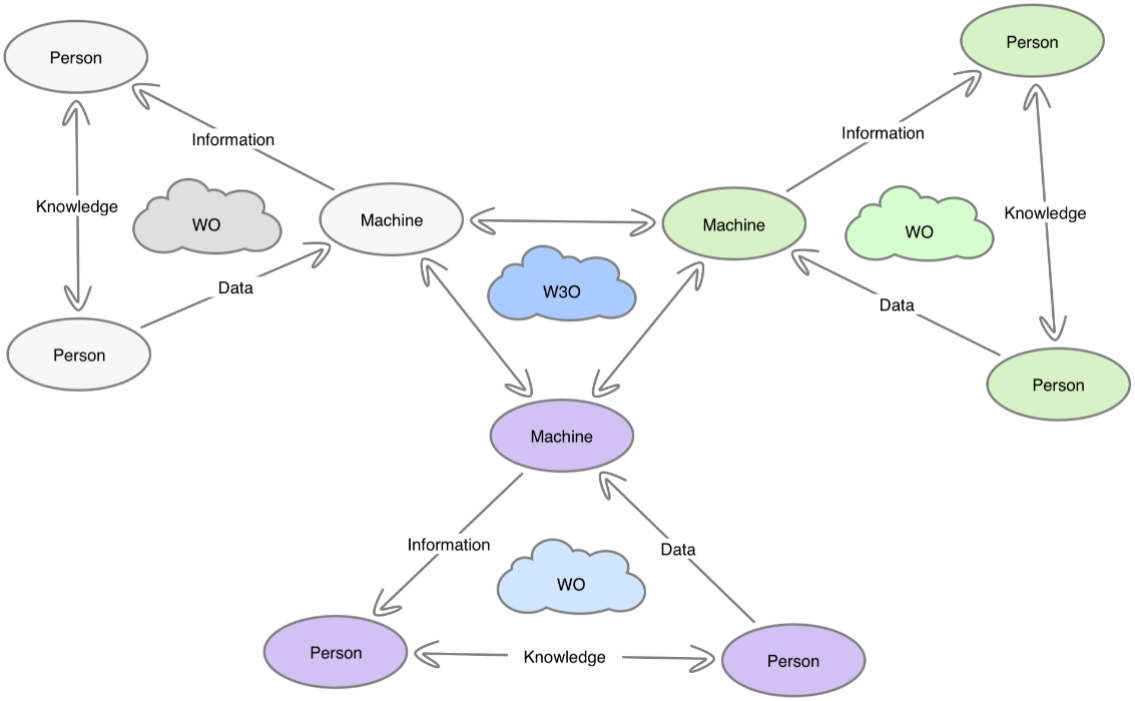


Figure 6-2 W^{3}O information interactions

(adapted from (Demarest 2007) on DSSresources.com accessed: Aug 2014)

This suggests that the different "modes" of operation as WO or W^{3}O may imply different (emergent) features and indeed a different definition. What is also noteworthy in Demarest's definition is the assumption that knowledge is gained, not for its own sake, but rather in support of information systems which can support the *objectives* and key *decisions* of users (workers) in given contexts. He also talks about the "politics of Data Warehouses[32]" through which projects often experience problems where:

1. They cross **organisational** treaty lines
2. They change both the terms of data **ownership** and data access, and expose the often-**checkered history** of data management in the IT organization
3. They affect the **work practices** of highly autonomous and powerful user communities in the firm.

We may assume that exposing WO data to W^{3}O may at least partially be affected by these (*social*) issues and so competition/political trade-off may be inherent in the process. Looking for a model/perspective from which to compare/contrast WO with other types of IT system the notion

[32] http://www.noumenal.com/marc/dwpoly.html

of supporting research decisions was considered to evaluate if WO might broadly be considered to be a class of decision-support system (DSS). Informal definitions such as Wikipedia's:

> ".. a computer-based information system that supports business or organisational decision-making activities"

appear insufficiently detailed requiring a more formal model of DSS to support a nuanced comparison. If we consider (Sprague 1980), (Power & Sharda 2007) and (Alter 1978) for a Taxonomy of DSS we may determine to what extent WO→W^3O fits another established model. I note, in citing Sprague to examine the distinctiveness of WO, that Sprague's own purpose (of establishing the novel existence of decision support systems 35 years ago) was quite similar to my own. The need to disambiguate new systems from old is itself not a novel endeavour.

In 1975, Alter studied 56 decision support systems and categorised them into seven groups:

1. **File drawer** systems: (e.g. Database) that provide access to data items.
2. **Data analysis** systems: that support the manipulation of data by computerised tools (e.g. Data Warehouse)
3. **Information** systems: (e.g. BI systems) that provide access to a series of decision-oriented databases and small models.
4. (**Numerical**) Financial Model-based: (e.g. Option pricing or predictive models) that calculate the consequences of possible actions. For goal-seeking, "What if?" or sensitivity analysis.
5. **Representational** model-based DSS: that estimate the consequences of actions on the basis of simulation models that include relationships that are causal as well as accounting definitions.
6. **Optimisation** model-based DSS (e.g. Deep Learning) that provide an optimal solution consistent with a series of constraints that can guide decision making. Examples include scheduling systems, resource allocation, and material usage optimisation.
7. **Suggestion** (Profiling) DSS based on logic models that perform the logical processing leading to a suggested decision for a fairly structured or well-understood task. e.g. insurance renewal rate and credit scoring.

(Power 2001) refines Alter's taxonomy and clusters the seven categories into two groups: Data-driven and Model-driven and adds Communication-driven and Document-driven categories of system. With this extension, we start to see something that the WO might correspond to in terms of a compound or hybrid support system. Generic vs. Specific and Web-based were added to the perspectives.

The extended taxa above were mapped to a range of information systems that might be considered to share features with WO. i.e., search, data storage, analytics, social networking interfaces

Table 6-1 below compares WO/ W^3O against other systems to see whether there is support for the claim that WO is simply a subset of another type of information system according to an established Taxonomy. It should be noted that:

1. The analysis assumes that the comparative examples chosen are reasonable/useful in that they might also be considered to support decisions for the purpose of this comparison e.g., It would not be reasonable to offer PowerPoint as a DSS
2. The evaluation of systems considers only the apparent features of the systems e.g., Google Search provides search results it does not reflect what else may be captured/produced by Google internally.

The analysis is based on publicly available material and an evaluation of the features/architecture of the selected system types. It uses a broad indication of *intended* features rather than technically feasible adaptations. e.g., one might physically type Java code into a PowerPoint slide but not reasonably claim that PowerPoint is a Java development environment.

Our candidate systems are defined as follows:

- **W^3O**: the idealised emergent system (and morphological space) described by the content analysis and facets extracted from the earlier analysis describing openly discoverable data and apps
- **WO**: a standalone WO offering access to private/public data and apps to a given audience
- **Search**: a repository of URI's and their structure with a profiling mechanism to assist in collapsing the search space and determine which URI's are most valuable/relevant for a personalised search
- **Web Analytics**: a model of actual vs. expected events and behaviours used to model structure, trends and anomalous system events and user behaviours
- **CKAN**: a repository of open datasets and documents published by an agency to allow stakeholder transparency and the transformation/re-use of data assets to create new value
- **Big Data** (Hadoop): Allows users to view temporal/value pattern correlations to support decisions around causation models and resulting responses/action in big (streaming) data

- (Social Media) **Aggregators**: Collection of OSN messages/feeds allowing users to determine aggregate sentiment around a location, person, product or meme across multiple social networks and news sources
- **Sandboxes**: A restricted shared environment allowing users (from commercial and non-profit groups) to look for shared combinatorial value and opportunities between (non) open datasets and resources without the need to make the sources or results public/open.
- **Co-Laboratories**: An environment in which specific issues, datasets or grand challenges are presented for solution by the crowd either as best-solution-takes-the-prize competitions (e.g., Kaggle or Innocentive) or as socially focussed citizen + researcher projects (e.g., MyExperiment or Zooniverse).

6.1.1 Findings

	W3O	WO	Search	Analytics	CKAN	Hadoop	Aggregator	Sandbox	Co-Lab
	e.g. Facets	e.g. NeXT	e.g. Google	e.g. Splunk	e.g.data.gov	e.g. Datameer	e.g. GNIP	e.g. TDA	e.g. Kaggle
Data Driven									
File Draw	YES	YES	NO	NO	YES	NO	NO	YES	YES
Data Analysis	YES	YES	NO	YES	NO	YES	NO	YES	NO
Analysis Info System	YES	YES	YES	YES	YES	YES	YES	NO	NO
Model Driven									
Numerical Model-based	YES	YES	NO	YES	NO	YES	NO	NO	NO
Rep Model-based	YES	YES	YES	YES	NO	NO	YES	NO	NO
Opt Model -based	NO	NO	YES	YES	NO	YES	NO	NO	NO
Profile Model-based	YES	YES	YES	YES	NO	YES	NO	NO	NO
Knowledge Driven									
	YES	NO	NO	NO	YES	NO	NO	YES	YES
Document Driven									
	YES	YES	YES	NO	YES	NO	NO	YES	NO
Communication Driven									
	YES	NO	NO	NO	YES	NO	YES	NO	YES
Intergroup (Collab) Driven									
	YES	NO	NO	NO	NO	NO	NO	YES	YES
Function Specific									
	NO	YES	YES	YES	NO	NO	YES	NO	YES
Generic									
	YES	NO	NO	YES	YES	YES	NO	YES	NO
Web-based									
	YES	YES	YES	NO	YES	NO	YES	NO	YES
	W3O	WO	Search	Analytics	CKAN	Hadoop	Aggregator	Sandbox	Co-Lab
Score	12	9	7	8	7	6	5	6	6

Table 6-1 Analysis of WO vs. related technologies using Alter's Taxonomy

No claim is intended here that the WO/W^{3}O *should* be defined as a DSS - merely that DSS is a convenient comparative framework. However, if these raw counts are reasonable (in the absence of inter-rater reliability figures), then the proposition that either WO or W^{3}O are a simple subset of the other types of system would fail - given both WO and W^{3}O claim more features from the extended DSS taxonomy than other classes of DSS.

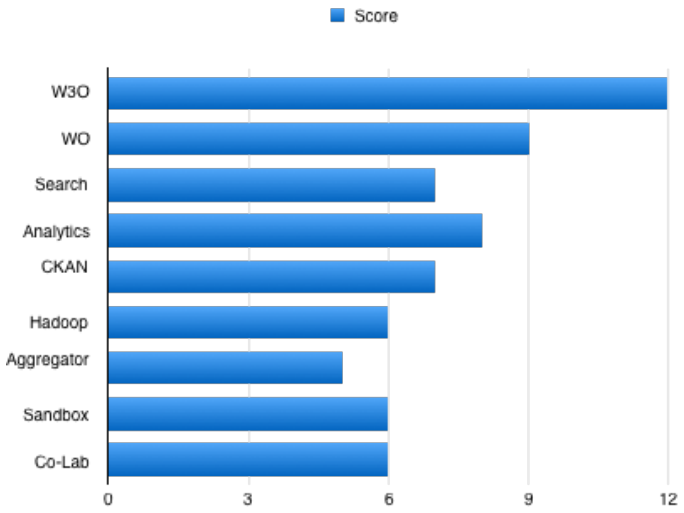


Figure 6-3 Number of DSS features by DSS type

6.1.2 Discussion

The inclusive nature of the WO/ W^{3}O spans several of the taxa defined by Alter and Power thereby qualifying this approach as a *composite* or Hybrid DSS. While no single comparative system appears to subsume all features of WO→ W^{3}O, two assumptions are made here:

- WO/W^{3}O systems are considered as notional (*idealised)* features of the WO node or emergent ecosystem and not using a particular extant Observatory
- No extension/enhancement of the other systems was considered though the idea of extending the other platforms is itself intriguing.

It is interesting to note that, when taken as individual WOs vs W^{3}O, the difference from other systems is minor. There may indeed be an argument to classify individual WO instances as a "type of" Web Analytics package or as a "type of" open data repository - which are precisely the challenges offered by some members of the community.

In comparison to the WO technical (design) features pre-supposed by each of these systems the novelty of WO, like the VAO before it, appears to be the *application* of the underlying system design. The technical novelty with respect to other systems may be quite low compared to the innovation or application novelty.

(Djorgovski & Williams 2005) note of astronomical VOs (which inspired WOs) that:

> "..any of the individual functions envisioned for the VO can be accomplished using existing tools (e.g., Federating massive datasets, exploring them in a search for particular objects, outliers or correlations but in most cases such studies would be too time-consuming and impractical and many scientists would have to solve the same issues repeatedly…VO serves as an enabler of science with massive/complex datasets and as an efficiency *amplifier* [33]"
>
> **(Djorgovski & Williams 2005)**

The key insights here are:

- The marked difference in the evaluation of WO and W^{3}O underscores the importance of maintaining the distinction between the WO and W^{3}O concepts
- Given the closeness of functionality between WOs and other systems not originally intended as WOs, there emerges a broad selection of systems that might *participate* in the W^{3}O with relatively minor adaptations compared to architecting/building a new Observatory.

6.1.3 Summary

In this section the WO/W^{3}O were compared to notionally similar systems to check for functional "novelty". The standalone WO appears functionally similar to other types of Web Analytical tools (though it remains novel in terms of its varied application and ecosystem). W^{3}O appears (perhaps unsurprisingly) more functionally novel as it represents a totality or superset beyond the capability of any individual WO. Similarities suggest a broad range of existing system might easily join the W^{3}O ecosystem if reasons/benefits could be found for them to do so. Given that even within the WSTNet the level of WO adoption/interoperation has been relatively low, the additional importance of non-technical factors seems likely.

[33] An usual term, also heard from a participant describing the WO

6.2 Data Demand vs. Data Supply

It became apparent during this project that it has been far easier to determine which datasets and resources are available to use (supply-side) on WOs/VO's than to determine which resources are actually used or even requested (demand-side) and even less data on why they may have done so. A notable comment from Joy Bonaguro (San Francisco's Chief Data Officer) was that in their open data journey the one thing that they had learned was that the measure of success in open data was "not simply releasing more of it" and so in this section I present a demand-focussed data set in the context of UK open government data.

6.2.1 Open Government Data Demand

I started by looking for proxies to understand which groups are asking for data and why (and hence why they might use a WO) to compare our motivational model. I reviewed the data.gov.uk site for an indication of themes and actual usage and also the ODUG open data request app/lists (Tableau app) covering an additional 789 requests for information release (ironically) no longer open/available from http://odug.org.uk/open-data-request-roadmap/. Though not a rigorous division, one might characterise the data.gov.uk site as providing information on both supply and fulfilled demand whilst the ODUG data represents the unfulfilled demand and includes reasons/justifications for the request. The request mechanism is open to individuals, groups, business and government itself and reflects the potential types of communities we might see for WO though we must consider the caveat that public stated reasons for wanting data (a Goffman frontstage reason) may not always match the real reason (backstage reason).

At the time of writing the data.co.uk site lists 32'677 datasets (usage data for only 29'367 is available) which the UK Gov currently publishes. This site runs on the Open Knowledge Foundation CKAN platform. In its default setting, CKAN could be characterised as a simple 'file repository' rather than an Observatory, yet with the addition of a harvesting capability to automate the discovery and upload of datasets plus 389 visualisation/analytical apps on the data.gov.uk site, this profile more closely matches the conceptualisation of WOs as a means of discovering, hosting and visualising/analysing data. CKAN (at least as it is implemented at data.gov.uk) could be classed as a VO for open data, and could perhaps be adapted to work with the WO network as a source and/or a participant.

Figure 6-4 shows the unusual typology adopted by data.gov.uk to characterise datasets that are downloadable (e.g., it is unclear which spending is not part of finance or why linked data is held separately)

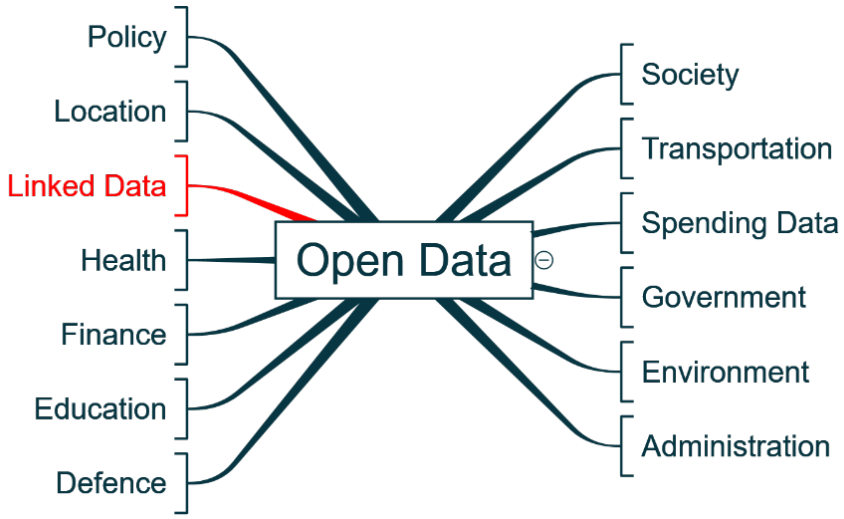


Figure 6-4 Open Data request topics

Looking at the corresponding ODUG dataset (Figure 6-5) reflecting new data requests for different types of data by different user groups (ranging from personal to government).

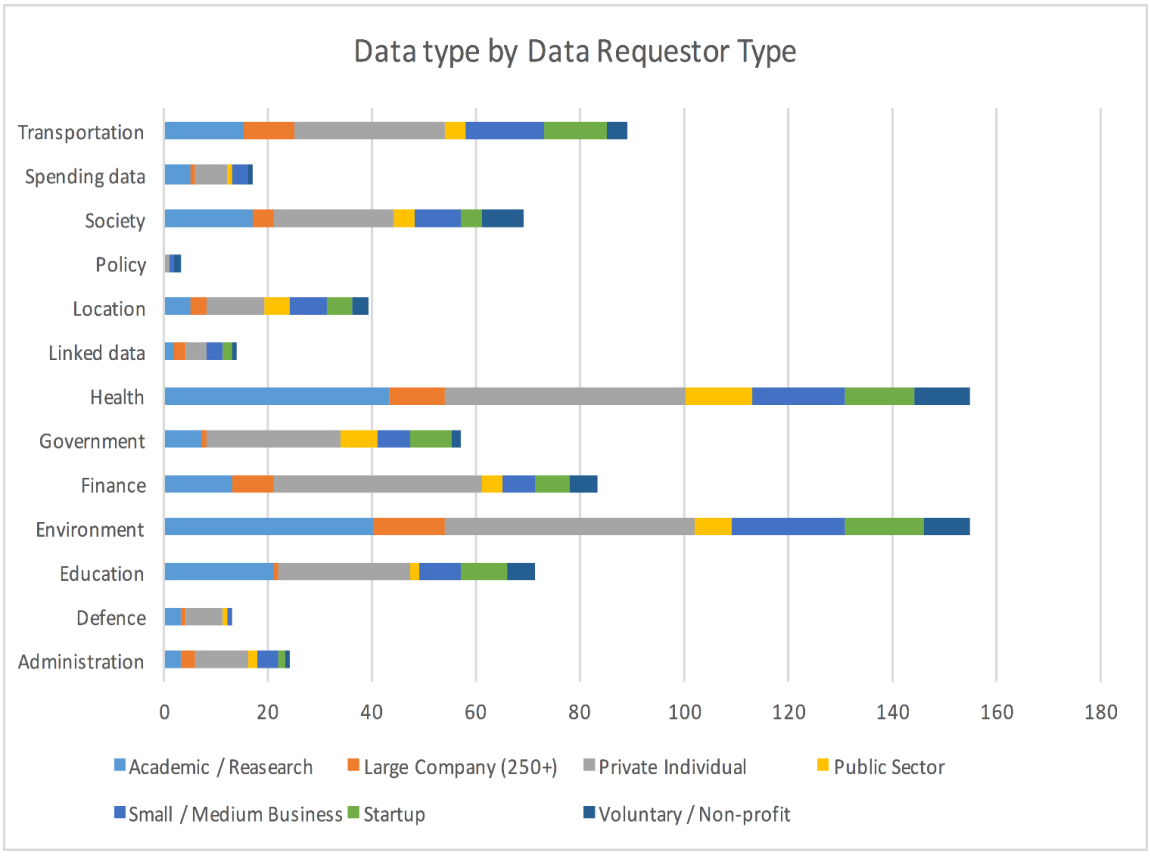


Figure 6-5 Data Type by Requestor Type

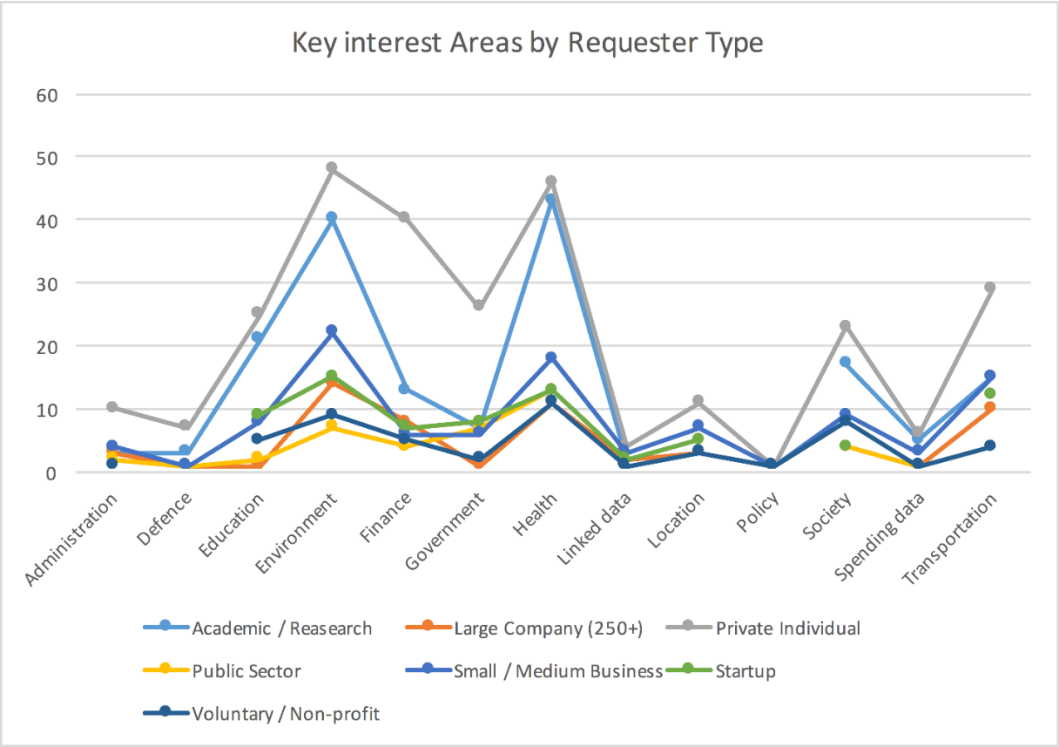


Figure 6-6 New data requests by organisation type. Source ODUG.

We can see distinct spikes of interest across all requestor types for datasets relating to environment and health and, of these, the majority are personal or academic requests. One might be cautious before generalising from this result to WOs more broadly since data.gov.uk offers specific data only and is not a generic site containing data about all topics. Notably:

- Ironically relatively few requests are made to understand government spending (one of the key drivers for the service is to be transparent around spending).
- Despite the assertion heard in the participant interviews that UK Gov is one of the biggest consumers of its own data this is not reflected by the figures[34].

While CKAN offers data on WHAT or WHO has been involved in downloading data, it is notable that no insight is offered (or requested) as to WHY the dataset is valuable. Such data might allow the provider to value the cost of provision vs. the derived benefit. Hence the ODUG dataset (while smaller) offers an unusual perspective and an analysis of reasons/motivations was performed as a potential proxy for the reasons users might engage with open data on WOs.

As mentioned in Ch4 the adapted (Reiss 2004) model offers a multifaceted model of why individuals are motivated and these have been extended/mapped from purely personal framings + motivations to social group framings in Table 6-2 Original/Extended Reiss Motivations.

[34] Alternate interfaces (not shown here) may apply here.

No	*Original Meaning*	*Adapted Meaning*
R1	Acceptance - the need to be appreciated	Promoting, highlighting a cause
R2	Curiosity, the need to gain knowledge	Research, to know a thing, transparency
R3	Eating, the need for food	Resources generally
R4	Family, the need to take care of one's offspring	Responsibility for constituents / members
R5	Honour, the need to be faithful to the customary values of an individual's ethnic group, family or clan	De Facto standards, common practice, i.e., “how we do it."
R6	Idealism, the need for social justice	Making this right, appropriate, easy, complete, saving time/cost
R7	Independence, the need to be distinct and self-reliant	The process of acting without control/influence
R8	Order, the need for prepared, established, and conventional environments	ΔSafety/∇Risk, better decisions
R9	Physical activity, the need for work out of the body	Exercise, Movement, Exploration, Travel
R10	Power, the need for control of will	Take action, exploit (Business) opportunity, offer/improve a service, app
R11	Romance, the need for mating or sex	
R12	Saving, the need to accumulate something	Curation, having for the sake of having
R13	Social contact, the need for relationship with others	
R14	Social status, the need for social significance	Acting for, on behalf of a community

R15	Tranquillity, the need to be secure and protected	
R16	Vengeance, the need to strike back against another person	Accountability, Consequence
B17	Liberty	the belief in permission to act (vs. the desire or the action itself)
B18	To Know the difference	To know more about the structure, Δresolution and differentiators/limits between things
B19	To Know the extent	To know a complete set of facts - everything about a set of things

Table 6-2 Original/Extended Reiss Motivations

These are used to code [R1-B19] the data requests though it should be noted that some latitude has been applied to Reiss' sometimes strange nomenclature.

e.g. Under the slightly poetic term "Vengeance" we include the more prosaic concepts of accountability and consequences which in particular are applicable to the owners of budgets, government actions and the perception of value-for-money - see also "justice" as fairness/equity rather than purely as the product of the court system. Three additional codes Brown (B17-B19] were added to reflect the data found earlier in this study:

B17. Liberty - freedom to engage in 1-16

B18. Territory - boundaries between concepts in 1-16

B19. Knowledge (as a resource) vs. the desire to obtain knowledge.

As part of adapting the Reiss model to reflect a cognitive process (from perception to behaviour) motivations were grouped into four perspectives following a cognitive path from:

- Artefact/experience (what a thing is)
- Cognition/meaning (what a thing means)
- Reaction/emotion (how thing makes you feel)
- Behaviour (how the feeling makes you act).

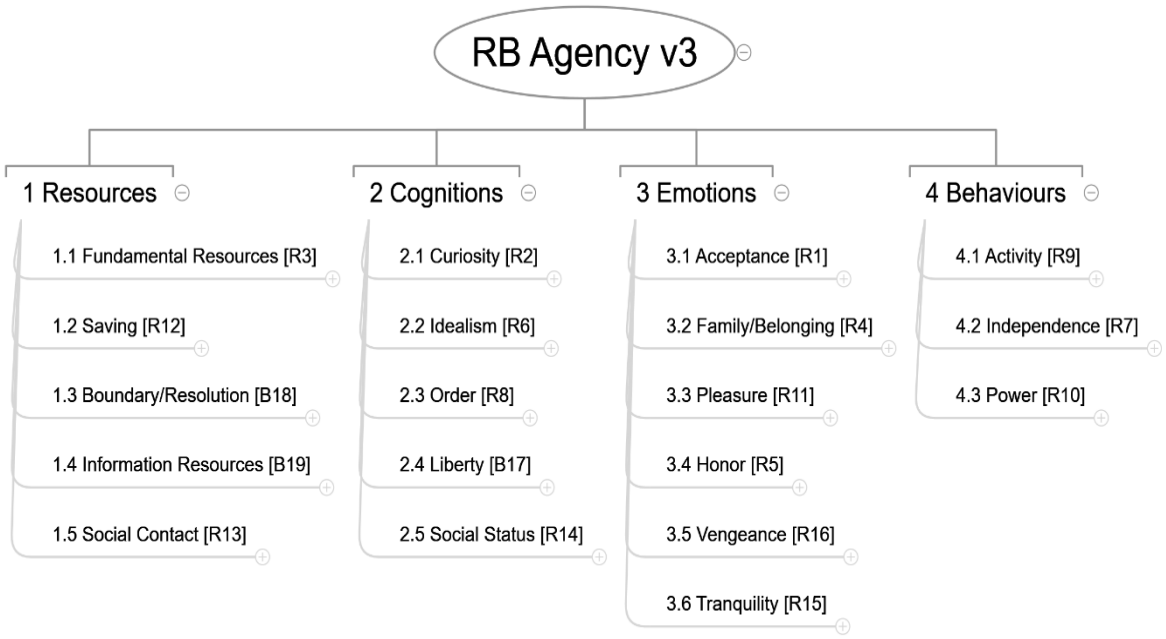


Figure 6-7 Extended Reiss model of motivation/agency

In terms of data robustness the request dataset from ODUG is incomplete with 549 of the 1318 datasets marked as classified. Only 393 of the remaining 789 data requests contained non-blank reasons for the data request, and I cannot determine the accuracy of the responses (such as businesses claiming the need for data as personal/social) The key thematic groups were as follows:

- ∇ Risk
- Δ Capability/Service
- Δ Accountability/Transparency
- ∇ Time/Effort/Cost
- Δ Accuracy/resolution
- Δ Intelligence about the market.

These were mapped to the Reiss Structure with *two* reason (R) codes (primary/secondary) for each request to avoid an overly simplistic single coding. The total counts across both the primary (what appeared to be the most important factor) and the secondary code (other info) is combined below to a raw score below:

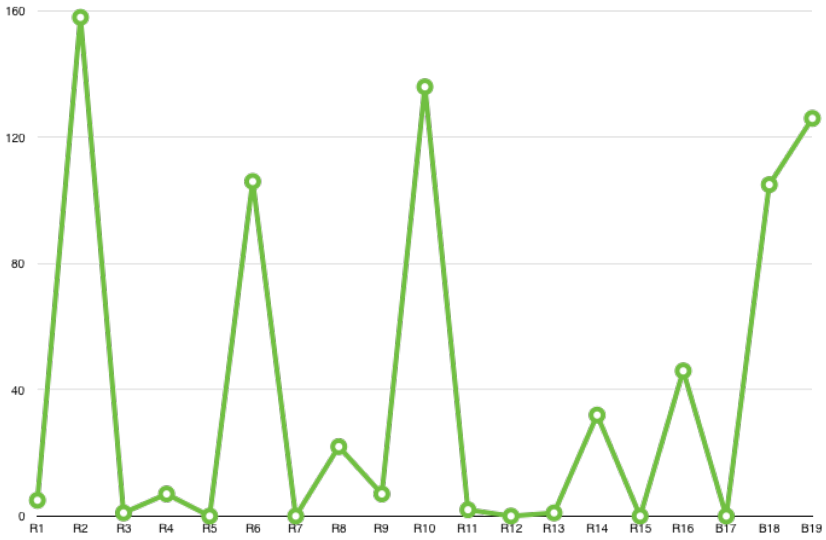


Figure 6-8 Breakdown of Primary/Secondary motivations after modified Reiss classification

Code	Count	Description
R1	5	A few organisations wanted data about their achievements to be released as a method for effecting recognition (possibly further funding)
R2	158	A large proportion of the requests are from academic institutions and labelled as queries for "research" purposes. Where these are "personal research", a co-coding for justice/fairness is often present intended to correct mistakes /challenge perceived bad judgements/decisions This code is to "know a thing."
R3	1	Very few people signalled their request as directly leading to "resources" though note the much more common R10 which is empowerment to action
R4	7	The R4 "family" code was used to signify the provision of services to promote the welfare of constituents or members of a key demographic or group
R5	0	No-one explicitly requested data out of form, habit or to respect convention
R6	106	A large group of requestors flagged improvements in speed, cost, ease or to achieve "the right result" (Justice)

R7	0	No-one flagged that data made them independent of control (though several cited independence from other sources or additional work which was tagged as R6)
R8	22	A number flagged avoidance of risk as being a driver for the data they requested
R9	7	Several requestors stated that data would enable them to exercise (more easily)
R10	136	A large number of requestors were looking to create/improve services and apps to generate revenue and/or improve the scope/capability of their service.
R11	0	There were no data requests for reported reasons
R12	0	Notably, all requests had a reason for applying the data beyond the act of curating it
R13	1	Making contact with others was given in one request
R14	34	Social benefits, performing civic duty, responsibility to a group were cited here
R15	0	No-one offered tranquillity (beyond safety) as a reason
R16	46	Accountability (for actions, spending) were cited here
B17	0	The notion of freedom per se (vs. free to do a specific thing) was not offered - though this is an often cited theme in interviews and thus may not be offered front-stage
B18	105	The desire to know the DIFFERENCE or BOUNDARY (structure) of/between things was cited here
B19	126	The desire to know ALL INSTANCES of a group of things was requested here .

Table 6-3 Raw coding counts for primary/secondary reasons combined

Overall we see three broader groups emerging from the peaks underpinning:

- A solution/service (an outcome focus)
- Data feeding a more accurate/predictable model (a design focus) and
- Data as the search to know a thing or set of things (the knowledge focus).

The profile of tribes: academics, business (small/large) and communities (small/large) reflects our original user group profile but while an element of:

- Academics → research
- Business → apps/services/profit
- Government → policy

is both expected and present here, there is further diversity in the reasons reflecting additional framing within the occupational/tribal frame. This will be considered in Ch8.

This subset of requests is valuable in that it represents the **demand** element for open data rather than the more easily observable **supply** element. It must be seen in the context of a much larger set of resources that are already available on data.co.uk (i.e., which do not need to be requested as new data releases but therefore unfortunately do not tell us WHY the dataset is needed (if at all). Thus we are restricted to proxies such as actual usage being approximated by download figures.

Usage information on data.gov.uk is poorly differentiated: Figure 6-9 below is based on the site usage statistics gathered by CKAN and shows the total (page) accesses for the top 20 datasets (7/6/16) on the site which forms part of a reported lifetime total of 23'876'526 page views over 7'581'208 visits . This tells us little about the *usage* of these datasets (total downloads are not reported) and the aggregate demand across the full spectrum of datasets (vs. the top 20 shown here).

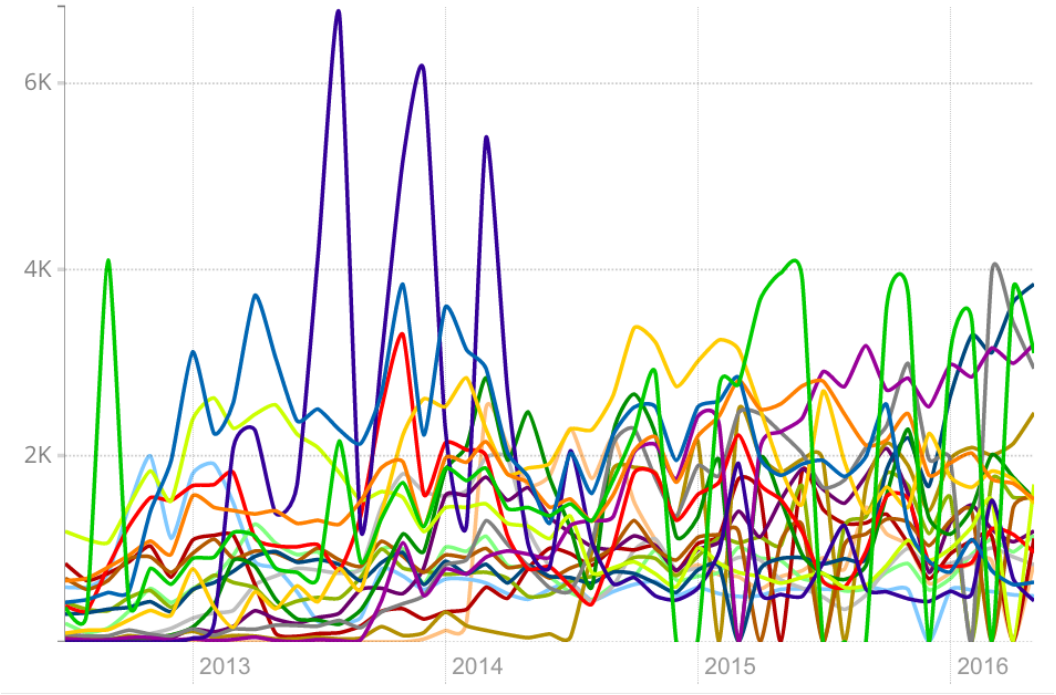


Figure 6-9 Top 20 dataset (page) accesses (2012-16). Source data.gov.uk accessed 7/6/16)

In Figure 6-10, Figure 6-11 and Figure 6-12 we see the mean dataset usage (downloads not views) for consecutive tranches of datasets to map the long tail of dataset usage:

- Top 10, 20, 30 .. 100
- Top 100, 200, 300 .. 1000
- Top 1000, 2000, 3000 .. 10 000
- Top 10 000, 20 000

Specifically, what is the mean usage per dataset in the top 10, 20, etc.?

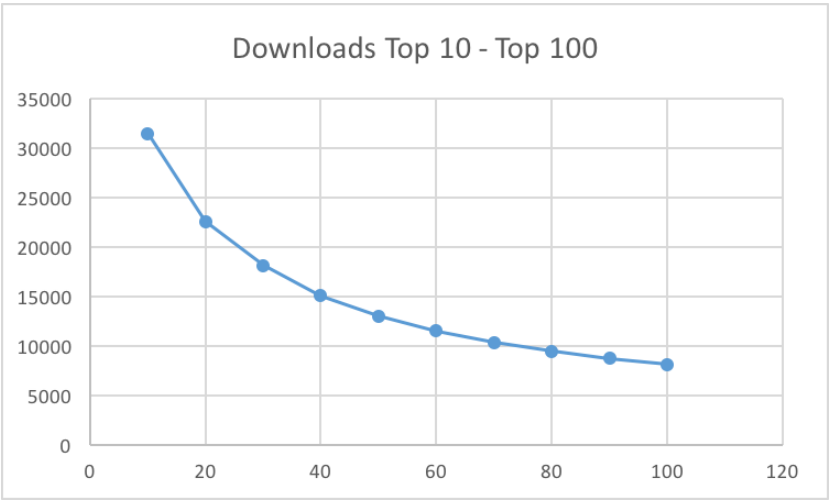


Figure 6-10 Top 10 - Top 100 dataset mean usage (2012-2016)

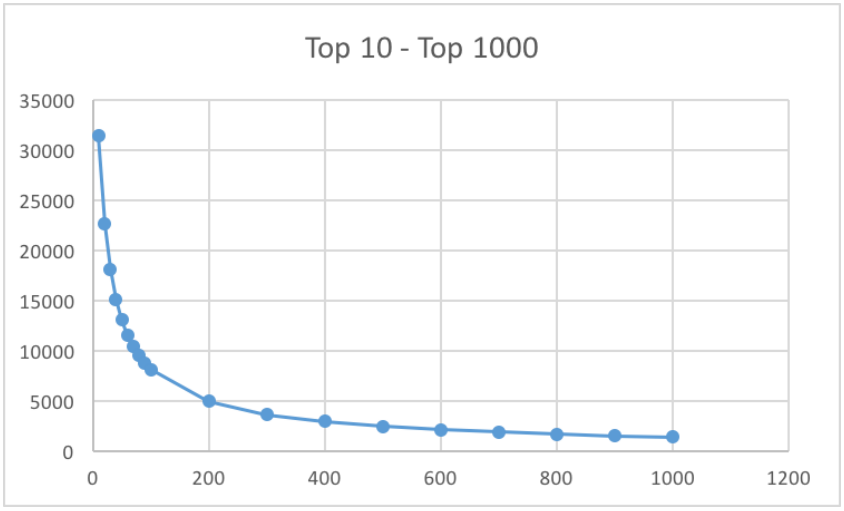


Figure 6-11 Top 10 - Top 1000 dataset mean usage (2012-2016)

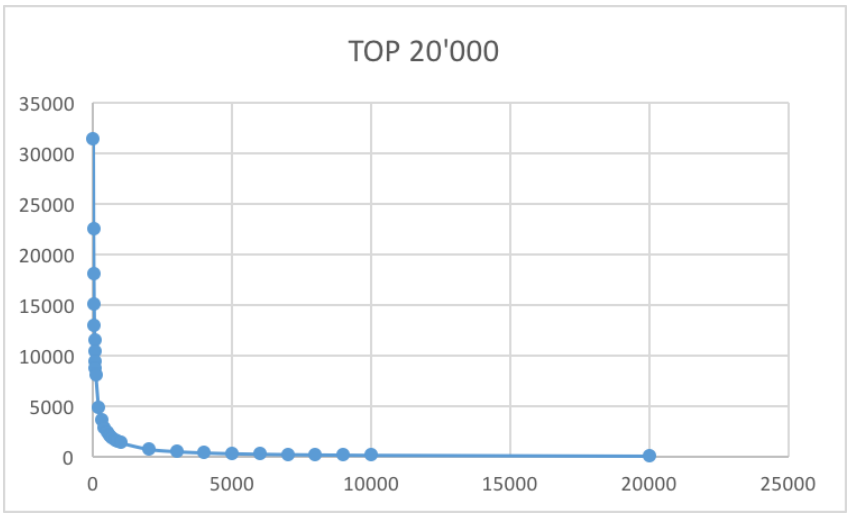


Figure 6-12 Top 10 - Top 10'000 (<20'000) dataset mean usage (2012-2016)

In summary terms we see a pronounced 'long tail' for the 32K reported data.gov.uk datasets

- 34.1% of the datasets show 0 downloads over the lifetime of the system - i.e. no usage beyond 19'334th ranked item
- 57.2% of the datasets have been downloaded 1 per year or less on average during the period 2012-2016
- 42.43% of all downloads (1.78Million) are accounted for by the top 100 datasets.

Whilst it is challenging to value open data in general and specifically the value of, say, the 101st most popular item on this list, it is notable that in Ch8 we will hear [Ivan], an academic participant, talk about funding for data services predicting:

> "These things die when the money goes away."
>
> **[Ivan]**

implying that the list of items which will ultimately be funded for hosting (and particularly for curation) will be limited by budget. This perception is limited by perceived value which in turn is often associated with the usage proxy - (it is indeed hard to argue the value of something which is hardly ever (never) used / or searched for). Thus in Figure 6-13, I note the corollary of the long tail in terms of where the focus on downloads and usage does occur.

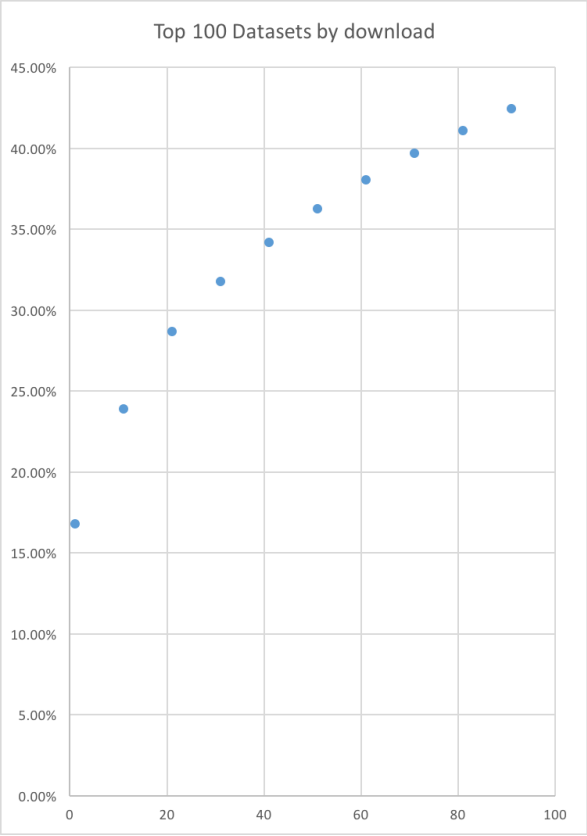


Figure 6-13 %Total downloads accounted for by top-ranked datasets

Thus 753'867 downloads (42.3%) are accounted for by approx. 0.3% of the datasets.

6.2.2 Data Quality caveats

- Of 32'677 datasets, declared usage data for only 29'367 is given
- When examining the data for views/page hits these, we note the data.gov.uk definitions
    - "Views" is the number of times a page was loaded in users' browsers.
    - "Downloads" is the number of times a user has clicked "Download" .

and when controlling for data quality/plausibility a number of anomalies of the ratio of page views to downloads. Assuming a dataset must be viewed before being downloaded (unless the user accesses the data from the URL directly but that it still gets logged as a download via the page), the ratio must logically remain <1.

Figure 6-14indicates a number of anomalies in the top 500 datasets where 20 show download/view ratios >1.

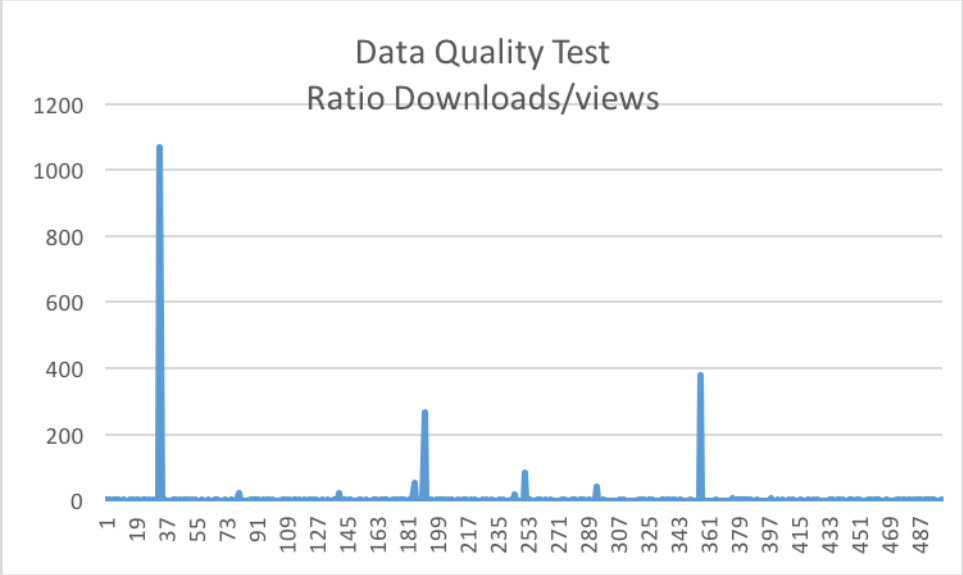


Figure 6-14 Data quality test for download metrics

With 180 datasets of r<1, 17 where r10, three where r100 and one where r1000 this indicates potential issues with data acquisition or with the use of other data access methods mixed in with the page statistics.

6.2.3 Discussion

Reference architectures such CKAN and SOCRATA offer the functionality to detect and harvest datasets and, through integration libraries, a method for hosting/linking to apps that build on this data, thus bringing them beyond the functionality of a simple repository and towards the notion of an observatory. While CKAN/SOCRATA are not focussed on Web Science or Web datasets and much of the data hosted on data.co.uk has little relevance per se to Web Science there is a path along which some organisations may choose to host both their Web and non-Web data in a single CKAN/SOCRATA instance. It is interesting to note that while individual WOs may restrict datasets to particular groups, the W^3O is, by nature a completely open endeavour since *only* datasets that are shared would be seen on W^3O and so in some respects, the W^3O is an open data concept.

The construction of this space (confirmed by interviews) has been one in which the role of the specific technology has been made transparent and subsumed by a focus on the format of the data and the desired attitudes to value and sharing. Thus the VO becomes transparent insofar as it works - claiming centre stage only when, (as noted by Star) ”in moments of breakdown”, it *fails* to offer required functionality, connectivity or analytical tools.

The systems reviewed here focus on the supply challenge and do little to reflect or support the demand challenge. This renders such systems 'context-free' to some extent as evidenced by the typology which reflects government departments rather than what people might want to do with the data. Indeed data requests may be invalidated if the requestor does not know which (named) dataset they need or whether it even exists. Social thematic elements may be missing here:

- A key challenge here has been around re-framing the concept of value and a sustainable model for curating and publishing data over time other than as a "point solution" - this may require changing the focus from the quantity of datasets to quality and impact measures
- There may be an active (if unconscious) avoidance around the issues of implementation for concern around excluding/marginalising potential participants, granting influence to one or other technical camp and the resulting loss of neutrality.
- A number of these interviews were quite challenging as the ability for systems to come together to create orchestrated cross-sector flows (the D and partially the N in DNA) has been assumed (sic), and little focus is placed here beyond making datasets "available".

The initial results showed that the ODUG set is probably a subset of the data.co.uk list though this is not initially self-evident. It is notable for an endeavour that is attempting to share/re-use data that:

- Disjoint reference number schemes are used between government and ODUG, and these are only partially mapped
- No overlap is found between the submission dates between the datasets
- Usernames (some potentially usable for re-identification) are used in one but not in the other dataset
- Despite citing the desire to understand more about the needs of data users (i.e., Why they are requesting data vs. What they are requesting) and specifically asking *why* each data is needed there is no analysis offered by the ODUG application nor via data.co.uk on the breakdown of reasons.

6.2.4 Summary

In this section, WO is compared to other conceptualisations to assess if the observatory is *subsumed* by existing technologies and approaches. The WO and W^3O platforms did not appear to be demonstrably sub-sets of other approaches implying that their application (if not their underlying technology) may be novel.

A source of data demand was identified to compare a 'straw man' motivational model adapted from (Reiss 2004) against >700 open government data requests. 13 of the 19 (approx. 70%) of the adapted model's elements were found at least once in the test data (which was limited to topics offered as open data) suggesting that this model may be useful as a nuanced measure of motivation and showing groupings suggesting three broad areas of focus which will be further investigated/validated though the interview process. An analysis of the data usage revealed a pronounced long tail:

- >30% of datasets had *never* been accessed
- Approx. 60% had only been accessed *once or less*

this raises political/economic questions of dataset value vs. the cost of collection and longer-term curation.

In the next section, a pilot project observing WO practitioners is described supporting the further validatation of candidate WO models.

Chapter 7: Pilot Project

In Short ..

In this chapter, the theoretical (*in vitro*) seed models are tested against (*in vivo*) products/projects and a number of candidate analytical/validation methods are attempted – some successfully and some unsuccessfully.

7.1 Introduction

Two Web Observatory workshops were observed (2013 China and 2014 Singapore) for the purpose of eliciting feedback on methods, suitability, perceived problems and opportunities around WO from research students in the field using WO services and data in a live (hackathon) environment. The second is documented in less detail than the first since it was used to gather longitudinal data to confirm/enhance the results from the first event and from individual participant interviews which are reflected elsewhere.

The Tsinghua WOW (Web Observatory Workshop) event ran for two one-week slots and took place across sites in the UK (Southampton) and China (Tsinghua University) in late November and early December 2013. It was aimed at fostering interdisciplinary and international cooperation within a Web Science framework and leveraging the newly available (prototype) Southampton Web Observatory (SUWO). The format involved two phases of travel/exchange in which 29 students/staff from S. Korea and China visited Southampton for a week with the intention of forming teams and completing technical training. They received a briefing on some core data science procedures and learned how to access SUWO. Thereafter, a week of hacking within cross-disciplinary teams in China was scheduled to develop and present the project.

With the support of local academics in Southampton as well as the international teaching staff, the 27 participants collaborated in teams to formulate research questions and to select from pre-existing (contributed) datasets in order to plan the development of an Observatory-based solution to be "hacked" and made live within the two-week period.

My goals were to observe researchers engaging with the theory vs. the practice of WO research, to experiment with different methods of gathering and analysing data and validate the nascent DNA models in order that early failures/successes would inform the research approach.

The Singapore WOW ran for one week at National University of Singapore (NUS) in December 2014 and brought together more than 50 students across 6 study groups to create applications/demonstrators based on WO and NUS social observatory data but was less focussed on the Observatories themselves and more on the available datasets.

In both cases I attended the event and observed the teams but did not join a team or work directly on the projects, selecting participant observation over action research.

7.2 Research Method

The methods employed were participant observation, focus group/individual interviews and narrative analysis from project outputs hence no attempt to 'improve the process' in the form of Action Research was made. Interviews were transcribed, coded and analysed using nVivo. After the event follow-up, additional interviews and iSurvey questionnaires were employed which revised the nVivo model.

7.2.1 Data Collection

Each group was observed during the hackathon period, and three focus group interviews were conducted. Following the interview, a SWOT analysis of the participant issues from these focus groups was prepared to generate further questions/themes for the later questionnaires and interviews. This was reviewed with the group for feedback/revisions.

A set of questionnaires based on the previous WO taxonomy and concept map (See Ch4) was prepared. These questionnaires sought to expose three aspects of WO:

- The function/structure of WO - what the WO was perceived to do
- The patterns of behaviours/processes - how the WO could be applied
- The priorities and motivations for usage - what problem was being addressed through use of the WO.

The questionnaires were previously rehearsed/reviewed with five academic researchers to elicit feedback on the design and layout, and these were then sent to representatives of each team for completion via iSurvey.

The project process consisted of:

- Concept modelling during five days of participant observation, focus groups plus 5 follow-up interviews and a review of academic papers subsequently written by the groups
- An automated lexical analysis used a pre-filter for typical "stop words" and considered the top 1000 concepts/terms by frequency from the transcripts. These were manually adjusted to exclude further stop-words and to include only those terms accounting for >= 0.1% of the total corpus.
- A manual analysis and coding of the transcripts was performed and did not consider the frequency of terms - only relevance.
- The Top 25 were considered in relation to the 25 issues raised by each group.
- Responses to questionnaires on the structure, processes and motivations for using WO. Three questionnaires (see Appendix) were sent to two participants per group.

7.3 Group Theme Summaries

Three cross-institution student teams of up to eight participants each (to allow for a mixture of technical and non-technical skills in each group) consulted with local academics to choose topics around data obtained (or obtainable) for the WO event:

- The use of humour in social media during the Salt Crisis in China (characterised as an Academic **(A)** project)
- A study of product features highlighted in social media during the iPhone 5 launch (characterised as a Business **(B)** project)
- A study of anti-corruption themes in Chinese social media (characterised as Community **(C)** project)

The groupings (**Tribes**) were thereafter summarised as **A,B,** and **C**.

Humour during the Salt Crisis (Academic Tribe)

In the aftermath of the Fukushima nuclear incident in 2011, there were substantial concerns, particularly in rural communities, concerning the possibility of contamination and associated health risks from nuclear material. A meme arose during this period concerning the alleged protective properties of common salt which led to bulk buying of unusually high amounts of salt which in turn led to salt shortages. This project looked at the dialogue between those supporting and those mocking this idea through the lens of social media messages from Chinese online social networks during the period. This project was looking to test proximal theory (both physical and temporal distance) against test data.

iPhone 5 Product Launch (Business Tribe)

With the release of the iPhone5 Apple added a number of new features to the new devices along with several incremental improvements. This team were seeking to model the "buzz" before, during and after the launch date of the iPhone 5 in order to gain insight into market reactions to the new phone and in particular the importance of different features. This group were looking to understand markets in order to assist in influencing market share and spending/resource patterns (not only with Apple vs. iPhone users but also with journalists focussing on writing about key features).

Anti-corruption Themes on Chinese Social Media (Community Tribe)

Chinese on-line social networks are used not only for typical mundane conversations but also for political debate extending to open critique of the Chinese government and even allegations of malfeasance and corruption. The Chinese government openly practices censorship of various media channels, and the team were interested to see how the discussion of corruption allegations might diffuse and be interrupted/taken down by government agencies. This group were looking at issues of transparency and calls to action.

7.4 Findings

In the following section concepts specified by the group as “important” are extracted manually from the focus group discussion including:

- Underlying data issues
- Issues around analysing/presenting the data as an "app"
- Non-technical (social) issues
- Researcher observations.

These are compared with an automated transcript analysis of important (sic) issues based on frequency. Automated/manual analysis is compared and an example of the groups visualisation/analytic output is provided with the groups SWOT model. The SWOT factors are expanded in each case into an nVivo model.

7.4.1 Humour in Crisis (A-Tribe)

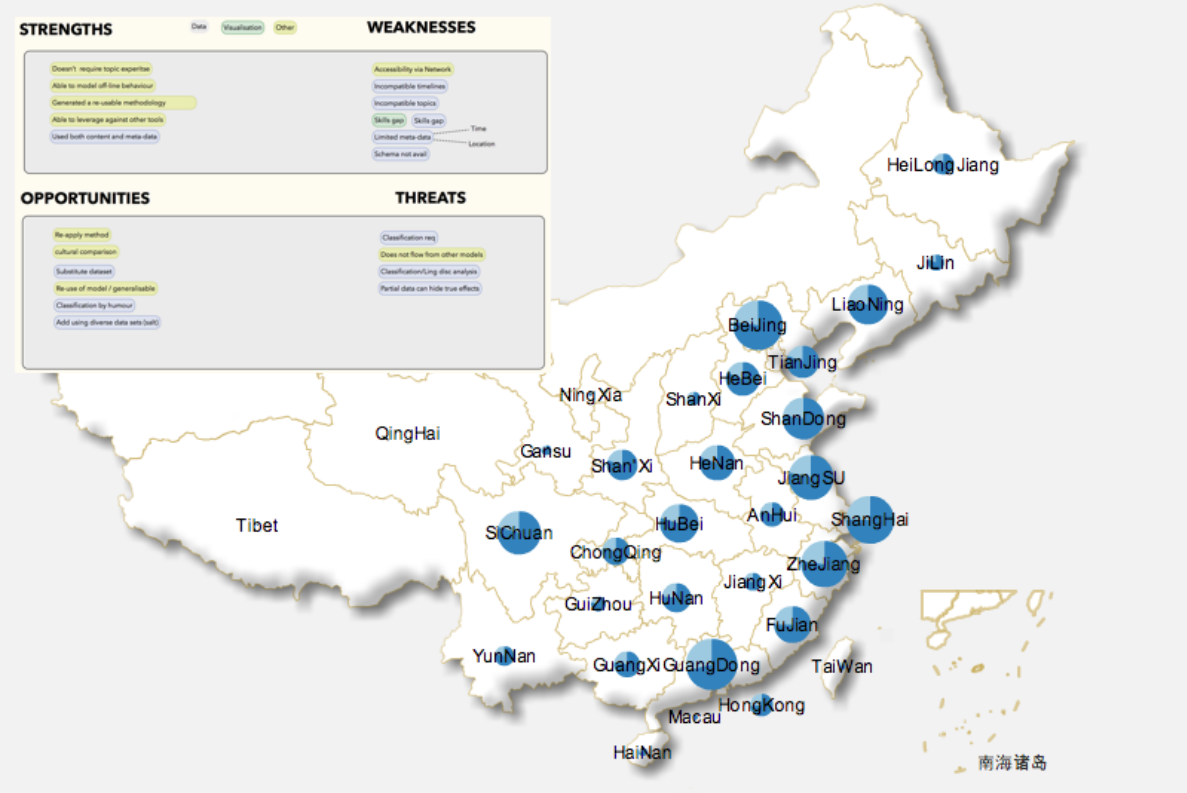


Figure 7-1 Humour Group output of concepts and visualisations

The group were able to use Chinese language social media sources to rank the level of humour associated with the meme of using salt vs. radiation threat over time and by location to test the theory of proximal and temporal distance.

The discussion points were coded to 24 SWOT terms that the group rated as "important" (Figure 7-2)

This group focussed on developing a research question/method to match the available data with some scope changes on-the-fly as additional team members (and datasets) were added later in the project. This effectively opened up opportunities to attempt cross-cultural analyses. Incompatible timelines and, more broadly, the problem of disjoint datasets made this difficult.

The group were tackling micro-blog data in Chinese and Korean and attempting to identify (and even characterise types of) humour leading to a focus on WO offering in-built lexical analysis and translation tools. Key issues around the completeness of metadata were raised and the reliability of certain metadata in describing live behaviour was questioned. Tweet "locations" were actually the location in which the user had *registered* rather than where the Tweet was *authored*. This was identified as a potential error factor and underscores the importance of accurate metadata.

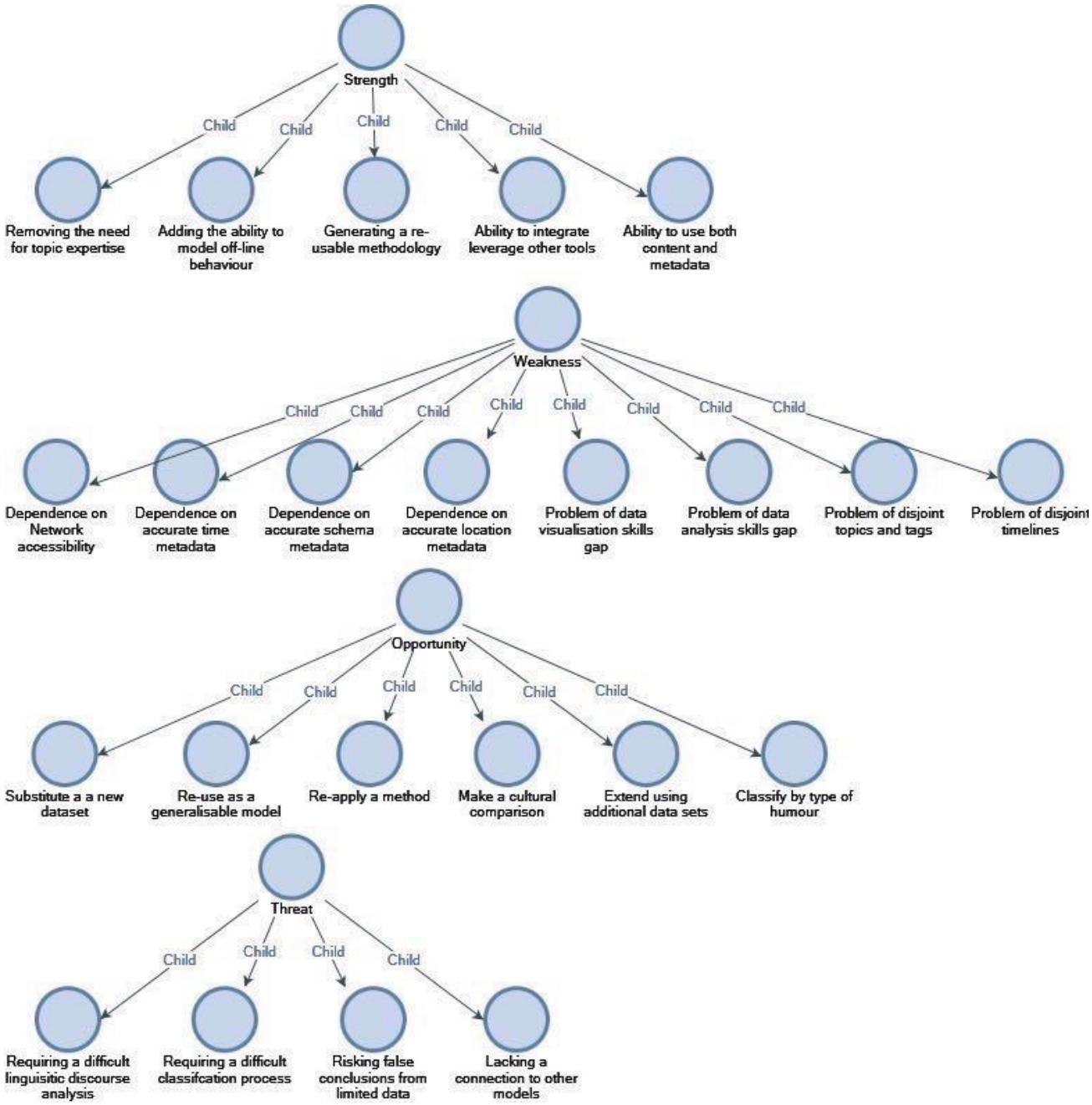
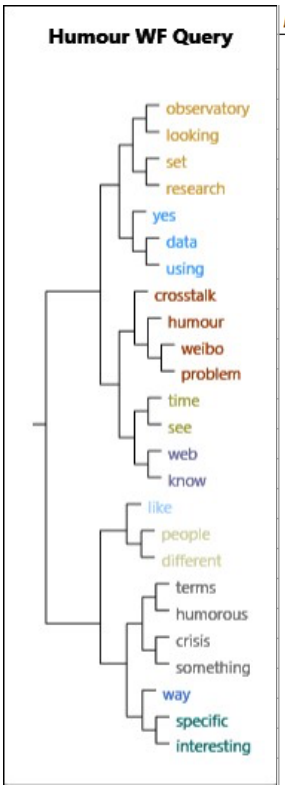


Figure 7-2 Humour Group SWOT Coding Scheme

During the follow-up interview, additional stress was placed on the need for additional built-in support/tools for data visualisation.

Notably, in two additional papers written after the event the use of WO as a technology is largely transparent though both offer broad support for the concept while requesting "more data" and "more tools" but give no other details or further perspective on the WO itself.



Focus Group Terms	Payload	Extracted terms
Ability to integrate leverage other tools	tools	crisis
Ability to use both content and metadata	content and metadata	crosstalk
Adding the ability to model off-line behaviour	off-line behaviour	data
Classify by type of humour	humour	different
Dependence on accurate location metadata	location metadata	humorous (dup)
Dependence on accurate schema metadata	schema metadata	humour
Dependence on accurate time metadata	time metadata	interesting
Dependence on Network accessibility	Network accessibility	know
Extend using additional data sets	Extend data sets	like
Generating a re-usable methodology	Re-usable methodology	looking
Lacking a connection to other models	connection to other models	observatory
Make a cultural comparison	cultural comparison	people
Problem of data analysis skills gap	data analysis skills gap	problem
Problem of data visualisation skills gap	data visualisation skills gap	research
Problem of disjoint timelines	disjoint timelines	see
Problem of disjoint topics and tags	disjoint topics and tags	set
Re-apply a method	Re-apply a method (duplicate	something
Removing the need for topic expertise	Removing topic expertise	specific
Requiring a difficult classifcation process	difficult classifcation process	terms
Requiring a difficult linguisitic discourse analysis	difficult linguisitic	time
Re-use as a generalisable model	Re-use model (duplicate)	using
Risking false conclusions from limited data	limited data	way
Substitute a new dataset	Substitute new dataset	web
		weibo
		words

Figure 7-3 Humour Group transcript vs. auto-coded themes

In Figure 7-3 we see the themes as highlighted by the groups and the key term or “payload” of the theme compared to the keywords extracted automatically. Matching/related concepts are highlighted.

7.4.2 iPhone (B-Tribe)

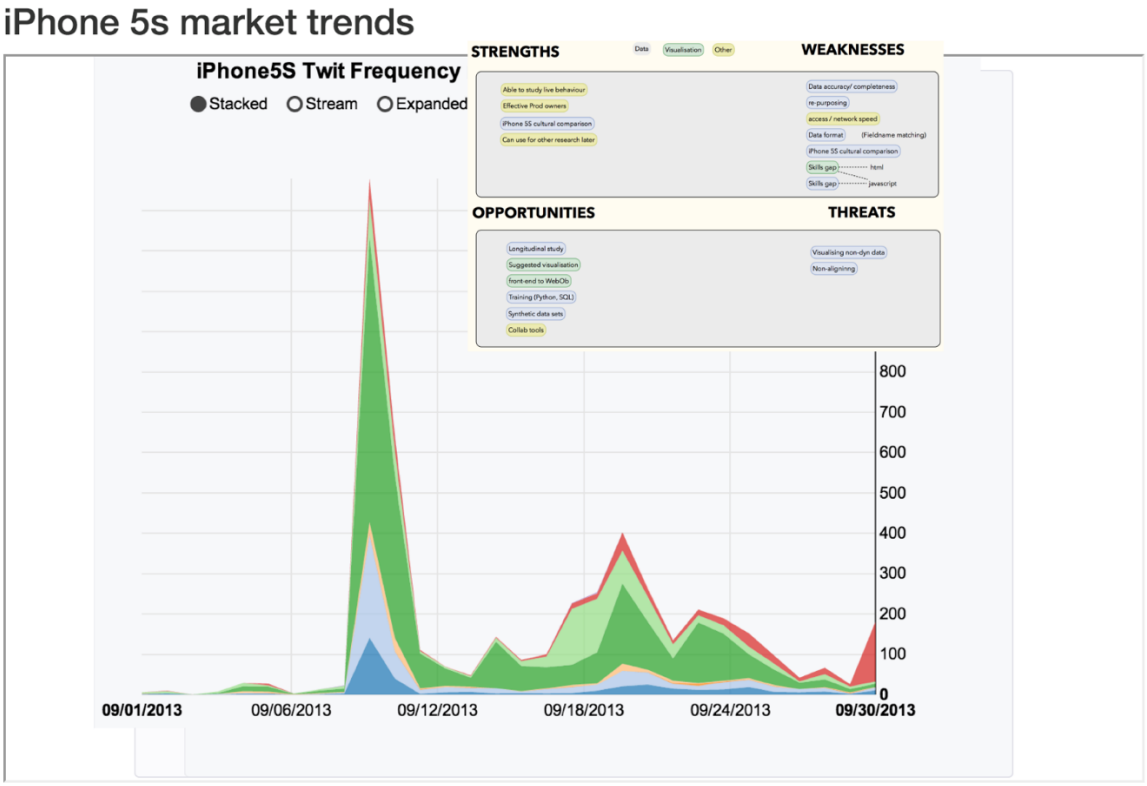


Figure 7-4 iPhone (Business) group output of concepts and visualisations

The group were able to use English language social media sources to rank the level of interest in iPhone5 features over time (pre- and post-launch) to inform the activities and focus of technical journalists and marketing groups.

This group had intended to trace micro-blog posts relating to product features leading up to (and subsequent to) a launch across different markets but were unable to source matching data and instead focussed on the Chinese market only. Basing their work on Chinese material, the inclusion of language translation tools was rated highly. The inclusion of "drag-and-drop" graphics support was also noted as this would have removed the need to hand-code the visualisations using the D3.js graphics library (not thought to be a key deliverable of the research) and would have saved time for more core research. During the group interview, the issue of Signal-vs-Noise in data was raised and the need to balance the volume of data against the need to clean/filter. Overall the team were optimistic about longitudinal studies around long-term ownership and reviews and comparative country studies highlighting which features are of most value in different markets. When asked about also adding methodological and data-cleaning standards, the follow-up interview indicated that this would be too restrictive and that “raw" streams with no enforced processing or method would offer the most flexibility.

Notably, WO curated (pre-processed) datasets were thought to be of limited novelty/usability vs. raw streams. No additional papers were written.

24 SWOT terms were rated by the group as "important".

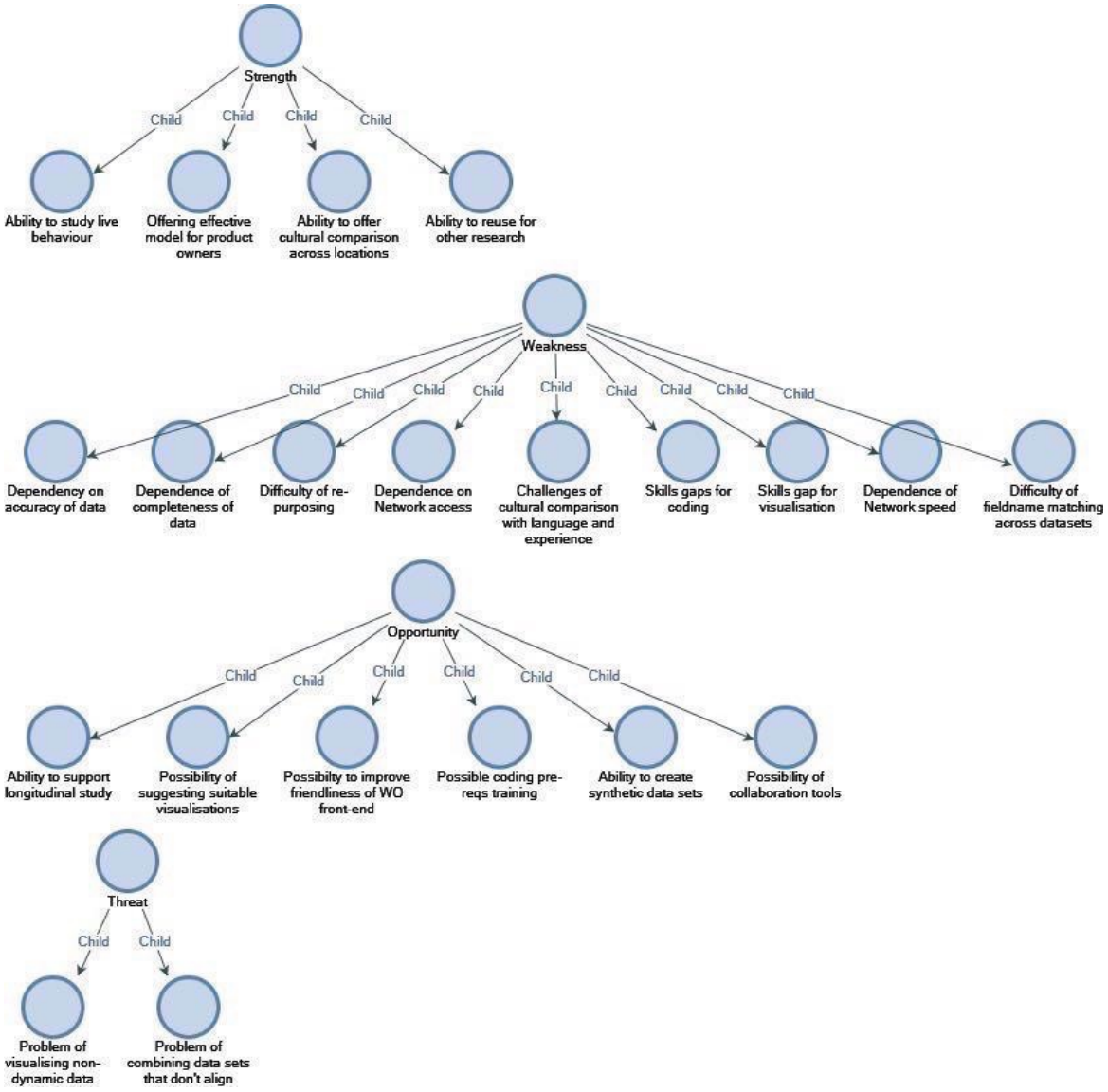
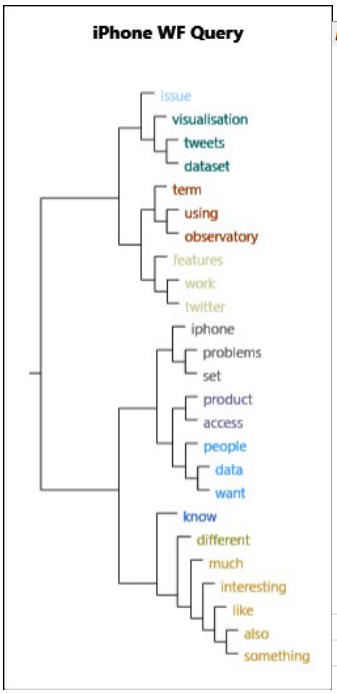


Figure 7-5 iPhone group SWOT coding scheme



Focus Group Term	Payload	Extracted Term
Ability to create synthetic data sets	Synthetic Data	access
Ability to offer cultural comparison across locations	Compare culture	also
Ability to reuse for other research	Reuse data	data
Ability to study live behaviour	Live behaviour	dataset
Ability to support longitudinal study	Longitudinal data	different
Challenges of cultural comparison with language and experience	Language/Culture Challeng	features
Dependence of completeness of data	Complete data	interesting
Dependence of Network speed	Network speed	iphone
Dependence on Network access	Network access	issue
Dependency on accuracy of data	Accurate data	know
Difficulty of fieldname matching across datasets	Matching data	like
Difficulty of re-purposing	Re-use challege	much
Offering effective model for product owners	Business application	observatory
Possibility of collaboration tools	Collaboration	people
Possibility of suggesting suitable visualisations	Analytics help	problems
Possibilty to improve friendliness of WO front-end	GUI help	product
Possible coding pre-reqs training	Training	set
Problem of combining data sets that don't align	Matching data	something
Problem of visualising non-dynamic data	Visualisation challenge	term
Skills gap for visualisation	Visualisation skill	tweets
Skills gaps for coding	Coding Skill	twitter
		using
		visualisation
		want
		work

Figure 7-6 iPhone Group transcript vs. auto-coded themes

7.4.3 Corruption (C Tribe)

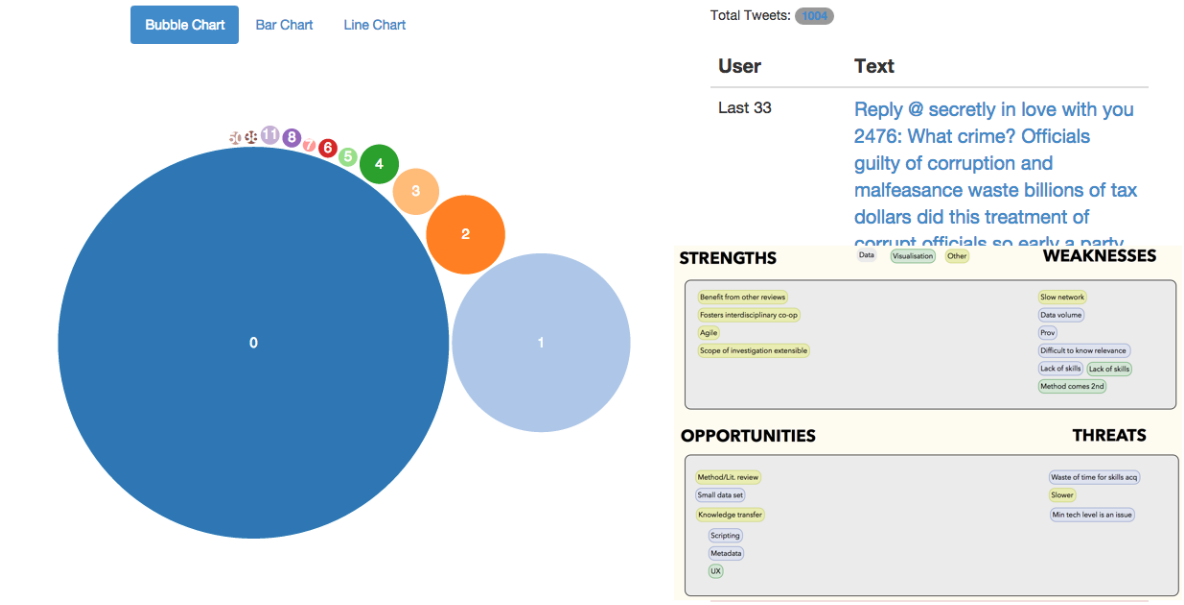


Figure 7-7 Corruption group output of concepts and visualisation

The group were able to use Chinese language social media sources to rank the level of interest in corruption themes gauged by the number of re-posting over time.

This group used the number of retweets as a measure of "virality" and importance but were hampered through not knowing what part of the total dataset or conversation had been included in the micro-blog sample (was this the 1st or 100th re-tweet?). Considerable time was spent

interpreting and profiling the available datasets due to lack of documentation/metadata, and ultimately two of three datasets were dropped due to mapping issues. Notably, the group wanted a way to make contact with the original research group to ask questions implying the need for a community around datasets. An important point was raised around metadata/provenance for APIs in addition to datasets since APIs also change over time, and this leads to dealing with deprecated APIs that may be associated with longitudinal data. During the follow-up interviews, the lack of documentation and insight into the datasets was re-emphasised, and challenges around using data across different political/legal domains were flagged. The group thought raw data should always be available as a basis for analysis even if pre-processed/filtered versions are also available. A vision of semantically-linked WO data was offered in which researchers searching by topic would be offered relevant materials and resources.

In an additional paper after the event, the authors offer a WO-centric view of their research process flagging the need for technical and organisational standards and raising the issue of copyright in databases. Re-use, they point out, may not only be limited by the technical and legal challenges but by additional ethical/privacy issues requiring anonymisation of datasets.

These equated to 20 terms that the group rated as "important".

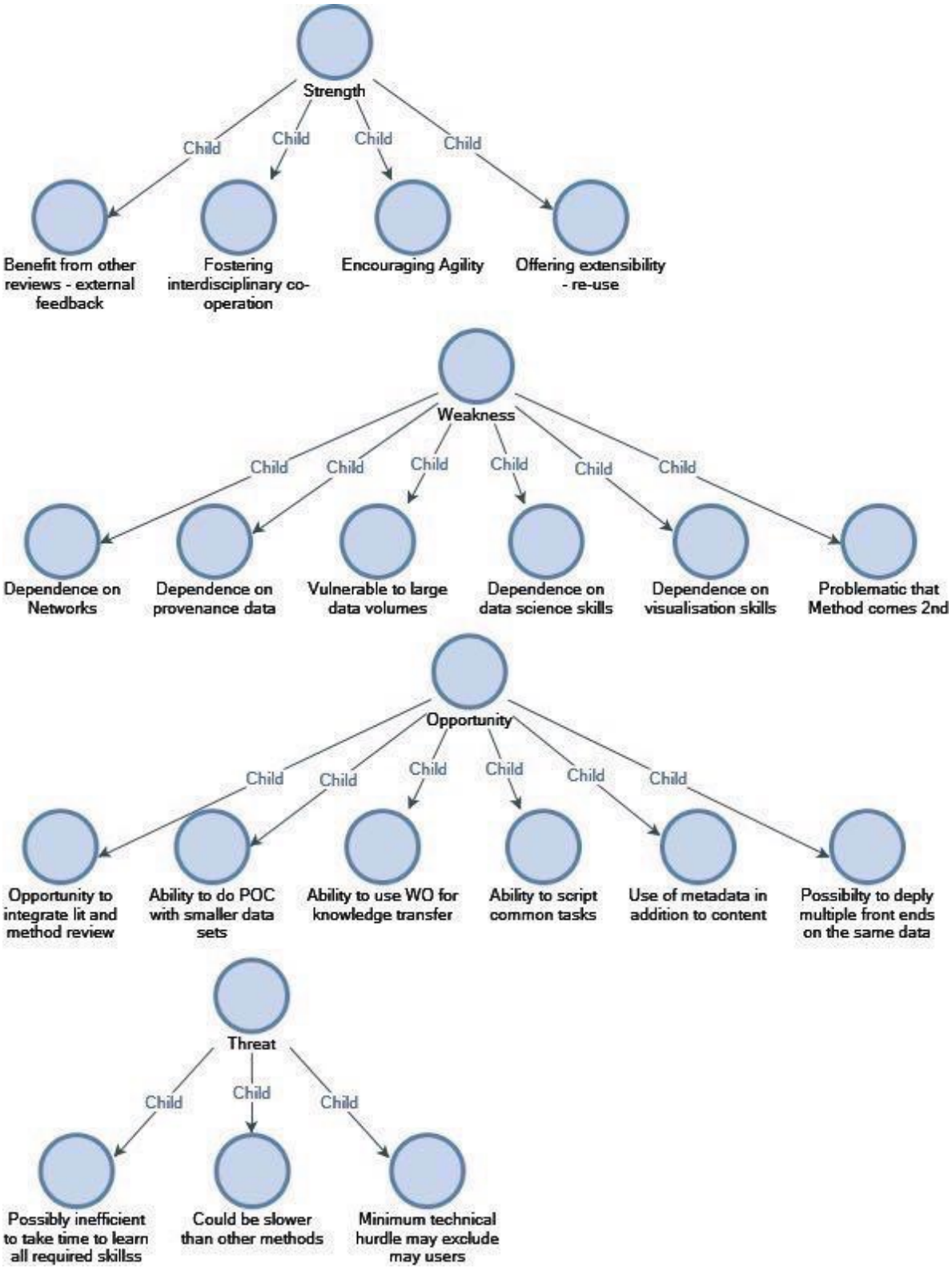
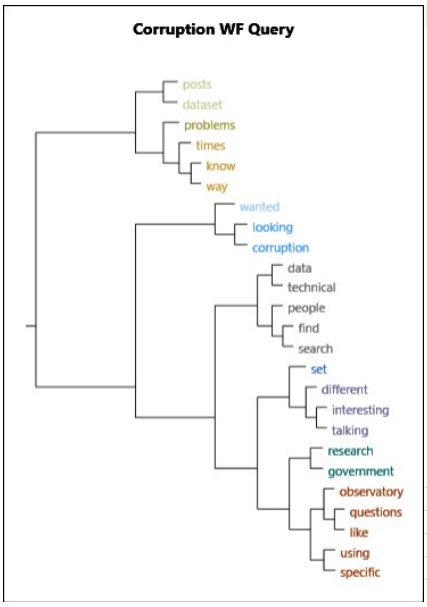


Figure 7-8 Corruption group SWOT model



Focus Group Terms	Payload	Extracted Terms
Ability to do POC with smaller data sets	test data	corruption
Ability to script common tasks	scripting/automation	data
Ability to use WO for knowledge transfer	documentation	dataset
Benefit from other reviews - external feedback	collaboration	different
Could be slower than other methods	slow risk	find
Dependence on data science skills	skills data	government
Dependence on Networks	depend network	interesting
Dependence on provenance data	depend prov	know
Dependence on visualisation skills	skills visualisation	like
Encouraging Agility	Agile	looking
Fostering interdisciplinary co-operation	co-operation	observatory
Minimum technical hurdle may exclude may users	skills technical	people
Offering extensibility - re-use	re-use	posts
Opportunity to integrate lit and method review	integrated method	problems
Possibilty to deply multiple front ends on the same data	multi UI	questions
Possibly inefficient to take time to learn all required skillss	slow skills	research
Problematic that Method comes 2nd	Method problem	search
Use of metadata in addition to content	Meta data	set
Vulnerable to large data volumes	Slow data	specific
		talking
		technical
		times
		using
		wanted
		way

Figure 7-9 Corruption Group transcript vs. autocoder themes

7.4.4 Overall Group Themes

Summary of the SWOT themes contributed by the three groups highlighted the following clusters as follows.

- Process themes
    - Skill sets
    - Familiarisation
    - Small (practice) datasets
    - Communication with other researchers/language
    - Communication with the individuals who harvested the data
    - Documentation.
- Motivation themes
    - Recognition/Citation
    - Curiosity
    - Social Status: Teamwork/cooperation
    - Ability to act (through learning).

- Positive themes
    - Teamwork
    - Interdisciplinary
    - Language skills
    - Extensibility
- Negative themes
    - Networking/Compute limits
    - Skill sets
    - Provenance
    - Data Quality/Completeness
    - Metadata
    - Time pressure
    - Language skills
    - Documentation
    - Shareability

7.4.5 Follow up & related papers

Interviews

A total of six follow-up interviews were planned (two per group with one unable to attend) and five were completed. The group themes were validated against individual responses and given the more private setting of the individual interview a more critical stance was enabled vs. the group setting.

Questionnaires

Three questionnaires were created (see Appendix) and sent to each follow-up interview participant to confirm facets around functionality, process and motivation as suggested by the content analysis and pilot project observations.

- From the three questionnaires, there were 18 possible responses from which 66%, 66%, 33% response rates respectively were obtained.
- 83% completed one or more questionnaire

- 50% completed two or more and
- 0% completed all three questionnaires.

Many responses were incomplete, and the decision was made that the questionnaire data lacked sufficient depth to include in a group comparative analysis. Individual responses are included in the discussion. I believe the questionnaires proved to be too detailed and oriented to the facet analysis itself and not to the participants' experience making a response very challenging for them.

Papers

Three additional papers were later authored relating to the 2013 research event, and these have been considered in the discussion below.

- Papadaki et al., 2014. (Tribe C)
- Halcrow et a., 2014 (Tribe A)
- Beeston et al., 2014. (Tribe A)
- No additional paper submitted by Tribe B

Longitudinal confirmation

Follow-up group interviews/observations were conducted with six student groups and members of research staff from WSTNet member labs at a subsequent WO workshop held in Singapore in 2014 to check for reproducibility of results and any significant developments in the student perceptions.

Whilst the research topics covered by the 2014 groups were different from the China event many of the issues raised in 2013 whilst "observing at moments of breakdown" were reprised. The general positive sentiment regarding the potential of WO and more negative sentiment regarding accessibility/skill requirements were flagged at this subsequent event though it should be noted that some of the themes may be inherent in "hackathons" rather than specific to WO development.

7.5 Consensus/Feedback

The Tsinghua and Singapore events created an opportunity to do end-to-end cross-cultural collaborative Web Science research in a compressed timeframe using the newly released (and at that time largely untested) SUWO. It is, therefore, important to differentiate between hackathon issues, specific WO issues, issues around "doing Web Science" and issues around cross-cultural collaborative research generally.

This provided an ideal basis to test/refine the seed model delivered by the earlier work. Since much of the discussion was centred around the pragmatics of usability/performance/skills I was able to engage in what (Star & Griesemer 1989) refers to as "observing during moments of breakdown" - something not afforded the researcher reviewing project reports, carefully orchestrated demonstrations or academic papers. Whilst all the interviewees found positive aspects and benefits from the WO experience I was aware of an element of “front-stage” behaviours/comments made in the presence of other team members and senior figures compared to later “back-stage” commentary offered anonymously.

The one-week format created artificial time pressure to complete the work while maintaining a much higher expectation on the deliverables than from a traditional one/two day ‘hackathon’. All of the teams reported on the time constraints relative to the volume of work required and the skills/resources available. Although these are not unimportant issues per se, they are not issues uniquely ascribable to WO. They do present an opportunity to see where failures occur "at the seams" of an evolving or unrehearsed process/system. The key feedback (for the WO at that time) comprised:

Skills acquisition - WO has little embedded support so far for researchers without technical and programming skills and so researchers with multiple skills (or multi-skilled teams) are required to complete the entire workflow from data to preparation/analysis to visualisation, deployment and documentation. Easier approaches to building interfaces, template code and integrated tools may be required to engage non-technical researchers.

Network Access/Capacity - most of the participants struggled with technical access issues to datasets on SUWO, which was hosted in the UK and was accessed via comparatively slow wireless internet access and via comparatively restrictive firewall policies set at the Chinese host university also making alternative sources of data difficult to obtain. This makes a highly significant point about the appropriateness of the WO→W^3O approach for users in locations with poor accessibility due to low bandwidth, poor infrastructure or artificial access restrictions or those with limited budgets to fund the requirements to access, host and analyse large datasets.

Cultural/linguistic issues - Each team was multi-lingual, and all had native Chinese speakers. While the communication issues between team members were apparently minimal, the ability to do cross-language and cross-cultural analysis would not have been possible within the WO itself which offers no language translation tools. One group solved this by manually translating/interpreting source material and defining suitable Chinese language hashtags/query terms via the team while another also leveraged third-party translation tools.

Trust: metadata, provenance, quality - all the teams reported 'dark data' issues in terms of understanding what their data was, when/where it came from or the scope/features it offered. It should be stressed that these critiques do not relate to WO or SUWO per se but to the challenges around using secondary data sources in terms of clear provenance, documentation and metadata. Issues of trust, quality and accountability come out as significant issues for WO→W^3O and must be reflected in the wider conceptual model of the WO beyond a technical definition.

Re-usability - The participants were forced to compromise elements of re-usability (using hard-coded datasets or queries) in their final systems due to time pressure. The wider issue here is the need to at least manually document resources within WO or ideally to include sufficient metadata around provenance and structure so that automated discovery of appropriate datasets and the terms under which they were created and can be re-used, might be associated with or packaged with the data itself.

Collaboration/Communication environment - the idea of sociality within the Social Machine and the ability to communicate key information about apps and sources through technical or viral means seems bound to the idea of W^3O. However few (if any) examples of WOs are designed ab initio to support communication/collaboration whilst those systems that do offer this (Zooniverse and other Citizen Science platforms) are arguably not WOs in the strict sense.

The key feedback (for the WO at that time) can be summarised along four main clusters:

1. The importance of data (quality, provenance, documentation) – no *dark data*
2. The importance of process/system (skills, capacity, access) – lots of tools, tutorials and bandwidth
3. The importance of outcome (ease, reuse) – better UI's and ways to curate/document/communicate
4. The importance of people (language, culture, collaboration, teams) – multi-cultural adaptations and ways to work together across cultural divides.

7.6 Reflecting on Outputs

SWOT characterisations - the idea of casting the WO experience in terms of abstract benefits and risks proved to be effective and much more accessible/engaging for the participants than either the taxonomy or questionnaires based on the taxonomy. This allows for the idea that users do not simply act or behave but rather they do so for a reason or in response to some other agent or piece of agency which makes this model inherently more social than the taxonomy. Problems or deficiencies here include:

- No linking of SWOTS to drivers
- No structure or links between SWOTS or drivers
- Difficult to model inconsistent and even contradictory evaluations
- Inconsistent understanding of SWOT terms.

This gives rise for the need for a more nuanced modelling approach which embeds a clear definition of notation, allows for multiple (even contradictory) viewpoints and evaluations which can model physical system, but goes on to integrate with abstract concepts.

7.7 Reflecting on Methods

While automated text analysis can function as a supplementary tool to highlight trends which might otherwise be missed, a poor overlap was noted between the SWOT concepts manually constructed with the focus group and the terms suggested as most relevant (frequent) by automated lexical analysis. Using a Jaccard distance [35] (1-Jaccard Index) where Index of 1.0 is completely similar) revealed a *dissimilarity* (distance) of:

- Academic = 0.863
- Business = 0.875
- Community = 0.927

While many of the automatically identified terms were helpful or "of interest" it would not have been sufficient to use this method without the accompanying qualitative analysis and desk-checking.

[35] The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets: J (A , B) = A ∩ B. A ∪ B.]

Coding analysis by team type (ABC) did not reveal obvious clusters - perhaps due to the artificial (simulated) nature of the ABC division. All the researchers were, in fact, academics (vs. entrepreneurs or community professionals) thus rendering any contextual ABC distinctions artificial and largely uninformed by B/C occupational experience.

Similarly, the questionnaires developed for this exercise proved to be too complex to complete in an unattended fashion and, on reflection, were too centred on the analysis of the WO rather than the experience of the WO and so were very difficult for non-specialists to complete.

These insights led me towards methods (such as IPA) consistent with:

- Using automated analysis only to seed models and to complete the analysis by hand and in conjunction with the "voice of the participant"
- Steering away from questionnaires and towards a method of engaging with something that participants were able to comment on – namely their own experience and idiographic views on WO
- Selecting participants with deeper contextualised "experience" in WO rather than discrete, 'individual experiences' available from less experienced (albeit more accessible) participants such as students/general public
- Visualising in a way that allows flow/structure/sentiment/annotation - all missing from SWOT diagrams.

7.8 Discussion

The groups were neither building nor operating the WO platform itself but rather building an application *on the WO platform* which was required as a last stage of the research process. This afforded the students a place to display/share their visualisations, but (for some) did not adequately supporting the earlier research process itself. As a consequence, some of the detailed questionnaire components were poorly adapted and too challenging for the student groups. Confidence in the completeness/accuracy of this questionnaire data was low.

The selection of projects (by the organisers) apparently from Academia, Business and Community was fortuitous as this reflected existing elements of the WO taxonomy and supported a differential analysis of themes and worked well to help design the analytical approach.

The ABC perspectives represented by the three research groups were to some extent *simulated* since all the participants were, in fact, Academic researchers/students at the time of interview. While several claimed previous business and government experience, these participants were not necessarily allocated to the corresponding A/B/C group.

It should be noted that while a significant proportion of the participants apparently had effective passive/reading knowledge of English, that a number were unintentionally (but effectively) excluded from participation in the focus group/interviews due to a lack of a shared language (Chinese/English/Korean) between the participants and the researcher.

The visual SWOT analysis was effective in determining the relative importance of events and presented a very low barrier of entry for participant participation. It did not, however, offer any way to connect or create causal or linked issues. From this, the more refined Triz notation was later adopted.

Despite a peer review of the questionnaire format beforehand, an attempt to confirm the range of extracted facets with the community of practice proved to be too confusing/complex via (an unattended) questionnaire. I noted a poor response rate of fully completed questionnaires and the responses were too sparse to offer a comparative ABC model. In response, an iPad solution for face-to-face interviews and workgroups was developed but the IPA interview/analysis was ultimately selected instead.

The use of lexical analysis to test concept frequency as a proxy for concept importance showed despite quickly generating a helpful *candidate* list, that manual qualitative investigation and validation is required to re-prioritise concepts from the raw frequency presentation to build confidence in model accuracy.

7.9 Conclusion

In this section a pilot analysis was conducted with three groups of international students to trial questionnaires, visualisations and the straw man models. The decision to discard questionnaires in favour of interviewing was made due to poor results from structured questionnaires and sentiment/narrative approaches were suggested instead. The speed vs. fallibility of automated techniques was explored so that such tools could be used with a clear sense of their limitations in order to achieve the best combination of machine/manual techniques.

Confirming this earlier feedback with six further groups, the ability to observe researchers using WO “in the wild” and the ability to compare the filtered ‘front-stage’ language used in focus groups and papers with the ‘back-stage’ language used in anonymous interviews, provides useful input to issues of accessibility/usability for WO.

Whilst some of the issues reported relate to the use of early (pre-production) versions of WO this nevertheless highlights an interesting broader set of issues that may be important for WOs and

WO adoption (consider innovation resistance) in the future particularly where network/compute resources may be restricted, intermittent or of poor quality.

In the next chapter participant interviews will be considered as they are used to validate and enhance the candidate models derived from the seeding and pilot phases.

Chapter 8: Participant Interviews

In Short ..

In this section, I consider three sets of interviewees split into social groups or "Tribes" with each having a distinctive occupational frame for his or her use of WO. Academic, Business and Community participants were interviewed variously in focus groups, one-to-one interviews and finally three participants from each group were singled out for a more detailed n=9 IPA analysis.

In line with the IPA method (Smith 2009), a summary/model of the themes and excerpts and commentary on participant interviews are provided representing the double hermeneutic: "the voice of the participant" and the interpretation of the researcher. Substantial transcript annotation, notes and analysis were generated during this process and much of the analytical output has been moved to appendices whilst the majority of the interview transcripts/recordings themselves are not included in the final report in the interests of both brevity and anonymity.

Themes/narratives are documented and cross-tribe perspectives are identified that may assist in managing targeted communication and interactions between WOs and WO user groups.

8.1 Academic Tribe

8.1.1 Introduction

In this section, I consider the Academic tribe perspective with focus placed on groups belonging to, or working with the Web Science Trust and also interviews around the Astronomical Virtual Observatory (VO), cited as a template/inspiration for the WO. A total of 31 academic interviews were conducted each lasting between 30-60 mins. All were reviewed, with summaries being thematically coded as part of the evolving draft model (Ch4). Nine were selected based on WO content/focus and were transcribed/coded in more detail against the emerging DNA facets (Ch9). Supporting/contradictory concepts were used to refine the model.

Three IPA interviews were conducted to develop themes from the WST group for a phenomenological (IPA) analysis highlighting ideographic perspectives of the WO meme.

8.1.2 WST

In this section, we consider three interviews representing the WSTNet. The Web Science Trust (www.webscience.org) operates as a not-for-profit organisation promoting Web Science as a research discipline and advises on Web Science related issues. WST comprises:

- The WST Board of trustees from academia/industry who produce guidance on research, education, policy and advise at a government level on matters relating to Web Science
- The WSTNet spanning 20 university Web Science research groups globally who actively engage in Web Science research projects and train Web Scientists
- The WST Admin team where I have (ob)served the board and the WSTNet over 4 years.

Three senior representatives from WSTNet: [Imelda], [Ted] and [Ivan] have been selected for IPA analysis to represent convergent/divergent conceptualisations of WOs framed within academia.

8.1.3 Findings

The interviews share common themes resulting, in part, from the semi-structured nature of the questions and the occupational framing inherent in the dialogue. When comparing the main themes emerging from each analysis shown in (Figure 8-1) there are several overlaps/clusters suggesting convergent themes.

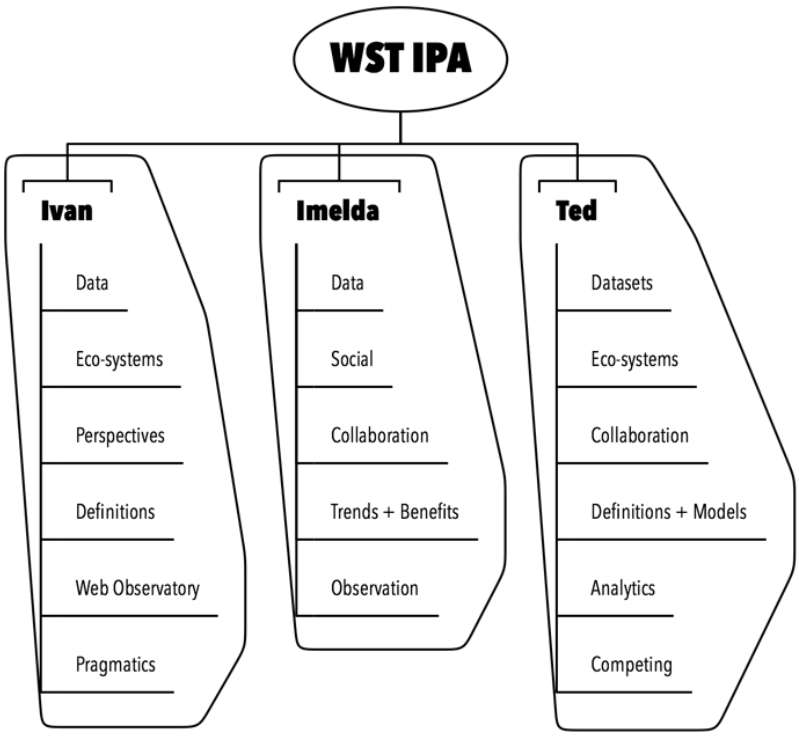


Figure 8-1 Main themes for academic tribe

This shared academic narrative at its most fundamental is that of diverse data, captured/standardised at a technical level and shared (or not) for social reasons, enabling co-operation/competition between social groups around the actionable insights that observation and analysis of combinations of this data may deliver. Each participant, however, emphasises these same elements from a distinct perspective/persona and has a different level of association/empathy with other (non-academic) uses for the data.

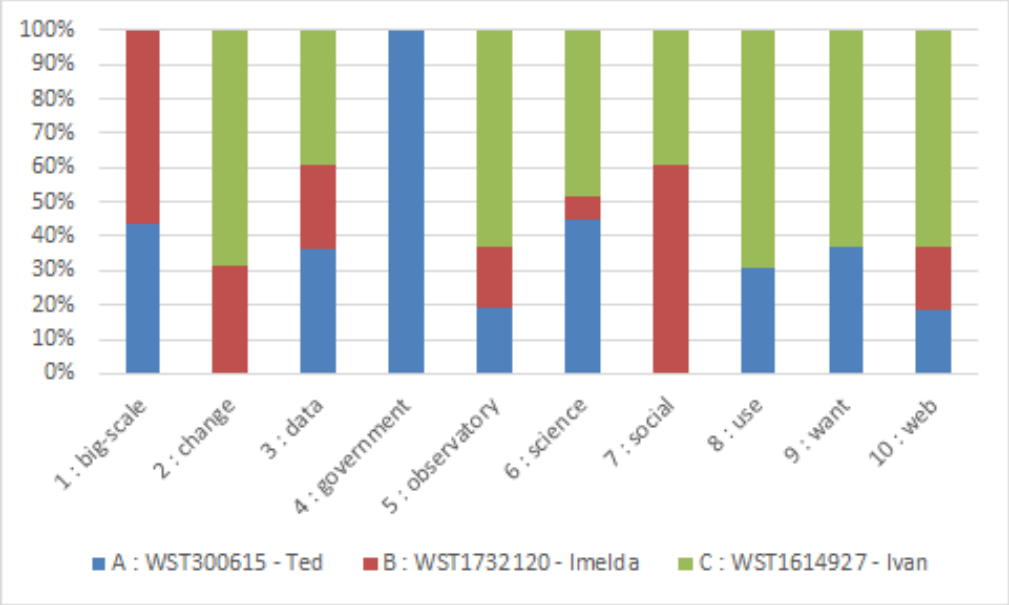


Figure 8-2 Coding Frequency across IPA Academic interviews

The qualitative/thematic analysis above is supplemented here by lexical analysis. Whilst earlier attempts in the project (see Ch4) cast doubt on raw frequency of unique terms as a suitable measure of importance - the technique has been refined here to employ groups of words grouped by meaning (the nVivo lexical tools identify similar meanings to ensure consistency across interviews) and these are clustered by topic. Relative importance is highlighted here with the scores closely matching my personal impression of the key areas of focus. Three narrative sequences were produced [36] from the word frequency analysis of the participants top 10 ranked terms (excluding stop-words) in order to create an intial characterisation or "gist" of the interview themes.

- [Imelda] Data-WO-social/people-sharing-big-science-globally
- [Ivan] WO-data-people-want-compare/research-[to]-use/apply-learning-[whats]-missing
- [Ted] Data-WO-understand-work/research-much-important-government

[36] Creating a pseudo-sentence based on the most frequent terms used in the interview transcript. This is akin to a sequenced word cloud for summary/focus analysis.

These are in a sense “emergent” ideas structured/summarised by frequency rather than by specific intention (i.e., they are **not** quotes) but this nonetheless offers an interesting comparison between the participants.

Comparing the recorded themes with the frequency data:

- **Big** - [Ivan] communicates less of the general scale of the WO application, framing WO as a pure academic application. [Imelda/Ted] use wider multi-tribe frames.
- **Data** - All three participants flag data as the key element
- **Going** [in the sense of working] - [Imelda/Ted] are more concerned with making WO work with other parties than [Ivan] who does not see WO as commercial
- **Government** - [Ted] specifically uses the Government frame both as a WO user and an enabler of WO
- **Observatory** - [Ivan] sees WO as a pragmatic work-in-progress and like a more traditional observatory rather than a dataset repository
- **Science** - [Ivan] frames WO in the scientific metaphor
- **Social** - [Imelda] is most concerned with the social/sharing aspect of WO. Given my impression from the interview I was surprised that [Ted] did not show up here.
- **Use/Want** - [Ted/Ivan] frame the WO as a pragmatic application/solution compared to [Imelda] who is less focussed on immediate goals with the data rather than what having the data long-term may enable.
- **Web** - All three participants include Web as a key component but [Ivan] particularly stresses data-ABOUT-the Web over non-Webby data.

[Imelda] is someone for whom keeping the data safe and available for posterity is paramount and who focuses on enabling the value/insight of this data to all groups who can benefit. She calls for broad collaboration and the release of data assets for the good of all in a visionary sense which transcends current technologies or applications to underpin future research questions/systems - playfully invoking Asimov’s notion of *Psychohistory* as an inspiration for the promise of a fully-mature global WO. She sees (data) observatories as a much broader concept than Web Science but views this not as a problem per se but rather as a source of greater scale and inter-disciplinary collaboration. [Imelda] sees this as a step-by-step endeavour starting simply with sharing data about data (as links/catalogues) in order to build trust and social ties that will underpin the development of collaborative and trading relationships.

Whilst [Ted] values collaboration, he is acutely aware of a "land grab" for data: the competitive environment both between companies but also between business and academic research teams for access to valuable data which is increasingly detailed, personal and therefore sensitive and

less tractable to sharing in the public domain. The move from "data-about-business" to "data-AS-THE-business" places a disruptive pressure on the way that academia interacts with business in terms of a new competitive/disruptive tendency. Ted is not oriented towards personal advantage here (he wants data to be made openly available for a scenario where "everybody wins"). He does, however, depict the current situation as a serious innovation race (or “land grab”) between multiple groups in multiple ecosystems to provide earlier, deeper and more valuable insights without tripping the increasingly sensitive alarms (and growing punishments/fines) around the abuse of private data. He speaks less about the resource or content itself but rather about the goals and what is enabled by the resource in research, business and government. New models based on openness and collaboration rather than outsourcing/subcontracting are stressed because the former promotes open data and open research outcomes while the latter concentrates and commercialises data in the hands of the few.

[Ivan] focuses on promoting the formal structures/standards that will support the endeavour and to some extent as a pragmatist - offsetting what would be *ideal* against what is *achievable*. He scales back and "socialises" grander automated visions of the Observatory to simpler cataloguing and point-of-contact solutions which he believes fulfil the essential utility of the WO. This more pragmatic WO answers such as “what is available?”, “where is it?”, “who do I talk to to get it?”. He cites broader limitations in funding and IT-as-it-is-actually-deployed (rather than what is theoretically possible) and believes that these simpler ideas, if managed consistently and with a strong focus on Web Science, will form the core of grander WO systems as richer ecosystems evolve. His focus on adoption and practicality shows a sensitisation to social dimensions over technical challenges and he expresses concerns about too much diffusion around the idea of the WO and the risk of becoming "another repository" rather than specifically a place where Web Science can (should) be done using data about the Web.

These three emerging themes of data vs structure vs application are identifiable across other academic as well as more broadly across tribes.

All three participants identify with the benefits/challenges of the Observatory metaphor taking, in some cases opposing views, but all recognising the attraction of recognisable paradigms for WO that produce rapid empathy/understanding and span "tribes".

All three participants seem a little uneasy about walking a fine line between openness/inclusiveness and the idea of WO being too generic/indistinct. This would perhaps risk WO being subsumed into the much larger and growing field of big data tools, social network analysis and digital marketing and, due to the close identification of WO with Web Science, blur perceptions of Web Science along with it.

All three participants suggest distinctive things about Web Science and Web (Science) Observatories which help to focus/characterise the work:

- The simple, but increasingly ignored (and perhaps transparent) criteria that data are *about* the Web and not simply delivered *via* the Web
- That they are persistent, repeatable and longitudinal - unlike search results in the Google-dominated Web ecosystem
- That they act as a focal point bringing researchers together to enable *collective intelligence* for results which are more comprehensive than those of any lone group or system and enable the study of Social Machines, which supports a new class of computing paradigm.

All three participants were asked about defining WO giving a number of interesting/convincing definitions of WO within different contexts including:

- WO-as-a-project
- WO-as-a-system
- WO-as-collections-of-data
- WO-as-a-paradigm-for-research
- WO-as-a-community-of-researchers

(which we shall return to in Ch10) and yet each broadly and repeatedly conflated, mis-defined or otherwise "airbrushed" the term 'Observatory' in other parts of the interview when not speaking directly about the difference/definition. It is both notable and important that despite [Imelda], [Ted] and [Ivan]'s atypically high exposure to the terminology and their unanimous assertion that "airbrushing" is largely unintentional in that conflation/context-switching, even for experts, seems very strongly embedded in the inherent ambiguity of the term:

i.e., Web Observatory as:

- Observatory-***on***-the-Web
- Observatory-***about***-the-Web

and this alone is highly likely to maintain a tension between the idea of practising 'a-Science-of-the-Web' versus practising (any) 'Science-on-the-Web'. There is ambiguity around the term Web (at least in English). This supports (causes?) cognitive flexibility around the idea of WO (seen in Ch4) that WO is represented/framed in a variety of different (potentially incompatible) ways by different users/stakeholders who nonetheless appear to be talking about the same thing. In other words, the WO meme has a primary central definition shared by users which "flexes" as

stakeholders apply the idea and the more specific requirements of their own environments/frame of reference. They do, however, return to the centrally agreed (if fuzzy) definitions when discussing outside of the personal/professional frame something (Star & Griesemer 1989) termed *a boundary object*.

Recognising WO as a potential boundary object gives us insight into how stakeholders can be expected to react, how consensus/adoption may be affected and ultimately how WO may decompose and evolve into new, more specialised objects with different central definitions/constraints as (Bowker & Leigh-Star 2000) predicts for classic boundary objects.

Similarly, when we consider the adoption of WO as an innovations, (potentially disrupting existing social models and power structures), it is important to consider how cognitions around risk/benefit (contextualised by tribe) are central to many of the explanatory adoption theories. This includes innovation diffusion (Rogers), innovation resistance (Ram), disruptive innovation (Christensen) and TAM (Davies) which are addressed in Ch2/10.

Earlier academic data observatories such as the VAO teaches us that socially-constructed meanings and contextual challenges are as important (if not more so) for adoption that specific technical solutions. I am indebted to IVOA team members for their frank insights around human factors that are generally not included in academic reports/papers. For broader adoption to be encouraged, the notion of winners/losers must be avoided. Neutral ground (through the auspices of W3C, WSTnet and the Web Science Trust) will help to promote agreement and the adoption of community-sanctioned approaches and platforms rather than engendering a tribalistic "not invented here" mindset around any standard/platform that might be suggested.

8.2 Business Tribe

8.2.1 Introduction

In this section, we consider the Business Tribe perspective. There is an introduction to the organisation being studied, a summary of the interview themes, key points and, where available, additional datasets/documents that have been studied.

A total of 23 business interviews were conducted each lasting between 30-60 mins. Each interview was evaluated/filtered for relevance, and all were summarised with the summary remarks being coded. Post-filtered interviews were transcribed and coded in more detail against the DNA candidate model and supporting/contradictory concepts were used to adapt the model. Specific focus was placed on [DataCo] where nine interviews were conducted to develop case-

specific themes and three were selected for a phenomenological (IPA) analysis highlighting ideographic perspectives of the Observatory meme. Additional concepts from the other business interviews (including TAMR, OKF) as they relate to this case are collected in a group summary, and the overall analysis of Business vs. other Tribes is presented in Ch9.

The pre-analysis and proof of concept study identified the class of business WOs as those intended to operate in return for financial remuneration - producing financial capital, adding value to raw materials, adding shareholder value and increasing market share. While these ideas typically also imply "making a profit", I have specifically avoided the simpler but less useful concept of "making money". This is a common motivation for *all* Tribes (and therefore not useful for distinguishing between them) since money is essentially a means of dematerialising value typically to enable the exchange of some other value (i.e., tax revenue → Government services, research funding → new knowledge in academia).

Specifically, in the [DataCo] case, the system explicitly neither makes a profit nor makes money directly - making it of particular interest and requiring a more nuanced model to understand why the business would choose to offer such a service.

8.2.2 [DataCo]

[DataCo] was formed approx. ten years ago by a merger between two major providers in the data market. [DataCo] are a news and media corporation (with additional capabilities in Tax, Legal and Patents). They provide curated news, analysis and pricing information for financial markets and also operate in the academic/knowledge management space.

[DataCo] are noted for their commercial services and proprietary technologies for the acquisition, curation, management and distribution of multiple sources/formats of data via standardised platforms and APIs. They are also deeply engaged with issues around open data and shared knowledge - a paradigm which figures strongly in this case study.

Integrating/disambiguating multiple information sources has been a long-running internal project for [DataCo], triggered in part by the original merger. This case looks at the motivation for opening up this capability to [DataCo]'s clients as an open framework designed to locate, marshal, disambiguate, integrate and deliver hybrid information sources from both within and outside the direct ownership/management of [DataCo].

The ambitions and the processes/challenges for [DataCo] show a striking resemblance to some of the processes/challenges highlighted in the initial analysis of Web Observatories and form a useful proxy to understand the journey from standalone systems (WOs) to co-operative WOs (WO → W^3O).

A note on context for the research

I had previously worked for a company acquired by [DataCo] and subsequently for [DataCo] directly over a period of several years prior to commencing this research project but at the time of interviewing was not employed or remunerated by [DataCo].

The original scope had included interviews with users of [DataCo]'s [DataNames] product to compare internal/external perspectives on the service. Access was not possible in the timeframe and remains open as potential future work.

8.2.2.1 Interviews

There were nine participants overall in this case whose roles ranged from IT management to strategic product/resource management to technical architecture and design. The common thread was a connection to [DataCo]'s strategy and the deployment of shared persistent identifiers (and subsequently shared open persistent identifiers) for users and resources under the [DataNames] programme. Each interview was coded and compared with the evolving DNA model to enable both a confirmatory analysis and also to enable an Academic vs. Business vs. Community factor analysis.

Three interviews, [Quinn], [Charlie] & [Thomas], were selected (purposefully sampled) for additional IPA analysis to gauge more subjectively how each was framing the WO concepts to investigate whether elements/approaches to WO are localised to one tribe/sector or can be observed across sectors.

While the notion of "Web Observatory" is not an explicit theme in the interviews that are reported here, the reader will note relevant parallels in which [DataCo] operates to discover, acquire, marshal and curate diverse data sources from multiple sources. This is both from partners and from within its own group of companies and involves assembling these into coherent data products (databases, data feeds and applications) in a way which offers notable similarities to the concept of scaling up WO→ W^3O.

I refer to [DataCo]'s [DataLake] and [DataNames] service as examples/analogues of approaches that mirror WO features to discuss the technical, operational and cultural challenges that emerge when creating Web-scale data platforms comprising several elements which are referred to

during the interviews. It should be stated that [DataCo] is currently offering Data-ON-the-Web and not Data-ABOUT-the-Web but even in this case such systems remain a potential source of data/services for W^3O.

8.2.2.2 Systems referenced

[DataLake]

[DataLake] is [DataCo]'s centralised content hub which identifies numerous sources, documents and dataset assets throughout [DataCo]'s federated estate and allows these to be combined into new products/services based on authoritative naming and publishing standards i.e., without the need to centralise the storage and processing of all the content.

[DataNames]

[DataNames] is an open standard/service comprising an API and registry of machine-readable identifiers that can be associated with a piece of information as a canonical term intended to transcend multiple human-oriented systems and schema and to bridge the gaps between multiple references to the same data item.

[DataTag]

[DataTag] is a commercial [DataCo] service comprising an API and Web portal to tag (semantically) textual material/content (i.e., extracting 'entities' from documents). This results in the identification of organisations, people and financial instruments from news stories, legal texts, patent filings and other documents resulting in an RDF encoding/linking to curated information about these entities.

TAMR

TAMR (http://www.tamr.com) is a commercial company spun out of MIT based on the Data Tamer project by (Stonebraker et al., 2013) and (Gubanov et al., 2014) addressing the challenges of data curation at scale through expert-mediated machine learning. [DataCo] are leveraging this technology to address the challenges of multiple data sources within their organisation.

Other Sources

- [DataNames] Website
- [DataTag] Website
- TAMR Website + white papers + a 2015 Stonebraker presentation at Hadoop World
- Creating Value with Identifiers in an open data world (ODI report)

- Unlocking innovation/performance with liquid information (McKinsey report)
- Going Open (My internal report for [DataCo])
- Open enterprise (ODI report)

8.2.3 Findings

[DataCo] previously purchased several competing/complementary companies and has acquired (rather than organically grown) a proportion of its data sources, technologies and platforms. It has required strong technical leadership and a culture of robust architecture/technology management to succeed in the assimilation/linkage of several diverse data management and data platforms. The interviews tended to support the impression, however, that due to process/data complexity some of the historical integrations remained superficial ("fixed at the desktop") in terms of connecting end-products/services rather than integrating more fully at an Enterprise Data Management level. Under more recent technical leadership, [DataCo] explained that this trend had been reversed with the drive to centralise, disambiguate and marry up key [DataCo] data sources across its multi-business estate under a single concept known as the [DataLake] platform. Underpinning [DataLake] is the need for more automated data curation and machine readable persistent [DataNames] which are uniquely "minted" by [DataCo] and refer unambiguously and forever(!) to specific pieces of [DataCo] data. This includes specific companies (legal entities) or financial securities (currencies, shares, bonds, options, etc.). These can be used internally to assemble services, feeds and products programmatically across a federated architecture without the need to store and process all aspects of [DataCo]'s large and complex estate of information assets centrally. Whilst this may initially seem to be a straightforward task, [DataCo] shared examples of how apparently unique identifiers have, in fact, historically changed and been reused over time as companies have become more/less powerful and have merged to form new organisations. Persistence of identifiers over time is a central challenge for WOs gathering longitudinal data and providing hindcasting services.

[DataCo] exists in an increasingly Web-oriented ecosystem where much of the content is openly licensed. They have previously been encouraged by customers and regulators to be more "open" about the re-use of standards and naming schemes which were historically licensed under strict commercial terms. Revised license terms and other concessions including the more recent decision to release a portion of the [DataNames] technology under a combination of creative commons licenses indicates [DataCo]'s commitment to openness.

8.2.4 Discussion

[DataCo]'s creation of [DataLake] enables easy access to reliably named canonical sources from multiple sources of information across the [DataCo] estate without physically centralising the storage, formats and processing of the underlying data. Applying the Observatory metaphor: the [DataLake] observes the locations where relevant data is to be found and enables access and linkage to these sources through a *lingua franca* that confers a high level of trust in the naming of entities across the sources. Through this process, internal application developers are able to maintain existing services and develop new ones as sources and internal implementations, storage and processing changes within the source systems.

The insights which [DataCo] brings to the WO →W^3O meme are that centralised storage is much less important (and much less achievable - even with a single organisation) than centralised naming/authority which can be used to refer/link back to data both internally and externally. [DataCo] are also pursuing a second strategy of interest to WO, which is that of automated curation through a hybrid ML (machine learning) + Expert human approach. An important insight here is that human elements in WO can be *producers and/or processors* (curators) *and/or consumers* of data and that this underlines the Social Machine nature of Observatories whose overall operation from source to output will be partly shaped by human decisions and hence socially constructed.

In the following section three interviews [Charlie], [Thomas] and [Quinn] are analysed for additional perspectives on this topic and the core problem of marshalling data at large scale. The technique employed below is based on IPA and comprises a classic 'double hermeneutic': the researcher's interpretation of the participants' interpretation of their experience. IPA stresses the validity of this additional perspective. In the W^3O scenario it is not only the explicitly stated (or indeed the objectively accurate/true) drivers, motivations and measurements that will shape the interoperation between different parties but also the perceived drivers/motivations which itself constitutes a double hermeneutic.

8.2.5 Conclusions

Comparing three IPA participants from [DataCo] gives rise to similar elements for the narrative but is interesting to note the perspectives and emphases are markedly different for individuals on the same team receiving the same interview brief within the same project.

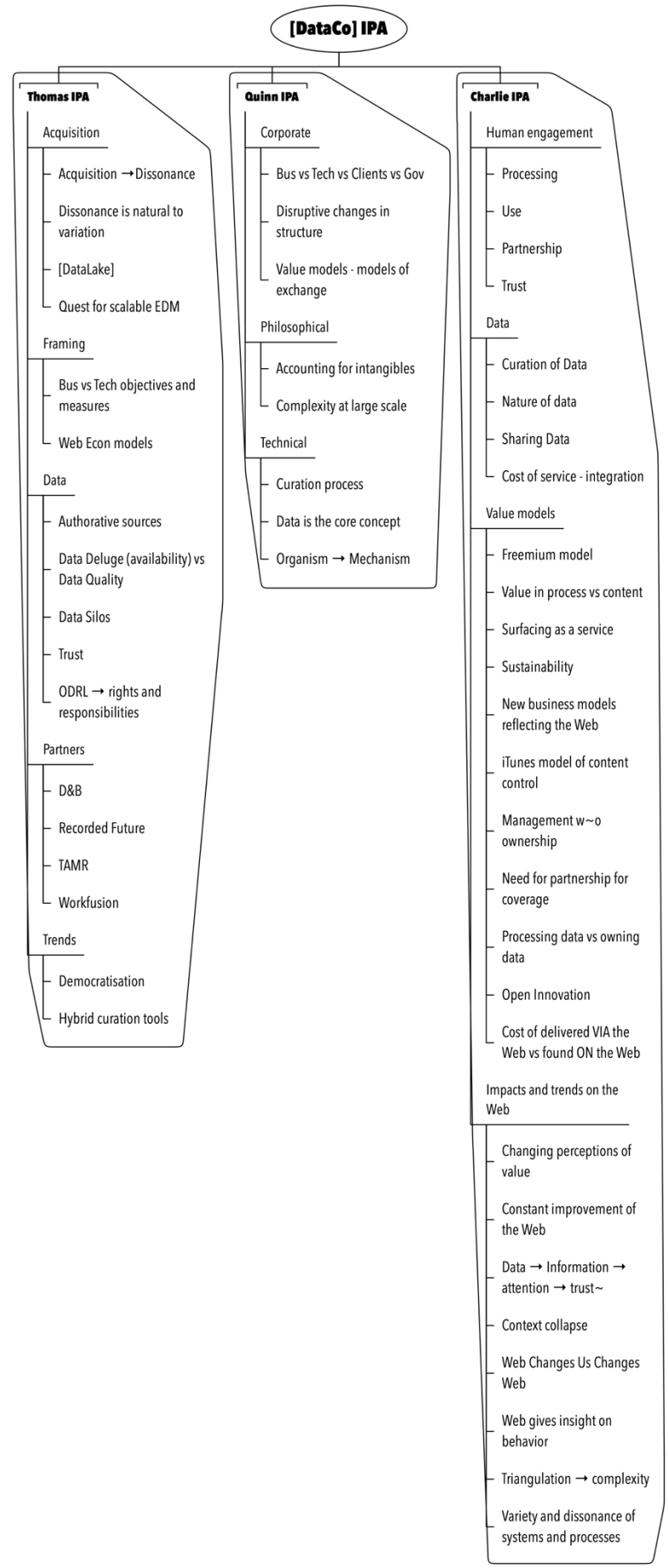


Figure 8-3 Main themes across the interviews

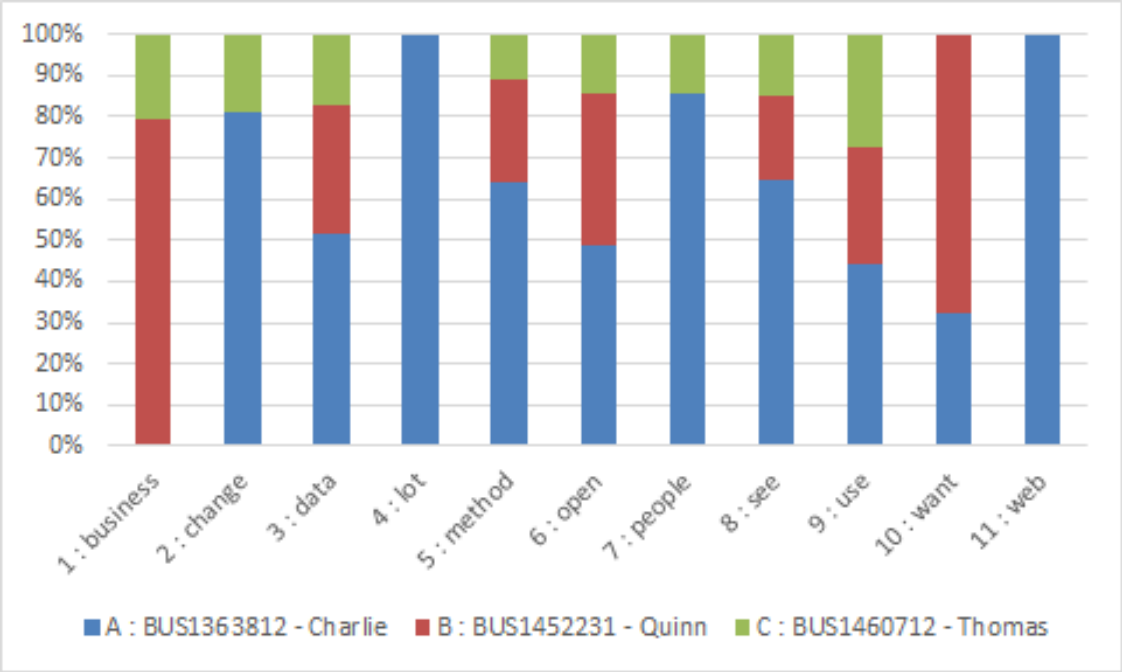


Figure 8-4 Relative theme frequency in Business Tribe

The compressed (gisted) narratives based on the participants Top 10 terms are:

- [Charlie] - Data-Web-people-change/going-direction/way-use/see-sets-open-want
- [Quinn] - Data-right-open-business-use-value-want-way-search-need
- [Thomas] - Data-going-open-market-see-use-licenses-part-take-client

[Charlie] is talking predominantly about the broader Web rather than internal data and stresses the wider Web ecosystem of data and people whilst [Quinn] is focussed on specific methods for the business to meet objectives. [Thomas] offers the most focused/compressed narrative relating to specifically how licenses and open data will affect clients.

There are common themes across all three interviews, and at the essential level, these are technical change →DISRUPTION, a focus on best SERVICE and the notion of how the service fits within a CONTEXT (a corporate context, an innovation space or the Web). Each participant specifically flagged the notion of perspectives between technology, business and markets and this is what Grounded Theory calls "a sensitising concept" around which the success of the project revolved. To understand how each of these groups valued and measured the service was seen as key to a successful deployment and engagement with that group. Each of the interviewees spoke with noticeable professional pride at the level they were framing the concepts under discussion. This came over as a strong commitment to deliver a robust solution to the problem as they saw it - albeit with each participant I saw a slightly different perspective on how they framed the specific problem.

[Charlie] focusses on how the data interacts with other data in the wider Web ecosystem and it is access to, and confidence in, this data which confers value. The specific subjects which are covered in each place are relevant and how they are blended or "re-mixed" is also seen to be relevant to the process. The specific structure and technologies are seen as transitory, as are the applications.

For [Quinn], the individual elements of the data appear less interesting than the structure/process which confers value. By adopting sound design and structural principles, these data can be combined with others to address a range of applications (no specific solution is implied) making them inherently more valuable.

For [Thomas] the data and the structure/process through which it is delivered are simply the most appropriate tool to address specific problems and pursue key opportunities. The elements of the solution are given a lower priority than the fact that a solution is found.

These three elements (content, approach and outcome) are seen as distinct key foci of the [DataCo] team and embrace both open and commercial data with a wider eco-system. In meetings/interviews with the Digital Catapult (DC) about their Trusted Data Accelerator (TDA), we saw a more tool-oriented, almost content-free approach from the offering. DC brings its architecture expertise and template tools to 3rd party organisations who are encouraged to set commercial objectives in the engagement which, it is hoped, will stimulate the UK economy.

Of particular note with this set of interviews is the notion of 'tribes-within-tribes' - evidenced by dialogue around competing business units, differing goals/approaches between commercial and technical groups. Hence a complex and potentially political decision process may underlie adoption of WO for larger commercial organisations representing multiple perspectives (or Laminations as Goffman calls them). Such competing interests may indeed form a proxy for WO adoption overall. (Davis 1989) refers not only to PEOU (perceived ease of use) but also PU (perceived usefulness) and thus considering perceived alignment to objectives must surely be a key factor in understanding/promoting WO adoption in addition to technical compliance.

The challenge of maintaining revenues in an open/shared data ecosystem and a growing resistance to proprietary systems and interfaces will be fundamental to the adoption/participation by commercial groups in the WO ecosystem. Groups like [DataCo] believe they have the experience/architectures in place to interface with open data and transition some users from free to paid services through the provision of trust, provenance and quality services. Considering the vital role that advertising and other paid services play in the support of the Web overall - such cross-funding models may be vital to the broader adoption of WO.

8.3 Community Tribe

8.3.1 Introduction

In this section we consider the group defined as "Community ". A total of 26 community interviews were conducted each lasting between 30-60 minutes. Each interview was evaluated/filtered for relevance, and all were summarised with the summary remarks being coded. Post-filtered interviews were transcribed and coded in more detail against the DNA candidate model and supporting/contradictory concepts were used to adapt the model.

It should be noted that government virtual observatories (VOs) are centred around open data and that the concept of "value" in terms of 'why engage?' in terms of impact, and financial sustainability is a key issue in this group:

> "One of the most frequent questions I get about open data is - 'why?' Most folks understand the need for transparency and openness in government, but some question the need to invest the effort in a comprehensive open data effort .. and to be perfectly fair - very few folks have measured the impact of open data."
>
> **- Joy Bonaguro - Chief Data Officer, City and County of San Francisco**

Open data may be considered to sit in two broad camps: the social/financial objectives of the community (government) groups which emit the data at no explicit cost to the user and the financial/social objectives of the groups who leverage that data sometimes for an explicit fee. An IPA analysis of three project members building a WO solution for the South Australian government is presented.

8.3.2 South Australian Government WO

The Australia and New Zealand School of Government (ANZSOG) has a particular interest in digital literacy within government and in putting research into practice. Through a collaboration between ANZSOG, the University of South Australia (UniSA) and the South Australian government a project was initiated to consider the issue of ageing population in Australia and to build a WO to consider questions in this area:

- To develop the data publishing and governance structures to enable the SA Government to publish its data on the Observatory.
- To develop a methodology to use that data to inform policy making.

- To develop cases which underpin a 'digital literacy' education programme to be developed by ANZSOG together with the SA Government for delivery to other jurisdictions.

Since the UniSA WO is built from the SUWO template, no separate discussion of the WO functionality (the D in DNA) is attempted here.

8.3.3 WO project interviews

11 interviews were conducted across the organisations involved: Three with the SA Government, Four with the University and four with ANSZSOG. The themes from all the interview and document sources are summarised in a thematic overview and feed into the overall community model presented in Ch9. Three interviews were chosen for more detailed IPA analysis to highlight/contrast their conceptualisation of WO in this context. Additional Sources used were:

- UniSA WO site
- Published WO papers
- ANSZOG project proposal
- ANSZOG final report.

8.3.4 Findings

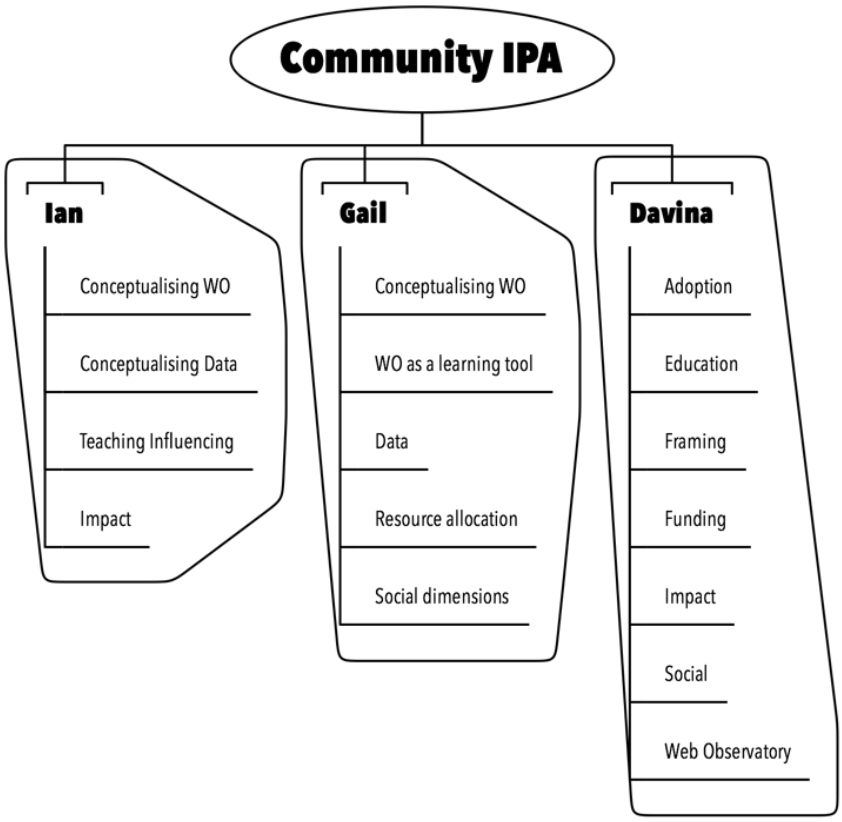


Figure 8-5 Comparing Community IPA theme

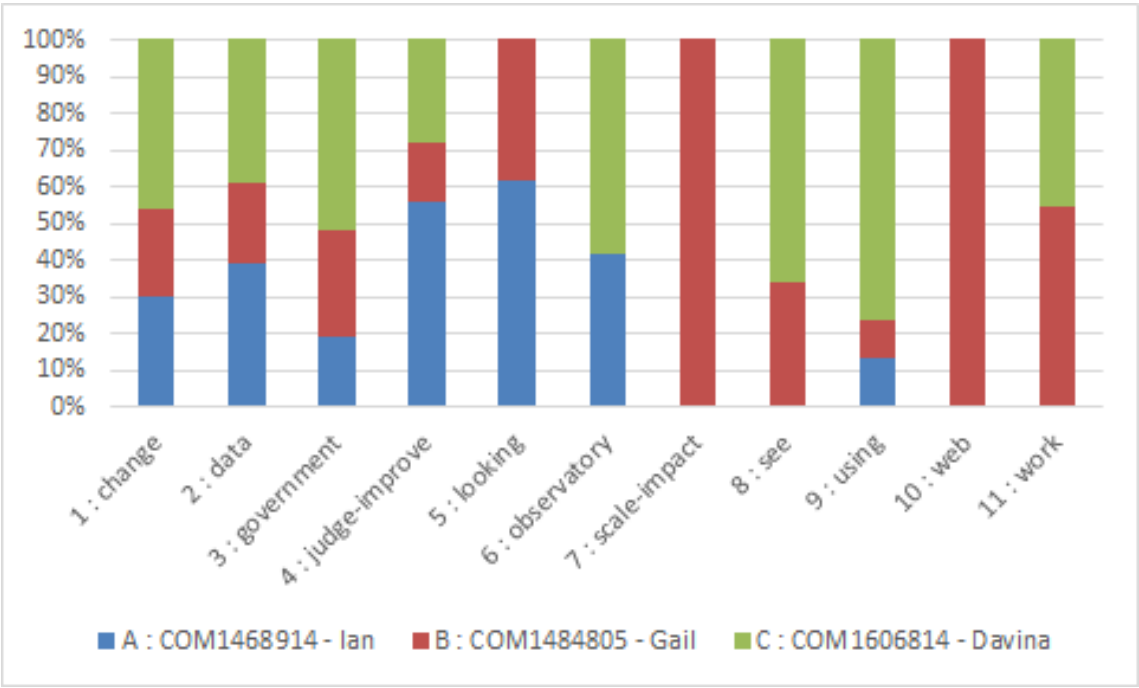


Figure 8-6 Matrix Coding frequency for government tribe

Three condensed narratives emerge from the frequency analysis:

- [Gail] - Data-work-web-looking-scale-change-government-well-using-see
- [Ian] - Data-like-change-open-need-Observatory-effect-understand-people-right
- [Davina] - Data-need-see-people-problems-change-improve-use-make-Observatory.

Our three participants overlapped closely on main occupational frame (agreeing on the key theme of government developing better insights from data but as seen with other cases above offered varying perspectives on the meaning/significance of WO. [Gail/Ian] emphasise having/considering data whilst [Davina] strongly frames WO as a solution for applying/doing something with the data. Gail stresses the Web and its data whilst Ian/Gail are looking at the way in which the Web mediates behaviour and other types of non-Webby data. [Ian] is looking for effective/efficient methods of use to address occupational challenges and along with [Davina] sees this via the WO rather than purely the data.

Less apparent from the frequency was the distinct emphases on:

- The **data** itself
- The arrangement of **systems and processes** to handle the data
- The **impact** of the data in context.

While each of the participants admitted to being vague/unclear about WO and its purpose at the outset, the group when interviewed had an unusually similar (coherent) understanding of WO not only within the content of their own project but in a wider conceptual sense. I concluded this was due, at least in part, to the educational programme of lectures that were tied to the deployment of the system. It is noteworthy that the team felt that the more traditional computer science academic colleagues (although more used to the details of such systems) had failed to appreciate the value/potential of WO - viewing it rather as a familiar (mundane?) collection of technical components rather than something exciting/novel. The contextualisation of WO as an IT system (an apparatus) rather than as an approach to underpin evidence-based policy (a solution) drew markedly different reactions. [Davina]'s remarks point to the flexible re-contextualisation that is associated with classical Boundary Objects:

> " ..It was the fact that they imposed their own concepts on it ..[and]..they put on an overlay .. a bridge between a need and a solution."

Also noteworthy was the issue of the group all placing government agencies *outside* the centre of the WO development community while doing so for different reasons:

- [Gail] felt Government was not the right place to focus on WO for financial (budget) reasons deferring instead to academia as the place where funding was more consistent.
- [Ian] also felt it was not the best place to situate WOs but in his case due to a notable lack of technical skills and insight to deliver innovations compared to academia.
- [Davina] did not specifically exclude Government from taking a leading role - though she did express some frustration at the speed of moving from planning to action - and instead suggested "not re-inventing the wheel" by re-using existing technical templates and leveraging organisations with established skills and capabilities to deliver results to Government.

What we learn from this WO engagement centres around:

- The need to wrap education (digital capabilities) with the introduction of digital concepts and value models (digital literacy) in order to drive a willingness to participate, an openness to the new ideas around value and value exchange
- The need to situate projects within flexible/responsive contextual boundaries (what is valuable to one group may be far less so to another) and an appreciation of potential complementarity
- The pragmatic necessity of suitable resources and funding. A challenging chicken-and-egg problem comprising "no funding without impact" vs. "no impact without funding" which relies partly on the first two elements.

The final summary report associated with this project also touches on the importance of framing/context for analysis and intersubjective agreement:

> "The challenge is how to analyse and interpret this data within the context that it was created, and to present it in a way that both researchers and practitioners can more easily make sense of ..[given that] ..having 'open' data is just the beginning and does not necessarily lead to better decision-making or policy development. This is because data do not provide the answers – they need to be analysed, interpreted and understood within the context of their creation, and the business imperative of the organisation using them."

Thus we see the critical aspect for WOs of tools/models which reflect the contexts in which they are used if we are to understand their adoption, sustainability and successful interoperation.

The report highlights the opportunity here for WO to be a "neutral space" supporting both the education of community members and leaders to access and interpret digital data in a meaningful

and socially enriching way in support of better policy and services. Education is the most fundamental message of the report.

> "The challenge was bringing people on the Web Science journey with us and demonstrating the value of the Web Observatory as both a research and educational resource."

The report cites the existing deficit for citizens, managers and even academics both in skills and also conceptually in understanding the potential and application of such even at current levels of technology where "working with data directly is both daunting and confusing". It calls this "the tip of the iceberg" predicting ever greater automation/datafication/integration as:

> "There need to be people within government who can appreciate the value of what the Web Observatory can provide, at all levels, and are literate in the information and knowledge that it can provide."

Hand in hand with education comes the ethical use of data such that society will volunteer its use:

> "One of the challenges with the opening up of data of all kinds is 'trust': with Government Open Data initiatives all too often there is a suspicion that they will be used for compliance or enforcement activities; with Commercial Open Data initiatives the suspicion is that companies tend to want to corral their data for commercial advantage ..[and]..it is when public and private datasets are combined that the potential value is significantly *amplified*[37]. This is where the Web Observatory comes in, as a platform upon which to view public, private, open and closed datasets with a view to solving a problem or answering a question, but also to providing a far greater transparency as to how data are actually being used."

The report characterises WO as a channel to bring citizens together with the data about the society they live in and with the officials elected to govern that society:

> "By opening up data, citizens are enabled to be much more directly informed and involved in decision-making. This is more than transparency: it's about making a full "read/write" society, not just about knowing what is happening in the governance process but being able to contribute to it."

Ultimately the ANZSOG/UniSA WO project recognises that the ultimate goal of WO is W^3O in which border insights are achievable for those with the skills/mindset to invest in development:

[37] I note the use of this word in the astronomical VO interviews

> "The true value [of WO] comes in bringing together various datasets to gain greater visibility on a specific policy issue or question. This requires not just having the technical skills to publish the data, but more importantly, the digital literacy to understand the broader value of doing so, and therefore championing and supporting such initiatives from higher levels of management."

8.4 Comparing/Characterising ABC users

In this section a cross-tribe and cross-participant analysis is presented which does not replace the researcher's qualitative characterisation but compliments it. Indeed, it highlights different aspects of the data through the measurement of frequency/similarity, though frequency is not always the same thing as importance. As discussed, automated content analyses are to be used with great care if significant (typically semantic) errors are to be avoided, but they are nonetheless very helpful in avoiding bias and ensuring a cross-check that emerging theories are grounded in data rather than the researcher's desired outcome.

Insofar as motivational factors were poorly reflected/differentiated in the study of the technical WO literature (Ch4), motivational models (including Reiss) in turn poorly reflect the idea of technology as a goal in of itself without relying on other more general occupational/status/resource goals which would offer poor differentiation between groups here. Thus a composite model has been structured combining relevant Reiss motivational elements (which were validated against UK Government Open data requests) with other technical elements from the interviews. This reflects the emerging foci on data, technology and outcome which are relevant to WO.

In order to characterise the motivations/goals of users of WOs (the A in DNA) Figure 8-10 shows what CGT calls selective->theoretical coding using sensitising concepts.In this case these are an orientation to Data (Curator role), an orientation to Systems/processes (Architect role) and an orientation to Application/outcome (Innovator role). This model draws on elements of the Reiss motivation model as well as repeated/common themes across the IPA interviews.

This focus is core to the research (RQ1/RQ3) and in particular, is grounded in the desire not to *show that* As, Bs and Cs are different but rather *show how* they are different with respect to WOs and whether this informs the process of encouraging interoperation and the development of a global Observatory paradigm.

Syndicates	0	0
Architecture	9	1840
Activity	9	49
Belonging	9	247
Method	3	155
Observatory	3	274
Order	9	359
See-Access	2	214
web	5	542
Curation	9	1213
Resource	9	44
Saving	9	86
Curiosity	5	10
Data	9	1073
Innovation	9	1435
Improve	3	187
Power	9	303
Response	9	58
Status	7	25
Tranquility	9	51
Use-apply	7	533
Work	4	278

Figure 8-7 Collection of emergent role indicators

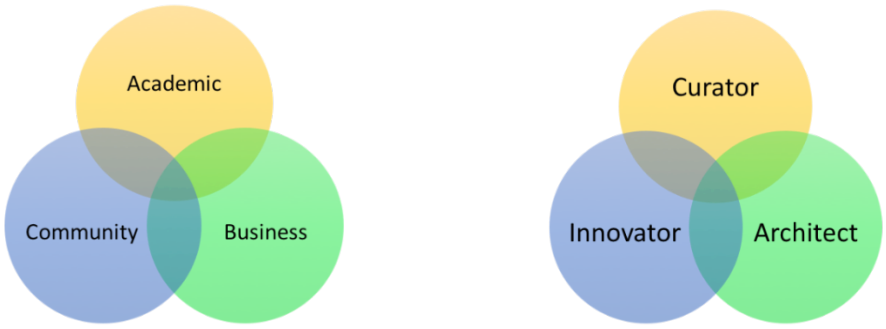


Figure 8-8 WO tribal roles and syndicate roles

The interviews were set against the (ABC) perspectives shown above (left) representing an expected occupational context/framing for the interviews. Against this, three perspectives of shared interest (syndicates – shown right) were identified in all three groups that centred around an interest/focus on data, an interest/focus on structure/process and an interest/focus on outcomes and these syndicates have been named:

1. **The Curator** - whose focus is on the collection/preservation/sharing of data
2. **The Architect** - whose focus is on methods/apparatus for enabling the sharing
3. **The Innovator** - whose focus is on extracting some capability/advantage

The overlapping figure suggests these are not intended as mutually exclusive interests.

While ABC distinctions are far from irrelevant and may, in fact, be dominant as they frame the social context from which each participant is speaking, the additional narrative roles which emerged from the interviews are superimposed (as *Laminations*) and may also be significant in determining the level of participation or resistance to cooperation between tribal observatories.

These are perhaps also viewed as nested/hierarchical frames or contexts and thus a participant might be framed variously as:

- <French> with the influences that French law and culture impose [Nationality]
- <a Business person> whose company has objectives, style and culture [Occupational]
- <An Innovator> who is interested in outcome over method [Focus/Interest/Speciality]

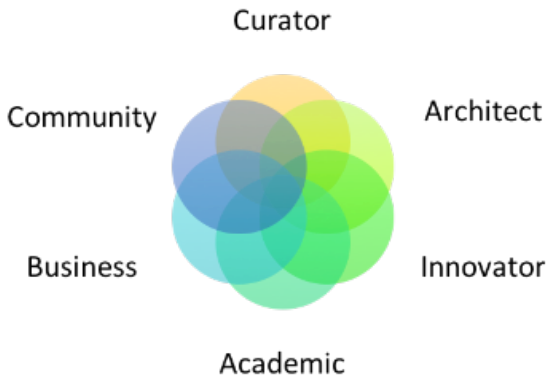


Figure 8-9 WO Syndicated roles intersecting with tribal roles

In order to address how the tribes and syndicates report their experience we consider both the structural (content) profile as well as the semantic coding (meaning). In terms of the language used, (Figure 8-10) interviews are objectively most similar within Tribes (an uncontroversial finding given the project-based theme throughout the interviews) and yet there are also correlations *across* Tribes (Figure 8-11) which the coding analysis suggests relate to the syndicate topics.

	Imelda	Ted	Ivan	Gail	Ian	Davina	Charlie	Thomas	Quinn
Imelda		0.2174	0.2174	0.2013	0.1996	0.1948	0.2059	0.1793	0.2041
Ted	0.2195		0.2469	0.2026	0.2212	0.2225	0.2317	0.1600	0.2151
Ivan	0.2174	0.2469		0.2138	0.2024	0.2418	0.2479	0.1630	0.2258
Gail	0.2013	0.2026	0.2138		0.2259	0.2181	0.1984	0.1577	0.1991
Ian	0.1996	0.2024	0.2212	0.2259		0.2279	0.2164	0.1662	0.1936
Davina	0.1948	0.2225	0.2225	0.2181	0.2279		0.2288	0.1631	0.2115
Charlie	0.2059	0.2317	0.2479	0.1984	0.2164	0.2288		0.1894	0.2670
Thomas	0.1793	0.1600	0.1630	0.1577	0.1662	0.1631	0.1894		0.2070
Quinn	0.2041	0.2151	0.2258	0.1991	0.1936	0.2115	0.2670	0.2070	

Figure 8-10 Structural (Jaccard) distance across all participants

Notable as an outlier is Thomas' interview which though averagely (at the mean level) coherent/similar to one of his tribe's other interviews shows significantly lower similarity with all other interviewees. Looking at coding for each participant in terms of their orientation to one/more of the syndicate topics we see percentage coding counts reveal the following:

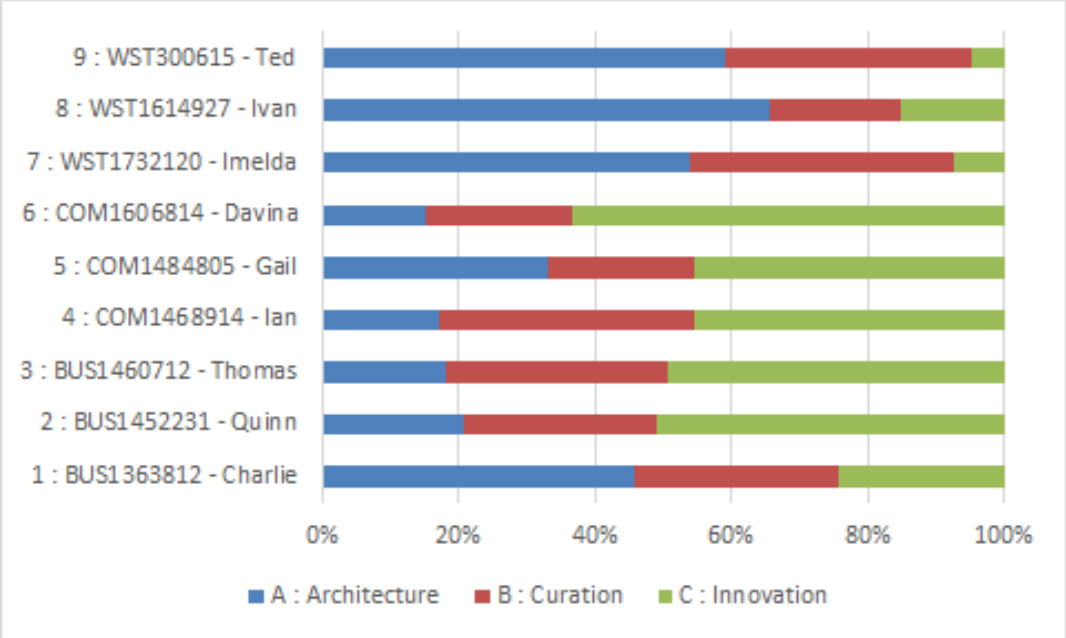


Figure 8-11 IPA participants syndicate orientation

In terms of groups we see a strong orientation to Architecture from the technical (Web Science) with little focus on Innovation (application) in comparison to much higher focus from both Business and Community on what the Architecture is used for. The area that seems more balanced and broadly shared is the focus on Curating/preserving data. An ordered pattern of relative interest can be determined for each participant (ACI, CAI, ICA etc).

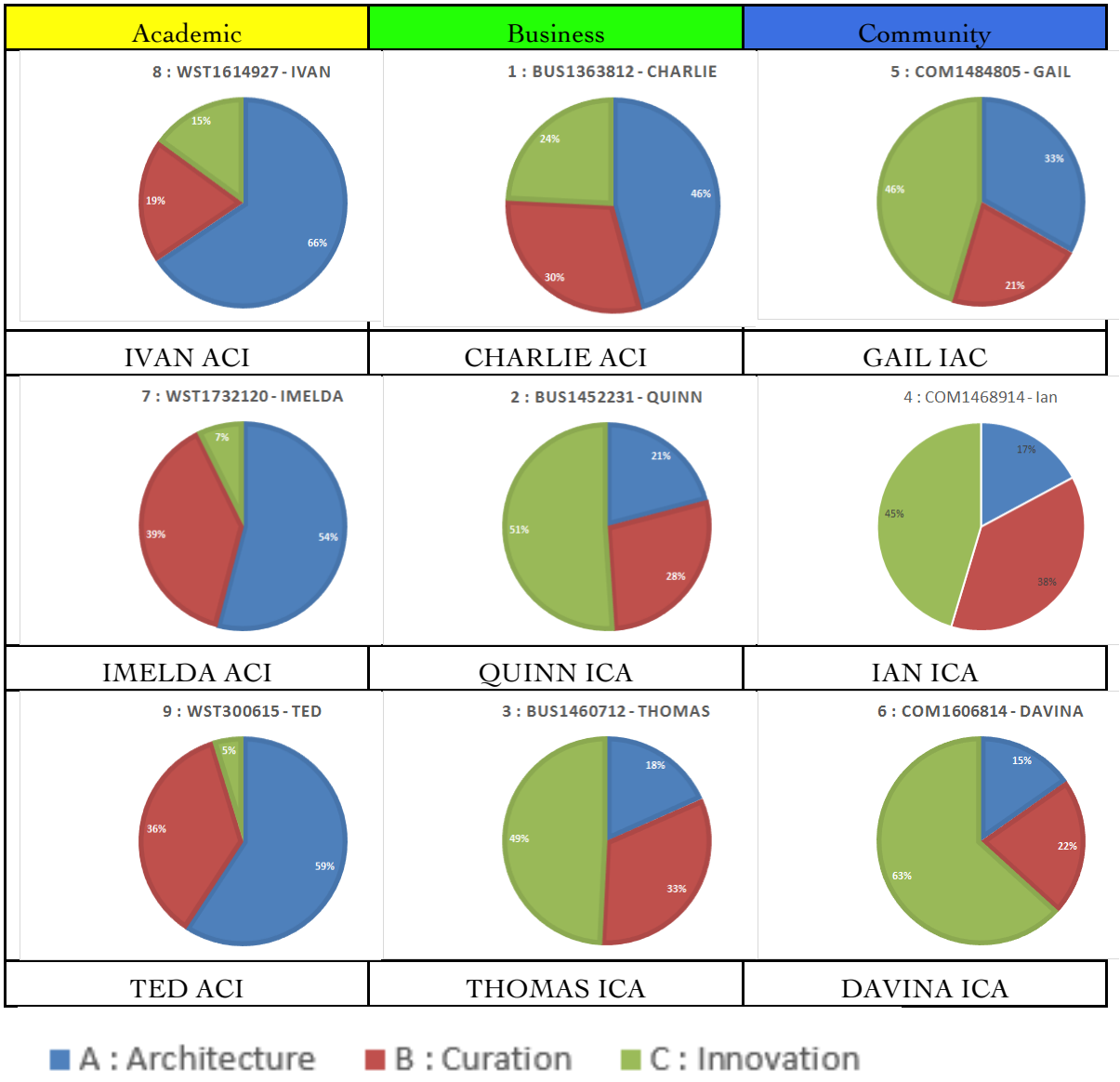


Figure 8-12 IPA theme priority by tribe/syndicate

This analysis shows general trends/distribution across tribes (Figure 8-13) and also the anomalies of individual users within tribe (Figure 8-14) that share similar focus at a syndicate level.

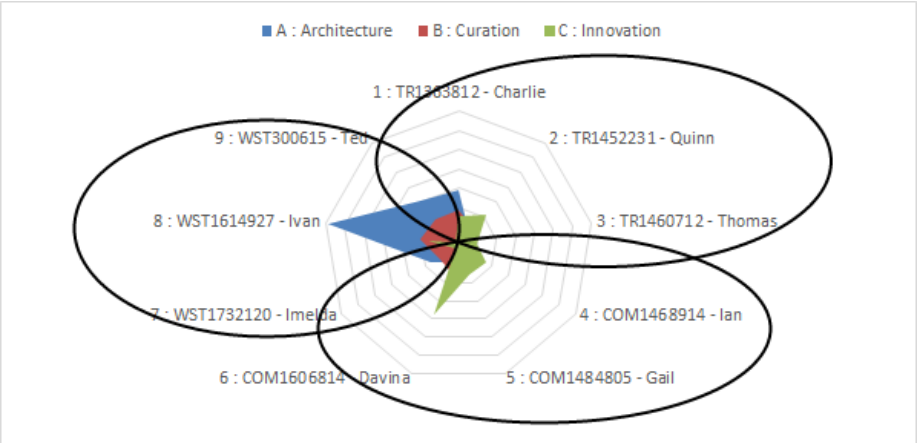


Figure 8-13 Distribution of focus across tribes

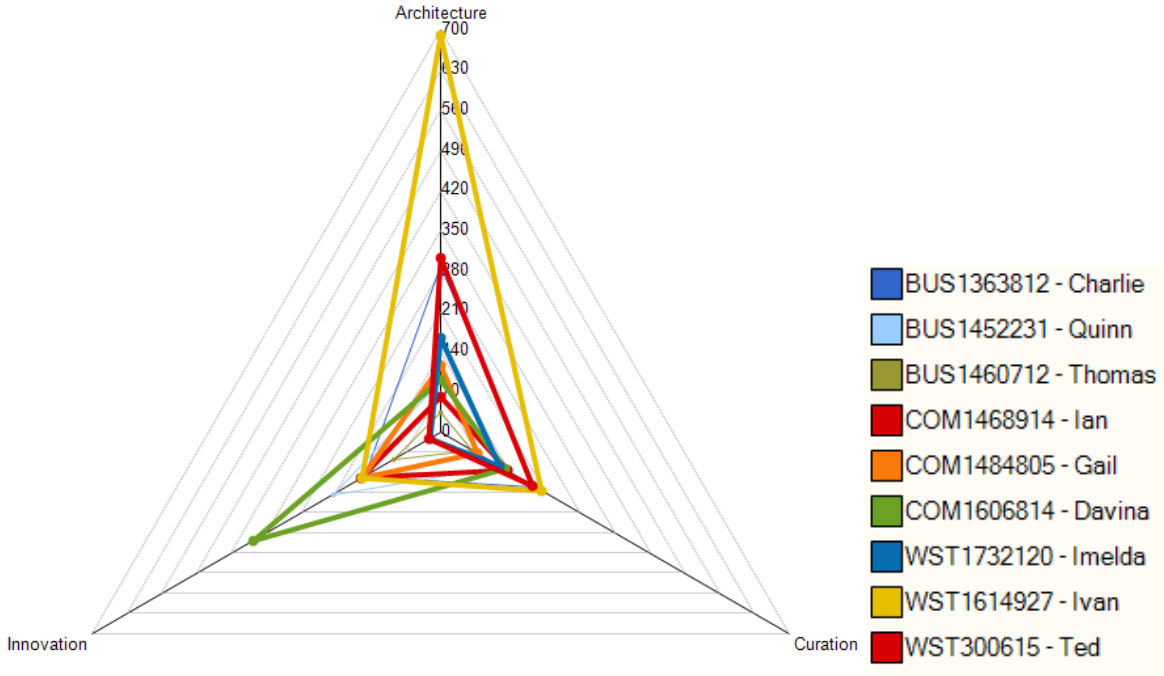


Figure 8-14 Focus profile by user

It should be noted that each tribe/syndicate is itself embedded in larger frames of National, International, Intellectual and Philosophical contexts beyond the scope of this project. This richer model of WO users (the A in DNA) may offer new possibilities to convene/develop WO SIGs (special interest groups) through which improved communication, rapport and reasons to collaborate may be developed. In any case it expands the set of discovered interfaces from tribes to similar/different syndicates within a tribe and similar/different syndicate across tribes whilst keeping a manageable number for the purposes of group management.

For the WO we may take away a number of key insights underpinned by the IPA interviews supporting RQ2/RQ3 in relation to diversity and interoperation:

1. The recognition that any dataset may only be part of a wider ecosystem that cannot be maintained by any one system. Sharing is, therefore, vital for coverage/diversity.
2. The vital nature of agreed (preferably open) technical standards to underpin the sharing of (meta)data and linking between sources particular where data volumes may be too large to move or duplicate directly. The WO becomes a point of reference/linkage rather than a centralised repository since such repositories are by nature typically restricted to choosing between selected types/domains of data. While WO may contain closed sources W^3O may not. Thus WO $\rightarrow W^3O$ emerges as a concept centred around shared standards/tools/sources.
3. The growing need for organisation/structure around data collections beyond simply acquiring/storing data lakes. WO may use big data techniques and/or connect to unstructured data sources, but the de-referencing of "raw" data into a solution for a contextualised problem requires structural/contextual knowledge which suggests semantic Web technologies to define/manage/navigate the connections.
4. Trust and provenance become key differentiators when choosing between sources/providers and solutions for Provenance (in the face of scalability factors) and models for quality borrowed (and scaled up) from enterprise data management become a focus.
5. The combination of advanced machine learning and crowd-powered techniques in the ingestion, classification and cleaning of data assets.
6. The lack of open data/tools vastly decreases the scope for network effects as evidenced by the private VOs offered by the digital catapult.

8.5 Conclusion

In the section above I have established profiles/narratives for participants individually, tribally (occupationally) and also at a syndicated (focus of shared interest) level across tribes.

A combination of qualitative assessment and basic quantitative analysis is used here to determine potential profiles or affinities as they are mixed/balanced that may help in communicating with each tribe and members within the tribe.

In the next Chapter the completed DNA model will be presented which represents a vocabulary of facets/features from which Observatories may be construed and constructed.

Chapter 9: The DNA of Web Observatories

In Short ..

In this section, the refined three-part DNA model is presented which has evolved from the pilot projects, case studies and detailed interviews. Example models using the DNA modelling notation are given.

9.1 Introduction

In this section, I present the three-part (D+N+A) theoretical model grounded in the research reported in Ch4-8 and a notation/representation for each perspective is given. Facets are gathered from a wide range of sources (Fig. 9-1) to deliver a superset or vocabulary of features, and so this represents a *notional* view of a WO and not a specific target or *normative* definition. i.e., WOs are *not* required to exhibit all the features from the vocabulary of genes

Example profiles from the SUWO (one from each perspective) are given using the template framework. It should be noted that SUWO is not presented as a specific target design for all WOs nor as an example of best practice but is a suitable example of real (rather than theoretical) WOs in practice.

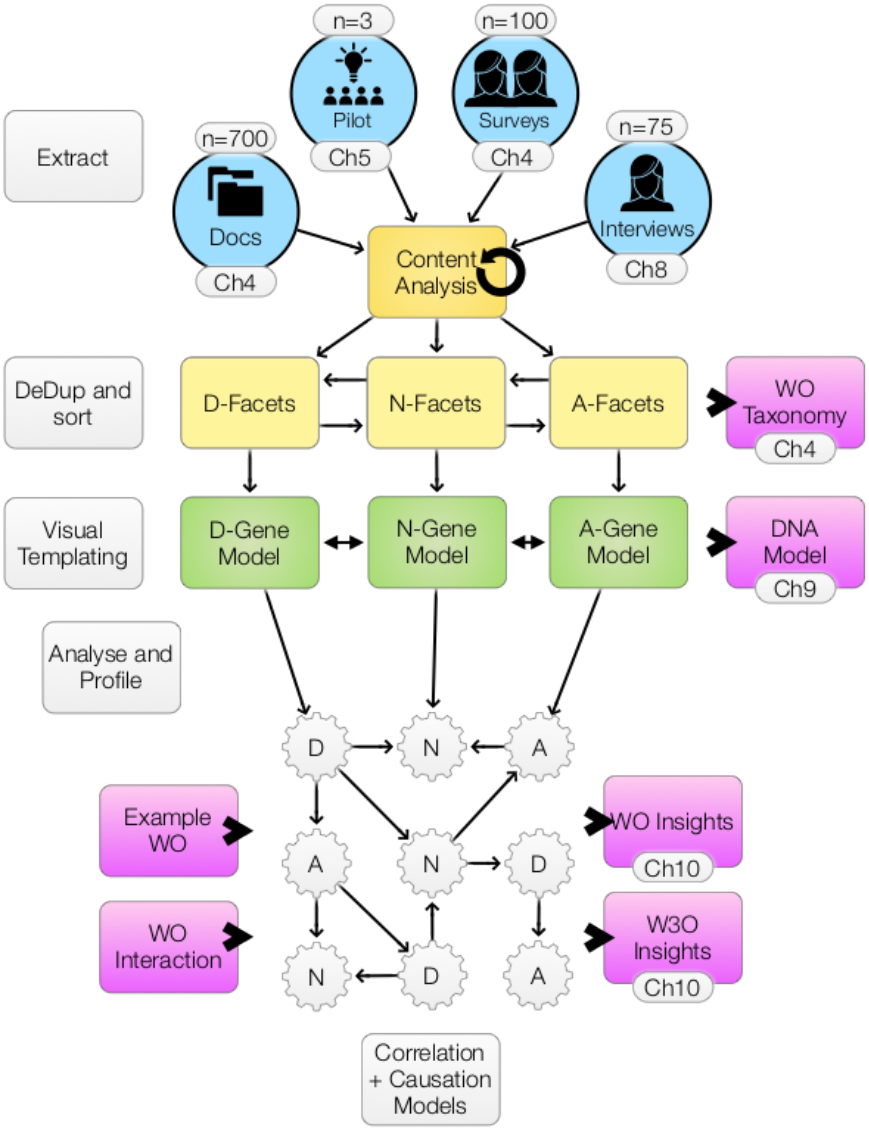


Figure 9-1 Overview of 'Extracting and sequencing' DNA

250 conceptual themes were identified from a review of the literature and observation of WO working groups and organised into three perspectives on WOs:

1. The **physicality/technicality** of WO (What is here? What does it do?)
2. The **meaning/ecology** (Who is engaging? Why?)
3. The **narrative via operational exchanges** which connect (1) and (2) (What is happening? What steps are playing out? What value is being exchanged?).

One might consider this to be a model connecting WO's technical and social elements via a socio-technical narrative.

Candidate facets were collected via content analysis of sources (organised by academic, business and community sources) and a series of smaller experiments including a pilot project (Ch5). The facets that were identified were grouped, de-duplicated and organised into a taxonomy (Ch4) which reflected the three perspectives:

1. A Definition of the form of the WO
2. The Narratives it executes and
3. The Agents (Agency) driving usage.

Each perspective was "mined" in a different way with the technical features/functionality being drawn primarily from published literature, the motivations and social grouping (which were less apparent from the literature) being drawn from interviews. The narratives which connect them were drawn (sometimes through observation/inference and sometimes explicitly) through a combination of the two sources.

The arrangement of features was suggested by earlier work on socio-technical systems and a biological metaphor where the complexity of systems is expressed through a combination of 'genes'. WOs therefore have the possibility to work in a variety of ways based on the set of all notional genes but each WO is not expected/required to express *all* possible features or implement all narratives (all behaviours) though, as with biological systems, features and behaviours might be expected to develop/adapt over time.

The set of all facets comprising this DNA of Web Observatories offers a rich contemporary 'vocabulary' grounded in current discourse/exemplars from which the set of WOs might be expected to be drawn. In biological terms this is termed a *Morphospace* and is strictly based only on facets found in current data sources/discourse and is not concerned with private speculation on the 'art of the possible' for future WOs.

Rather than a single normative definition for WO - the research findings uncovered a *set* of ideographic, contextual WO definitions that can be drawn from the DNA vocabulary of facets. This allows the identification of common elements, common transactions and common motivations whilst allowing for the diversity of WO perceptions that were apparent through the interview/analysis process. The WO shares a functional space whilst generating a much wider application space in the form of what Bowker & Star term *a boundary object* (infrastructure).

Such an approach supports all three research questions in terms of understanding the conceptualisations projected by different groups, comparing elements of WO to existing systems and placing the WO at the centre of a boundary structure comprising multiple social contexts and different frames of users. This affords different perspectives both on adoption and, in future, on how WOs may choose to interoperate within a wider network morphospace. The morphospace approach carries an additional advantage of representing the ways in which WOs are (in fact) populated vs. theory and can indicate which elements of the overall space may be (in) compatible.

9.2 DNA Definition, Notation & Method

As we have shown, (Ch4-8) there is considerable variability in the way in which users appear to conceptualise WO and how they might wish to apply WO systems across different contexts. The challenge within a definitional framework is to respect such diversity while retaining generalised structure to underpin comparisons. The DNA framework has three distinct perspectives:

1. **D** - **Definition** of the functional components (including human computation) of the system
2. **N** - The **Narrative** (negotiated exchanges) that take place when the system is used
3. **A** - The Agents/**Agency** of the system which are the human/technological components that instigate actions for a notional reason/motivation.

This model, in more prosaic terms, describes what a WO is, who drives the activity and why and how this activity is expressed. The facet groups are kept distinct (though connected) for the purposes of definition to reflect the possibility that the same system might be applied in different ways, or the same motivation might drive different processes.

In the next section, models of D, N and A are presented that are intended to represent the full range of reported elements encountered during the research project. Many more elements are thinkable, and indeed one might expect the definition of WO to grow over time, but it is not the objective of this project to *design* the WO but rather to capture how it is conceptualised by the community vs. how it is currently implemented across current exemplars.

Thus, when the full range of process elements are shown in the template form, it should not be inferred that all the listed processes must be present in all WOs (any more than one must use all colours in a palette for every painting). The claim is that a superset or vocabulary/palette of the processes can be represented in this way as a potential guide/road map for other WOs to compare/steer their own development.

9.3 Data Modelling Approach

The process of decomposing systems into three perspectives (technical, social and transactional) is summarised below before going on to a presentation of three Web Observatory visual templates and considering models from specific exemplars.

Three visual templates have been developed to represent the three perspectives from the taxonomy in the Triz notation as follows:

- System Concept Maps (after Novak) for physical systems (**D-genes**)
- e^5 list of narratives (developed in this project) for processes/exchanges (**N-genes**)
- Actor/Agency framing (adapted from Reiss/Goffman/Star) for social elements (**A-genes**)

These representations were created from the data in the following simplified sequence:

1. Extract facets from target systems/sources grouping as D's (function), N's (operation) and A's (intention) as a faceted taxonomy - preliminary analysis is done to establish facet frequency and clustering for typology. In the presence of more data, correlation amongst factors could be attempted informing the analysis of causation/behaviours. This is left as future work.
2. Establish a system scope/border and arrange functional (D) facets within the system, on the system boundary (interfaces) and outside the system as part of the ecosystem inferring exchanges between elements that are required for the system to function.
3. Arrange the set of (A)gents, both Human and Non-Human, around the (hierarchy of) frames within which they are engaging (Geo-political frames, community frames, occupational frames).

    - Natural behaviours for human agents exhibiting explicit or implicit motivations may be identified from existing work by (Reiss, 2001) and were validated against a demand model for open government data as presented in Ch4.
    - Non-human agents are recognised as valid agents and instead employ explicit algorithms or more general heuristics to fulfil goals or optimise values. NH Agents are considered as an algorithmic proxy for the programmer of the non-human agent for the purpose of this research.
    - Motivations (algorithms) are held to be the default driver for the use of the systems via certain narratives. Social (structural) factors (e.g., laws, taxes, customs, payments) may, however, promote/discourage different behaviours, and so a consideration of natural Agency vs. Structural influence may also be mapped.

4. Establish the set of (N)arratives or (N)egotiated exchanges that are represented by the interactions between actors, between systems and between actors and systems from the e^5 template "vocabulary" developed from the content analysis (Ch4).

    - These exchanges are grouped/sequenced and have a starting state determined/influenced by the relevant ecosystems and outputs that result from collective behaviours and emergent results that feedback into the ecosystem.
    - Facets may be annotated to highlight particular perspectives here. It should be noted that the same physical (IRL) exchange may be more than one notional exchange from the perspective of the Agents in the system since exchanges may be complementary in nature not homogenous:
      e.g., You buy a football because you want to give a gift: I sell the football because I want the money - we do not necessarily share any interest in Football as a sport.

5. The three models (or dimensions) are then analysed as a whole for interactions, perspectives, inconsistencies/tensions and, where possible, causal factors. The arrangement of the analysis may depend on the purpose:

    - Design of a new system
    - Analysis/monitoring of an existing one noting that any causative models arising from the analysis are NOT pre-supposed fixed by the steps above (see DNA AND NDA below).

From this broad 3-dimensional template, a large number of different systems may be described. To manage complexity, all elements should remain black-boxed initially and then unpacked and decomposed only as required.

9.4 Organising / Interpreting the models

The three sub-models do not explicitly require 'assembly' to form a single representation - but rather they potentially represent the same system from three different perspectives such that we may see an overlap between technical, process and motivational factors depicting different aspects of the *same activity*. i.e.,

1. [**D**] Database *retrieves* a record (technical element/feature)
2. [**N**] Record is *licensed* (sold) to a user (negotiated process)
3. [**A**] User (e.g., Aggregator) can clean/combine/resell the record for a *profit* (User/Motivation)

Following the identification of individual factors in each of the three perspectives, groupings and relationships between the D/N/A elements and the participant groups are considered.

At such early stage, this initial project has focussed on developing a single reference model encompassing *all* WOs generally but does not have sufficiently varied WO exemplars to underpin the identification of patterns/sub-types with meaningful statistical analysis. Future analysis to consider includes correlations between facets for *specific* WO's or specific sub-types through the construction of high dimensional models using techniques such as:

- Multidimensional scalogram analysis (in the manner of (Guttman & Greenbaum 1998)
- (Structural) morphometrics (Mitteroecker & Huttegger 2009)
- Generalised morphological analyses (Ritchie 2012).

It is, however, not an expectation of this project that the *template* models (i.e., the reference model containing the superset of *all* abstract facets) should necessarily exhibit meaningful correlations between WO facets/structures and hence this classificatory approach is left for future work targeted on specific WO's.

The identification of facet groups into functions, operations and intentions provides instead a catalogue or structural level of analysis. The next part of the analysis considered the distribution/incidence of facets between different groups. The distribution of genes was observed to be different for each type of gene and a representation/interpretation for each is provided below. This is an interesting operational observation around the categories that emerged, rather than an experimental finding regarding the facets/genes themselves but does inform the type of analysis that are meaningful for each category.

References to **D-genes** across the document sources appeared to be consistent across the (ABC) Tribe with the exception of interview transcripts which offered more references from technical experts than subject matter experts.

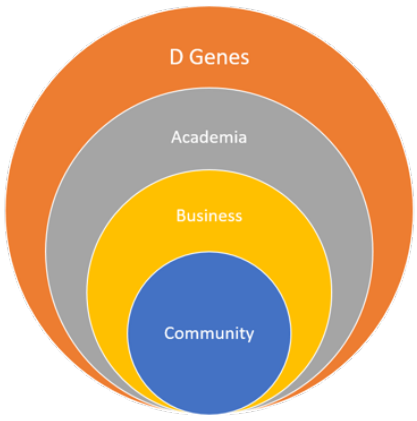


Figure 9-2 Broadly shared notional distribution of D Genes

All the tribes appeared to share technical conceptualisation of Observatories, and this suggests that the WO technical landscape may rely on *generic* approaches (open formats and general purpose hardware/networking). This is corroborated by the analysis in Ch4 which highlighted that the individual WO instance is less distinctive (more generic) and potentially less functional with respect to other web-based data analysis systems than the W^3O distributed paradigm.

N-Facets were observed to fall into two categories: broadly shared *generic* exchanges (like clarifications or canonical sources) and clusters of *occupationally-framed* exchanges (like citations or contracts). These occupation-specific narratives may be sources of friction for cross-tribal operations for groups that usually support them with the Tribal boundaries (e.g., Business systems supporting academic citations) and may be further complicated by the creation of synthetic data and services from multiple cross-tribal sources or providers.

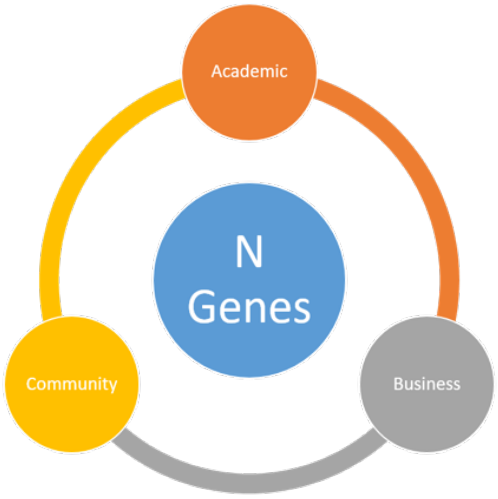


Figure 9-3 Localised + Shared notional distribution of N-Genes

In terms of **A-Facets** this type of information was rarely found explicitly in the document sources and depends more strongly on a double hermeneutic in which the researcher is overlaying his own interpretations/beliefs on the stated (apparent) beliefs of the observed/interviewed agents. During the IPA interview exercise narrative models were built and detailed textual analysis was performed.

Cross-tribal grouping of parties with shared interests (syndicates) were identified in the areas of curation, innovation and architecture. As Figure 9-4 indicates, other smaller syndicates may exist or indeed form over time but the three current clusters were identified as shown below.

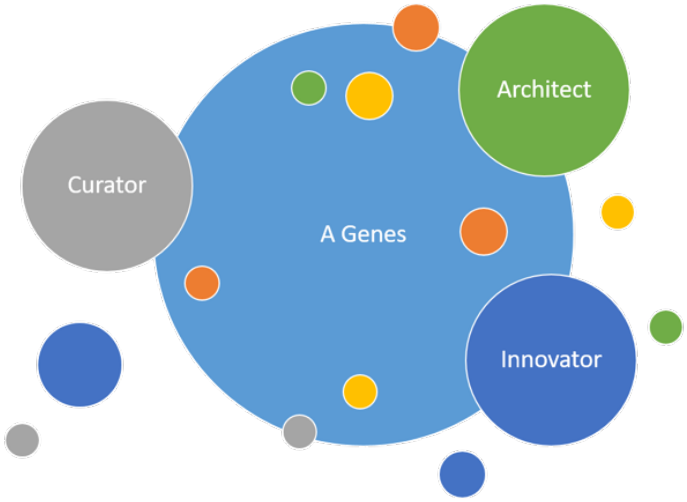


Figure 9-4 Cross-tribal syndicate clusters for A Genes

9.4.1 DNA-AND-NDA

In reviewing the approach and considering feedback from expert reviewers, it became apparent that the term 'DNA' might be read sequentially implying D→N→A i.e., that the process necessarily starts with technological features and is driven by them. This is, however, not the intention and so to reflect philosophically varied research approaches in this space it is of further importance that the often conflicting underlying views of technological determinism vs. social construction of technology be supported. This point is summarised by considering 'DNA AND NDA'. This sequence is intended to remind us that there are several ways to arrange D, N and A into recognisable forms that make sense for the researcher. This may include explanatory models driven by features/technology and/or driven by social elements as determined by the researcher herself. The model uncovered by the researcher may notionally include complex repeating series of genes (e.g. DANADNDA) though naturally only if the researcher's evidence supports the existence of such a pattern.

With this in mind, the broad spectrum of causality from technological determinism to socially-constructed systems can coexist (even within the same project) though meta-level diagrams which organise the facets according to alternative causal patterns of which many are possible. Three are given below as a starting point though others are possible.

The research team is free to consider the order or causality of perspectives including, but not limited to:

- **D→N→A** (technically deterministic) where the Defined technology drives the Nature/operation of the mechanism which shapes the agency and behavioural patterns of the Agents
- **A→N→D** (socially constructed) in which the Agents express their needs through a set of processes which shapes the technology
- **N→D→A** (process regulated) in which legislation, standards and other operational factors drive the functionality which, in turn, shapes the available behaviours and pursuit of opportunities (or process loopholes).

The intention is that DNA templates should be usable *by* teams and *between* teams to form a common understanding of the elements of observatories, providing a balance between standard/structured approaches and flexibility to allow interdisciplinary (and not just multi-disciplinary) insights and widely varied interpretative approaches to an understanding of what may be complex socio-technical systems.

9.5 (D) Design Facets/template

Functional/technical facets were identified and organised into a taxonomy (Ch4) as part of a content analysis exercise. In light of the poor visualisation options offered by the taxonomic form, this structure was refined over several iterations into a visual template using the TRIZ notation based on constant comparison of structure/boundaries across more than 1200 sources and confirmed via interview/observation.

- The resulting arrangement of features is intended to be implementation agnostic (in that it does not require/specify a particular technology or vendor) but rather suggests that a capability exists
- The arrangement of features is not intended to represent a technical design but rather a visualisation of processing/data exchange in the form of a logical architecture.
- The resulting features shown in Appendix comprises a vocabulary of features referenced or implied by the participant's discussions and literature sample.

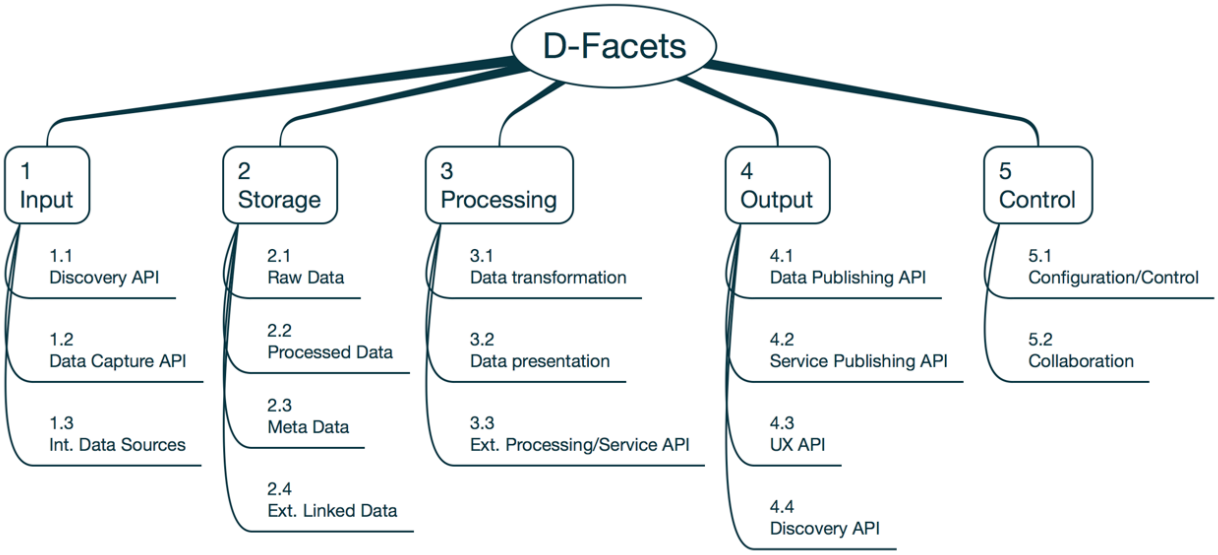


Figure 9-5 WO D-Facet Taxonomy (L1-L2 simplified)

These features are arranged in Figure 9-3 to imply an input/processing/output from top to bottom:

Figure 9-6 WO D-facet vocabulary

Below in (Figure 9-7) we see an example snapshot analysis for the SUWO system using a colour-coding extension to the notion, indicating (in this case) which features are implemented. Such a representation can be used to plan further extensions to the WO itself or as a comparison with other systems to which connections may be desired. As with other 'boxology' notions each box supports decomposition/black boxing and nodes may be decomposed for more detailed definitions of implementation details if required. The following snapshot was taken in mid-2016 and may require periodic updates to reflect the SUWO team's progress.

Figure 9-7 SUWO D-Facets: snapshot from Q22016

The figure above shows fully implemented features in solid green, partially implemented features in hollow green, features not (yet) implemented in hollow red and features that are specifically excluded (or not possible) would be shown in solid red (none meet the criteria here).

9.6 (N) Narrative Facets

The narrative process model identifies a large selection of narratives (definitions are given in the Appendix) which function to bridge the D and A elements by providing a concrete expression or transaction in which the Agents use the technological system to enact some more abstract motivation towards/away from something. These processes cannot logically be entirely independent of the technology, nor the Agents that underpin them and so are deliberately linked through a structure of five sub-processes shown conceptually in Fig 9-5. Two groups of factors (rather than exchanges) dovetail with the D and A groups as input/output links and three core groups describe the exchanges/processes themselves:

1. The Encounter section in which sources, users and services are discovered, discussed and disambiguated
2. The Enhancement section in which data are analysed, computed, enriched, visualised and stored
3. The Execution section in which data/services are mobilised for users and other systems and on-going updates and orchestrations are managed.

The two related groups reflect the inputs (sociotechnical Ecosystem factors) applying to the machines context and also the outputs (the resulting Emergent factors) that may result from the operation feeding back into the ecosystem in which the WO operates.

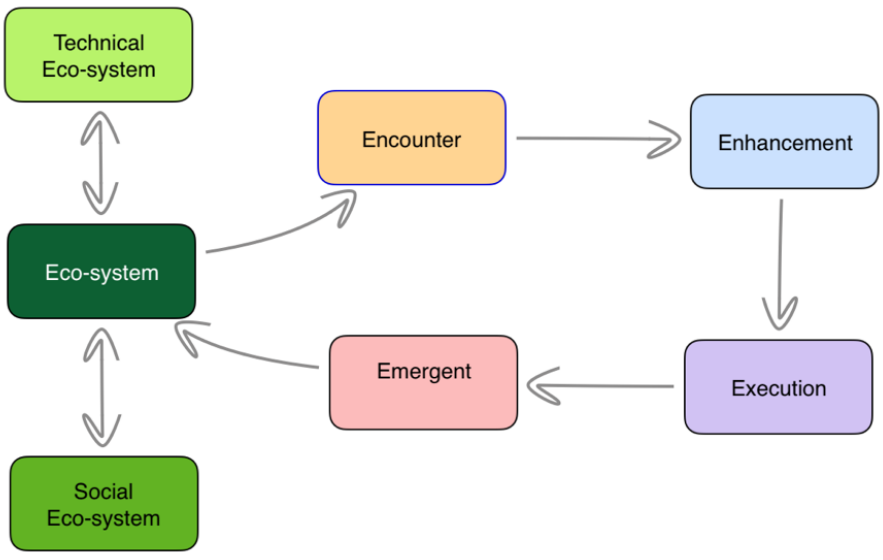


Figure 9-8 e^5 narrative flow model

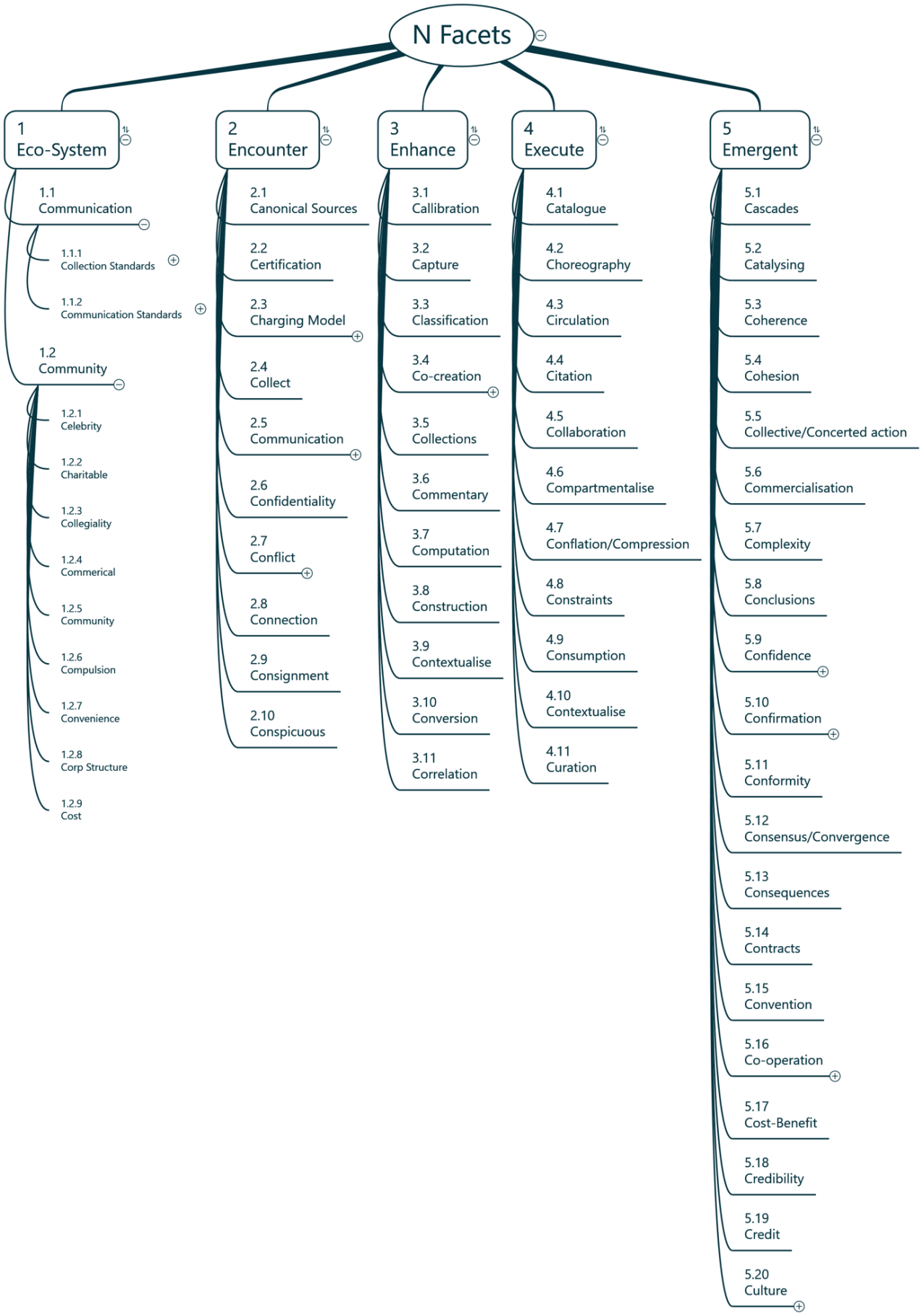


Figure 9-9 WO N Facets (L1-L2 Simplified)

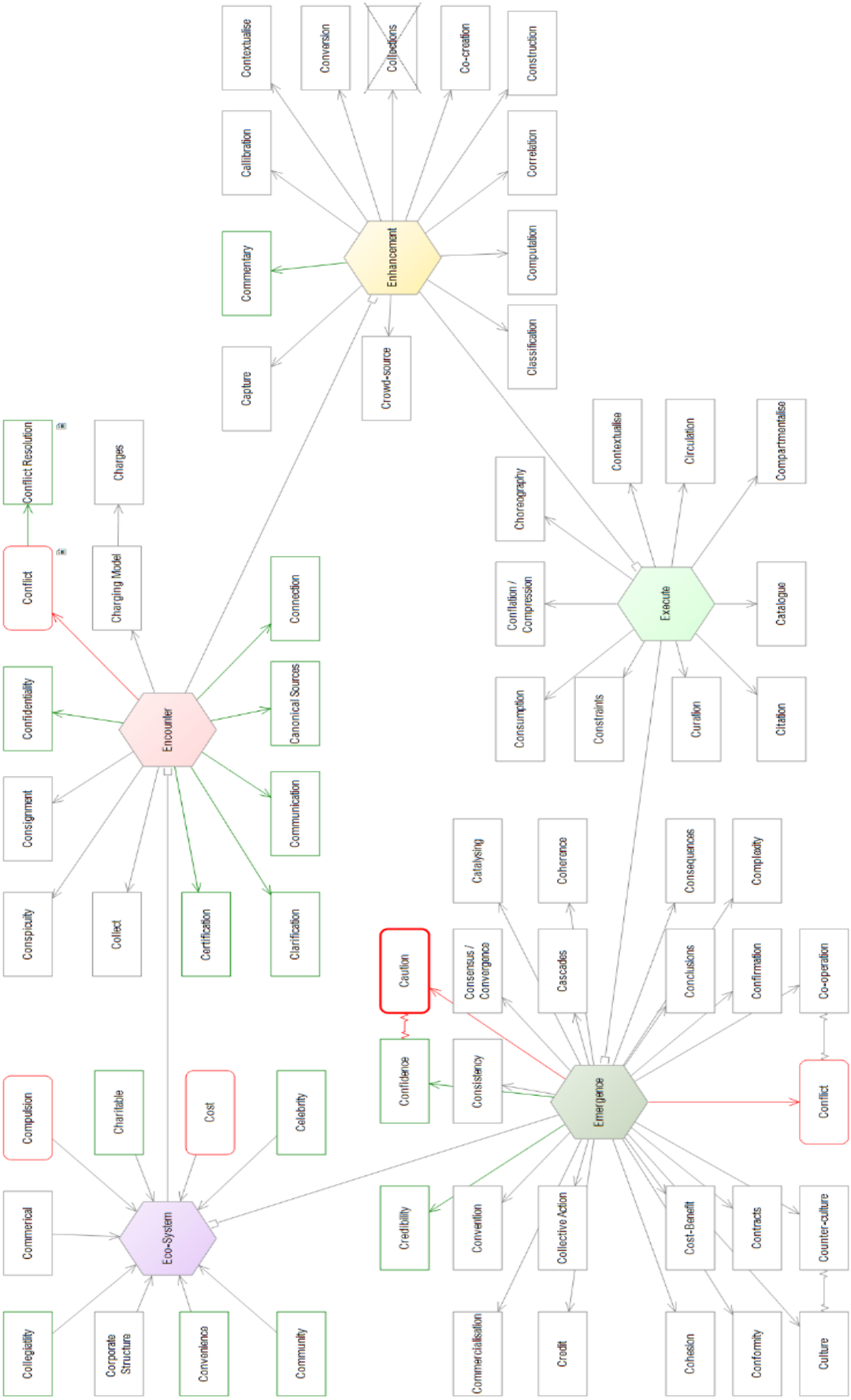


Figure 9-10 Narrative exchanges and processes

A note on the use of E's & C's

The reader will note that the processes/exchanges are described with terms beginning with a common letter (in this case C or E). It should be stressed that this is not intended as a flippant exercise but rather as a deliberate method of rendering the analysis of factors to be a higher (more abstract) level rather than simply quoting a collection of verbs/terms from source texts/transcripts. While retaining the original terms is technically closer to (more grounded in) the intention of the individual speaker/writer, there is an additional analytical challenge in aggregating such lists. The same term may, for example, be used in a different context or sense and at the point of presenting a grounded theory it is the higher level, more abstract concepts that are of most interest.

What has been attempted here is to apply the grounded theory technique of assembling terms at a more abstract (more theoretical) level by aligning them to *sensitising concepts* - (Charmaz, 2014).

In the following Figure 9-11the individual narrative exchanges are listed under each grouping and a definition/contextualisation for WO→W^{3}O is given.

SUWO has focussed on underpinning the development and deployment of minimum viable product (MVP) WO templates to kick-start the WO sharing process. This has enabled several other international institutions to set up WOs quickly and start to form a distributed network of nodes. The Narratives for SUWO are to some extent reflective of this MVP approach since in this initial case the WO *itself* is the objective rather than what is done with it, and we should view SUWO in this context.

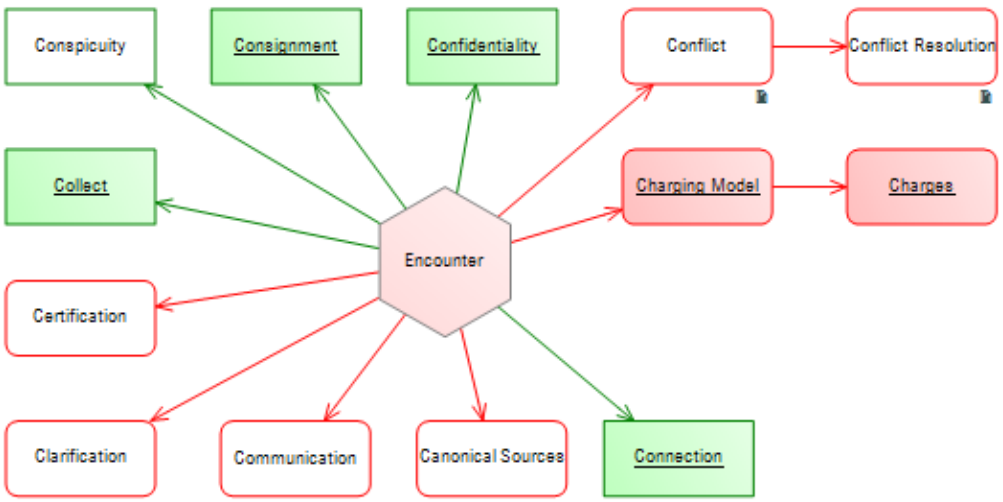


Figure 9-11 SUWO encounter profile: snapshot 2Q2016

It is notable here that SUWO includes the restriction of access to materials at different levels of sensitivity for different communities but has no charging model reflected for the use of the SUWO itself based on its own non-profit status. Freemium and support-based funding for SUWO have not specifically been excluded and third-party data/service providers may levy charges outside of SUWO.

Notable exclusions are around communication/collaboration and annotation of resources/apps and the understandable avoidance of certification vs. liability that would not be expected from an FOC system.

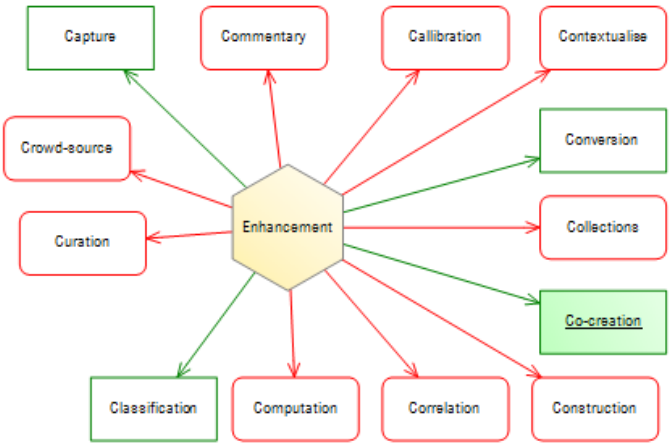
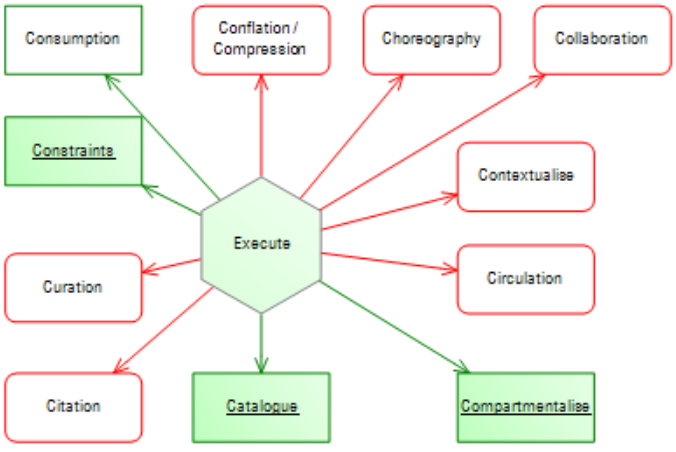


Figure 9-12 SUWO enhancement profile: snapshot 2Q2016

The MVP approach is also seen in the Enhancement phase where the creation of data resources vs. apps are hosted with the encouragement for users to "re-mix" them co-creating new versions of apps and updated/processed data resources. There are currently few built-in analytics and it is not clear that SUWO intends to go the route of embedding processing functionality in the platform rather than linking both external sources and processors.



.

Figure 9-13 SUWO execution profile: snapshot 2Q2016

The key features of the execution phase are to present a security-enabled model in which users can choose to publish material openly or require log-ins/credentials. While access may be restricted, SUWO updates local/remote global catalogues through the schema.org micro-markup standard.

9.7 (A) Agents & Agency Facets

The concept of Agents/Agency (who acts and why) is inherently broad, and an exhaustive treatment of psychological/sociological factors is beyond the scope of this project. The intention here is to focus on perceptions/actions around WO *as they are reported* through the body of interviews and literature rather than attempting a broader validation of social theories. Despite this, some of the concepts/factors included here (unsurprisingly) do mirror established theories/models of socio-technical systems. No attempt is made here to force the results to any existing preferred theory but rather the findings are grounded in a set of participants' experiences (purposefully) chosen for this project and may, therefore, be incomplete as result of sample bias.

The DNA Agent/Agency perspective supports the social element of the sociotechnical system and reflects several key elements which have emerged from the research:

1. The idea that Agents can be human or artificial and operate singly and/or collectively
2. The idea that social rules/constraints (Structure) will affect how Agents behave balanced against purely natural desires/programming (Agency)
3. The idea that collective action and structural effects operate in parallel within and across Agents giving 'net' behaviours
4. That both Agency and Structure can be 'framed' according to social groupings
5. That behaviours are consistent with, and are the product of, cognitive schemas (conceptualisations):
    - We run from something because we conceive of it as dangerous
    - We may buy something because marketers shape our conceptions of what it means to own/use this product

    There is hence a cognitive element to human factors not only in terms of meaning but also in terms of resulting behaviour (driven by the perceived meaning).

As the elements of function and exchange are stripped away from the interview narrative, there remains a static picture of who the types of player/agent are in the overall process. As such, this section tries to capture what was said/implied about the motivations/conceptualisations of WO specifically and how this fits into the agent-oriented view of the WO→W^{3}O ecosystem.

I consider two perspectives on Agency:

- The first is a conceptual macro-level of WO Agency and maps the collection of human/non-human agents acting individually or in groups and "naturally" vs. under the influence of regulation and proposes a social ecosystem within which agents will act
- The second is an expansion of part of offering a more specific micro-level representation of a "frame hierarchy" for individual objectives (adapted from Reiss) within occupational and other contextual frames (adapted from Goffman).

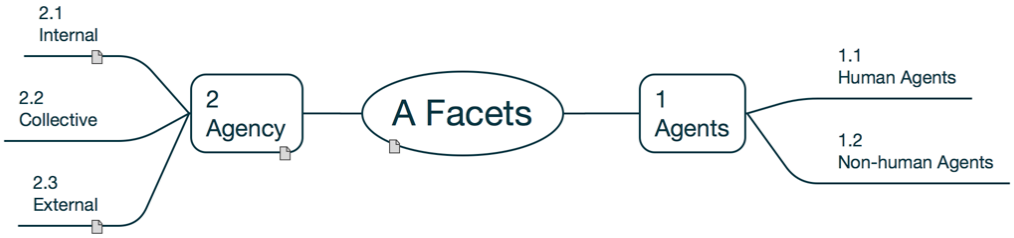


Figure 9-14 WO A Facets (L1-L2 simplified)

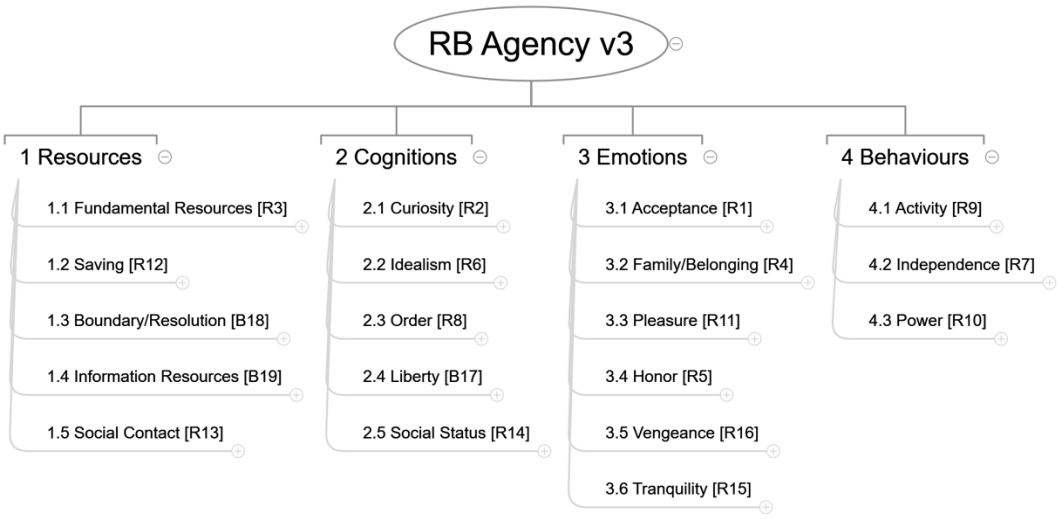


Figure 9-15 Adapted Reiss Model of Agency/Motivation

Along the first axis we may consider not only living users/owners/operators of systems but also algorithmic (active) systems, such as 'bots, and static technical artefacts such as data. On the second axis the actions/effects that these agents produce may be direct/individual, or mediated: either as net actions as part of collective/group effects or influenced by external structure/force/regulation.

These combine to give a “wider-lens” (Adner 2012) of the micro/macro effects on motivation and agency.

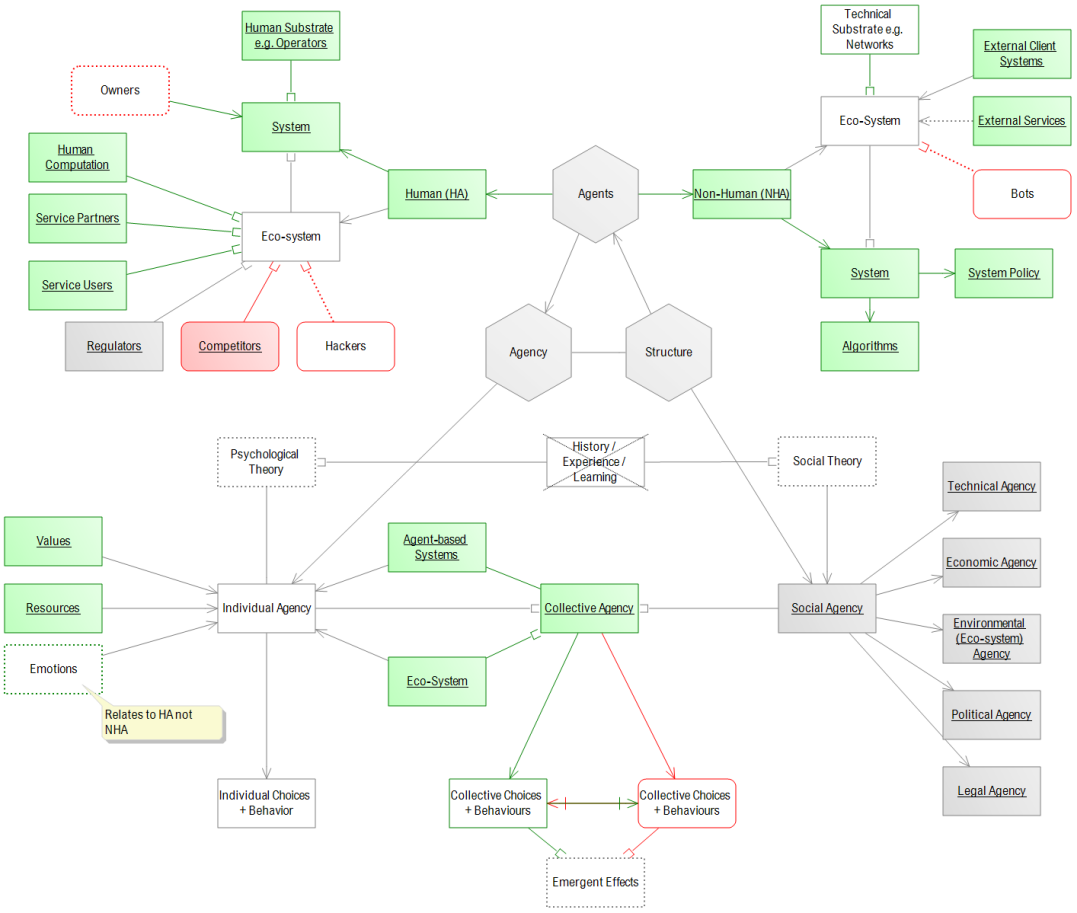


Figure 9-16 Macro level Actor eco-system perspective

The model in Figure 9-16 abstracts (according to grounded theory approach) from specific examples of agents/structures to general theoretical concepts. e.g.,

1. "<Company X> does our <process> for us because it's such a great fit." →[Ext Service Partner]
2. "We'd love to release <dataset> but we could have some real privacy issues."

I am not attempting a definition of all possible interacting structural/agency effects or restating a model of society and the social world in general. Rather the concepts and arrangement of the model are grounded in the participant interviews and contextualised for WO. It should therefore be stressed that I am only considering the WO factors/impacts that were discovered/observed during interviews albeit guided by existing models i.e., the effect of investment in WO depending in part on perceptions of economic opportunity vs. political sentiment about personal data privacy.

The overall model combines several interdependent aspects (Goffman calls these laminations) and represents a series of frames/contexts ranging from private (backstage) to occupational/public (front stage):

1. How a person conceptualises/frames the WO (e.g., as a tool, via a role or community) through associated values/schema modified by prior experience and current emotion (or how a programmer conceptualises the WO when coding a Non-Human agent with similar modifiers)
2. What a person seeks/avoids naturally (what a machine is programmed to do)
3. How the actions of others affect the default conceptualisation (if the NH agent is sensitive to its environment and its programming)
4. Whether one/other competing ecosystem elements/behaviours overrides another or if blended (net) elements/behaviours emerge
5. To what extent macro-level structural ecosystem guidelines/rules affect natural (net) behaviours to result in contrarian behaviour, adaptive behaviour or compliant behaviour.

At the micro-level I expand the individual agency nodes, adapting from (Reiss 2004), revealing a vocabulary of human motivations tested in Ch4 and also seen during the IPA interviews (Ch6-8).

Whilst the Reiss model attempts a universal characterisation of motivation, the detailed IPA interview process identified clusters of motivations both within (representative of) groups but also across groups.

We expect organisational and structural themes e.g., PESTLE (adapted from Aguilar 1967), to be represented here - these are what Goffman calls key (contextual) frames from which we make sense of reality and communicate socially. These might include making profit for a business, seeking knowledge in academia and serving constituents in a community.

From the literature search and pilot project, three ex-ante groupings, namely Academia, Business and Community were identified and this organisational framing is evident from the interview grouping. As stressed in Ch1, it is not the goal of this project to demonstrate that organisational groups generally act consistently with their own structure or aims (!), but rather to identify *extrinsic* factors that might offer a useful perspective when encouraging adoption and interoperation between WOs.

Three such groupings of motivations were found across all three groups: forming common interest groups which I have termed *Syndicates* [38] focussed on:

1. **Outcome** priority - where the content and methods are secondary to the goal/outcome in the participant's account (an *Innovators* syndicate)
2. **Data**/topic priority - where the methods and outcome are secondary to the content in the participant's account (a *Curators* syndicate)
3. **Structure**/process priority - where the content and outcomes are secondary to the structure/process in the participant's account (an *Architects* syndicate)

Thus we arrive at two axes/groupings. These are shown in (Figure 9-17) comprising common occupational (tribal) groups (ABC) and the common interest (syndicated) groups which emerged from the IPA interviews. I submit that whilst these groups are not self-evident they are plausible/unsurprising as to some extent they reflect the WO itself (Data, Technology and Application).

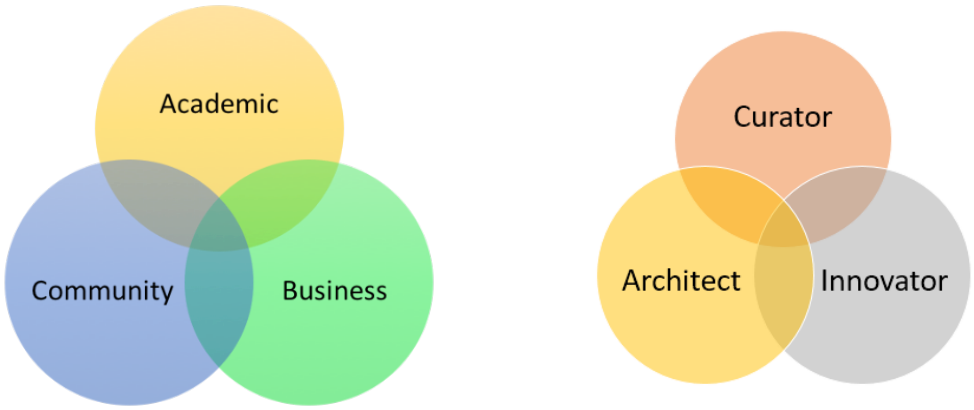


Figure 9-17 Primary axes for WO: Occupation vs Focus (Tribes vs. Syndicates)

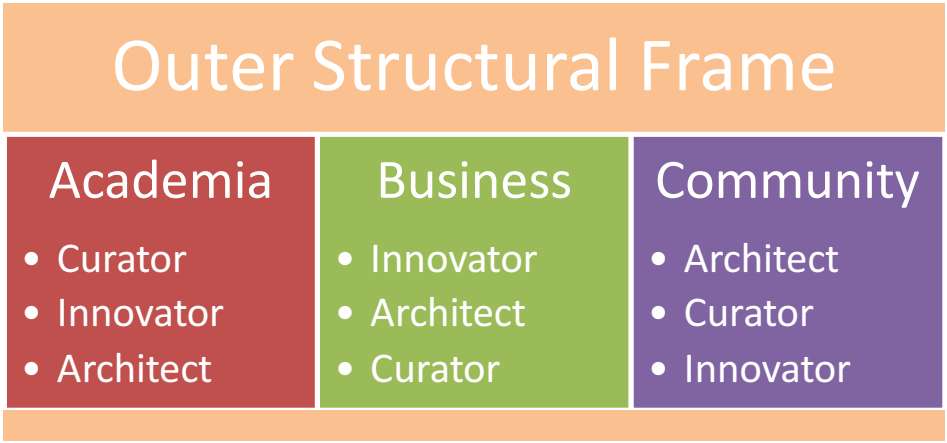


Figure 9-18 Hierarchy of frames: Structural(IRL) →Tribal→Syndicate→Individual

[38] a group of individuals or organizations combined to promote a common interest.

It should be noted the focus roles as well as the tribal roles may be considered as 'relative to one another' rather than mutually exclusive. Participants may express interest in both data and structure, and/or may act atypically for their tribe (e.g., employees of academic institutions who commercialise research outcomes or business-funded pure researchers). A *primary* rather than an *exclusive* focus is intended here.

- The **Curator** is characterised as the agent focussed on the content and resources within the WO and values the nature of the material itself for current and future uses. This is prioritised over the form of the current technical solution/tool employed and the particular objective or innovation that currently has focus.
    - It is represented by a set/sequence of motivational facets from the adapted Reiss set including the collection of raw (data) material (R3, R12, B18, B19), creation of a stewardship role underpinning Discovery (R2, R6, R8) and the creation of models to inform future behaviour (R4, R5, R10).
    - These elements are seen as significant themes in the interviews for [Imelda], [Gail] and [Charlie].
- The **Architect** focuses on the technical arrangement/orchestration of data/processes/people to enable particular outputs from particular inputs. She works with data/resources at hand and works towards the innovation/outcome at hand but focuses on the creation/perfection of the solution itself
    - It is represented by a set/sequence of motivational facets from the wider Reiss set including structure (B19, R8) the attainment of efficiency (R3, R2, R6) and effectiveness (R10, R7, R9, R16)
    - These elements are seen in the interviews for Ivan, Quinn and Ian
- The **Innovator** works with available data/resource and tools/solutions but is focussed in the particular outcome according to the marketing saying "no-one wants a quarter-inch drill, what they want is a quarter inch hole in the wall”
    - It is represented by a set/sequence of motivational facets from the wider Reiss set including taking action/application of what is available (R3, B19, R9), creating some with impact (R13, B17, R14 R10)
    - These elements are seen in the interviews for Ted, Thomas and Davina.

The nine IPA interviews were conducted to elicit specific perspectives across the three social sectors (Tribes) from individuals who were known to have experience/insight of virtual observatories and were selected to provide exemplars of the three viewpoints which are syndicated across the tribes. These syndicate priorities were rechecked against the full set of interviews, which were reviewed for the dominant characterisation from the interview summaries. This gave a broad indication of the three "Syndicates" across all interviews though in less detail than the IPA interviews.

Syndicates

Curator	24	31.17%		
Architect	29	37.66%		
Innovator	24	31.17%		
	77			

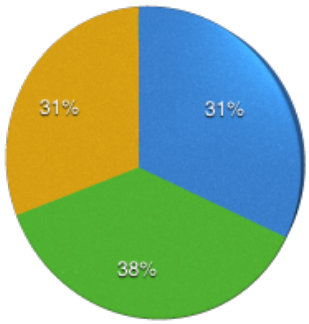


Figure 9-19 Breakdown of all participants by Syndicate

Syndicates by Tribe

	Academic	Business	Community	
Curator	38.71%	15.00%	34.62%	
Architect	35.48%	50.00%	30.77%	
Innovator	25.81%	35.00%	34.62%	
	100.00%	100.00%	100.00%	

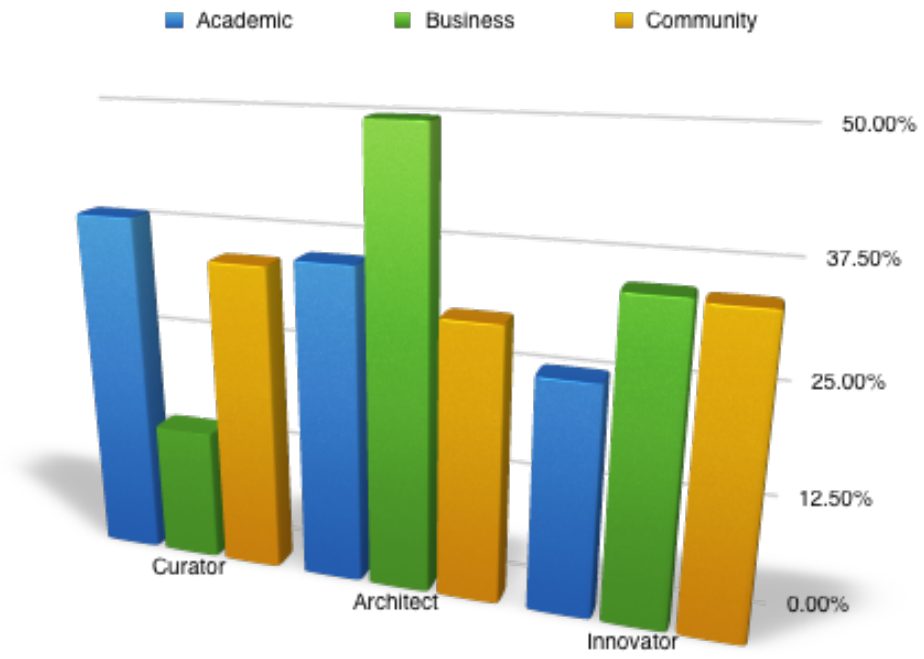


Figure 9-20 Syndicate members by Tribe

9.8 Conclusion

In this Chapter, the evolution of the DNA model has been outlined and the individual facets, the final structures and notations are presented through each of the three constituent axes/perspectives. Specific examples of notation in each case are given as snapshots from an example WO. The ability to flexibly arrange and interpret the models is presented through the idea of DNA-AND-NDA and the claim is presented that shared “syndicated” roles appear across all tribes that were studied.

In the next Chapter the implications and observations from the research and the DNA model are discussed. Limitations around the model are discussed in Ch11.

Chapter 10: Observations & Discussion

In Short ..

WO is considered from a series of theoretical perspectives/models – including those which emerged from the *ex-ante* literature review, from the research itself and those which subsequently bear comparison with the findings and the DNA model.

The first section offers an evaluation of the DNA theory according to grounded theory criteria and notes limitations of the research findings. The second section looks at various perceptions of WO which emerged during the research and which inform our understanding of the diverse responses that were observed.

I consider the implications of the research findings for WO adoption and wider establishment of *observation* as a new paradigm.

10.1 Evaluating the DNA model

DNA is a novel assembly adapted from a range of work covering structure/schema, agency and constructivist/framing perspectives. In terms of presentation/notation, it borrows ideas such as decomposition and black-boxing (information hiding) from system theory and thematically employs elements from:

- Ranganathan, Denton, Spiteri's work collecting facets into a taxonomy which offers useful grouping/structure and extended using concept maps.
- Abell's three-point perspective and Whitworth's model for STS perspectives/objectives comprising technical and motivational elements (social constraints).
- Malone's gene expression metaphor
- Narrative modelling using the TRIZ notation (Altshuller) rather than more traditional plot element or story grammar tools since the objective here is to identify/resolve tensions within the narrative between participants rather than simply document what the narrative is per se.
- Elements of Callon & Latour's Actor-Network Theory (ANT) allowing for both human and non-human agents (actors).

An ex-post literature review suggests parallels with matching social psychology and social theory models. These include aspects of structure/structuralism[39] (**D**), Functional[40] ideas (**N**) including social exchange and Symbolic[41] interactionist perspectives (**A**) such as Frames (Goffman, Bateson), and Boundary Objects (Star, Bowker).

DNA, therefore, offers a multi-perspective framework for studying observatories enabling team(s) members to focus on specific areas which contribute to a larger multi-dimensional analysis and to apply one or more theoretical interpretations to results based on the skills and perspectives of the individual researcher. This approach, while rooted in existing functional decomposition and system design approaches, is nonetheless a novel take on the specific set of challenges that result from interdisciplinary research of socio-technical systems.

Despite the conflicting perspectives on the *primacy* of social vs. technical effects and of agency vs. structure in socio-technical systems, a socially-embedded system (e.g., a "Social Machine") must be more than just the sum of its technical parts. Thus, while relevant, a purely technical assessment is insufficient to capture its essence, and by extension, a purely social view may be a no more accurate view than a purely technical one.

How then can we combine this array of different factors into a descriptive or even causative model? Considering the emergence of biological systems (organisms) from their genes leads us to consider whether socio-technical systems might result from "socio-technical genes" that correspond to the facets extracted from analysis/observation of the phenomenon. In this paradigm, genes may be present (or not), dominant (or not) and might be expressed (or not) relating to their relevance/impact in a socio-technical system. (Malone et al., 2003) previously considered genes within the context of collective intelligence, and so we extend this to broader (a higher level) analysis for socio-technical systems and Social Machines such as Observatories.

[39] What a thing is
[40] How is functions – its purpose
[41] How it is embedded in a social context/meaning

Having sought to define/characterise WOs through a set of documentary and interview sources I find three main perspectives from which to define WOs:

- **Technology** (Construction/Instruction) - the arrangement/agreement of technologies in standard/shared formats which **D**efine the WO instance.
    - This forms the minimum shared definition of the WO as it is instantiated providing a description of the underlying tool and the standards it uses
    - The specific technology/approach is subsumed by the required functionality/format
- **Narratology** (Expression/Execution) - the exchange of selected data/services via Narratives/Processes (in a series of steps/transactions over time)
    - This forms a bridge between the technical and social elements of the WO eco-system and comprises the sequences (both technical and cognitive) that represent the ways in which the WO is used. These may be simple data exchanges captured via business process orchestration but may be more complex exchanges of trust/assurances as with establishing a canonical source amongst many or determining quality, provenance or trust for a source.
    - The specific method/format (manual, automated, written, spoken) is subsumed by the values that are exchanged
- **Ecology** (Agency/Ambition) - enacting choices within "frames" to engage with socially-contextualised problems/goals
    - This forms an eco-system of homogenous/heterogeneous/complementary goals expressed by social actors driven by goal-seeking/problem-avoiding behaviours at the individual/group level and modified (subsumed)
    - Individual/default behaviours are subsumed by net collective behaviours and/or structural forces at the broader ecosystem level (e.g., the law may dominate the desire to express a personal desire)

DNA extends the Taxonomy discussed in Ch4 and evolves from the need to study WOs (and particularly W^3O) as a blend of technological and sociological perspectives beyond the static catalogue of facets, through arrangement/presentation and analysis of the interaction between facets reflecting the social, technical and socio-technical elements.

10.2 Evaluating the project and limitations of the research

It has been argued throughout this document that we approach and even construct our interpretations of reality from a social context/personal perspective, and hence an attempt has been made to access a wide range of sources spanning documents (both textual/visual), focus groups, participant interviews, observation and surveys. Not only has the format been diverse but a conscious effort made to reflect diverse perspectives and ensure multiple data sources within academia, business and government. In most instances, participants have been paired with at least one more from the same organisation, and each organisation has been paired with another from the same Tribe. Three distinct 'Tribal' views have been included spanning 30 organisations. The literature suggests a minimum of n=12 interviews to achieve saturation/validity for participant interviews whilst the totals for this project are n=31, n=20, n=26 respectively for Academic, Business and Community Tribes giving a total of n=77 for individual/group interviews and more than 100 participants overall.

Ultimately one must engage the community of Observatory builders to determine how accurate this initial taxonomy structure may be, but based on Spiteri's criteria (derived from Ranganathan and the CRG's criteria) for evaluating faceted classification I submit that the taxonomy for WO performs as follows:

1. **Differentiation** – Top level facets are fully differentiated
2. **Relevance** – partially. e.g., the focus on platform details may not be relevant to all users of the classification
3. **Ascertainability** – partially (platform objectives such as "scalability" are poorly defined in the literature)
4. **Permanence** – fully – whilst sources/topics may change we feel the top-level facets will be stable.
5. **Homogeneity** – partially. Topic Data and metadata may be homogenous (or converted to such) within a particular classification but all OSN sources will not be functionally equivalent
6. **Mutual Exclusivity** – partly. Interfaces may be thought to be a subset of Services, but we have chosen to pull this out separately for the purposes of understanding WO usage.
7. **Fundamental Categories** – fully. None of the facets function as a more general facet of the others

As for the DNA model itself, theories produced by Grounded approaches are evaluated not in terms of truth or validity but rather in terms of appropriateness of the final theory to the problem in terms of Fit, Relevance, Workability and Modifiability.

When we consider the grounded theory suggested by the DNA analysis of the example WOs I approach the four criteria for evaluation as follows:

Fit

The structure of DNA expresses the technical, performative and social elements of the WO experience reflecting key themes in both the literature and observed in projects and interviews. This represents a suitable reflection of the organisation of themes. It allows focus on the individual elements of interest or the model as a whole.

Relevance

WOs sits in a broader ecosystem of Web data and Social Machines that may overlap in terms of function and intention. Understanding friction/alignment between individual WOs and between other sources/services in the shape of WO-like systems (and sources more broadly) underpins an understanding of how to assess and potentially encourage/engineer participation in the global Web of Observatories, W^3O. The discovery of cross-Tribal roles (WO Curators, WO Architects and WO Innovators) similarly enables the outreach and incentive engineering around WO participation and the transition of WO→W^3O.

Workability

Sub-dividing into perspectives allows not only an appreciation of and focus on the three perspectives individually (potentially by different teams) but also the conceptual flexibility of arranging these variously into further causal/analytical models. It is practical/workable in that different viewpoints within and across teams can be allowed to diverge and explore while staying within the broader DNA framework.

Modifiability

DNA is based on a faceted analysis of exemplars and discourse. It can easily be extended to reflect new functions, processes or participants as required. The faceted taxonomy has been created to reflect types/classes of the element rather than their end values (i.e., 'Social Networks' not 'Twitter') and so the definition is not based on ephemeral sources or content but broad classes of facet. The Tribal model itself reflects generally-recognised broad social groups that have remained stable for centuries though these could be replaced with other perspectives (e.g., profit vs. non-

profit.) The syndicated (focus/interest) roles reflect current data but could also be modified or replaced to reflect other perspectives which cross-occupational/structural boundaries.

Beyond the raw definition of WO as a collection of features/functions, there is an element of contextual meaning and perspective to consider. Consider the example of accurately defining a simple object like a knife:

"as a tool comprising a blade typically set in a handle"

This relates to a (Chomsky 1957) "Deep structure" context-free (Blade + Handle) without referencing a context for the application of the tool giving us several potential "Surface structures":

- Spreading butter on toast/cutting vegetables → knife as cooking implement → nourishment
- Stabbing a victim/excising a tumour → knife as life/death implement → power/control
- Juggling/throwing knives to entertain children → knife as the focus of dexterity → skill/amusement

The surface meaning of the tool to those engaged in experiencing its use will vary widely and indeed throughout the research process various definitions of WO as a solution/application emerged from what (Goffman 1974), (Kahneman & Tversky 1984), and (Bateson 1972) called "Frames". Such Frames may be single artefacts or roles that contextualise meaning and dictate a reaction/behaviour. Examples are a single object like 'a prison door' or these cues may be stacked up as layered frames reflecting the multiple groups (e.g., UK/Academia/Curation or Australian/Government/Architecture) which participants use to encode/decode ideas around WO such as those we noted during the research.

"WO,.." we heard,:

1. ".. is all about government data"
2. ".. is all about digital literacy"
3. "..is all about Web Science"
4. ".. is all about sharing".

In the next section we will consider several additional frames for WO.

10.3 Considering The Many Faces of WO

Parallel to the various intrinsic elements of WO's DNA (Structure, Expression and Context) there are also eco-system (E) perspectives [42] that inform implications/recommendations for WO which emerge from a review of the narratives at the top level:

Conceptualisation		*Description*
WO-as-a-meme		What do people *mean* by WO? How DNA is positioned as a meme with the network of users in terms of the perception of functionality/quality and adoption/feedback.
WO-as-boundary-object		The extent to which a single WO functions as a central concept for diverse applications such that the definition or understanding "flexes" becoming less coherent/consistent further from the notional centre related groups (After Leigh Star)
W^3O-as-boundary-infrastructure		The extent to which the shared community of WOs functions as a broad mediator between diverse groups of users, providers (After Bowker & Leigh Star)
WO-as-a-novel-solution		The extent to which WO is seen as a solution distinct from existing offerings
WO-as-a-set of-genes		The extent to which the genome metaphor successfully meets the objectives of the research project
WO-as-a-project		Considering WO as a piece of WO to be delivered over time to meet some set of expectations and within an agreed budget
WO-as-a-paradigm		Considering whether the concept of Observing the Web is an evolution of web paradigm in the form browsing → searching → observing rather than a specific class of platform or technology

[42] DNA may be extended to EDNA

WO-as-a-social-machine		The extent to which WO/W^{3}O fulfils the criteria of being considered a Social Machine (After Berners-Lee, Shadbolt)
WO-as-a social-movement		The extent to which WO is an expression of needs, a deficiency of tools/resources for the pursuit of Web Science (After James & Van Seeters)
WO-as-an-innovation		The extent to which WO may be adopted, resisted and disrupt/be disrupted.
WO-as-knowledge-infrastructure[43]		The extent to which WO may be considered part of the socio-technical idea concept of knowledge infrastructure (After Edwards, Williams)

Table 10-1 WO conceptualisations

10.3.1 WO-as-a-Meme

The two worst things about Web Observatory ..

Saussure's work gave us the notions of *Langue* and *Parole* to highlight the difference between the rules and potential arrangements of *language* versus the chosen forms and meanings of *speech*/*communication.* I observed this difference as an important recurrent theme running through the project through the varied usage/understanding of the term "Web Observatory". This difference gives rise to a disconnection between the words used (Reference) and the physical object (Referent) that was "pointed to". (Ogden & Richards 1924) depicts a *semiotic triangle* (Fig 10.1) showing how individual conceptualisation separates/mediates this translation from symbol to object. Consider the ambiguity behind the term "ORANGE"

[43] Added at the suggestion of the Thesis reviewers

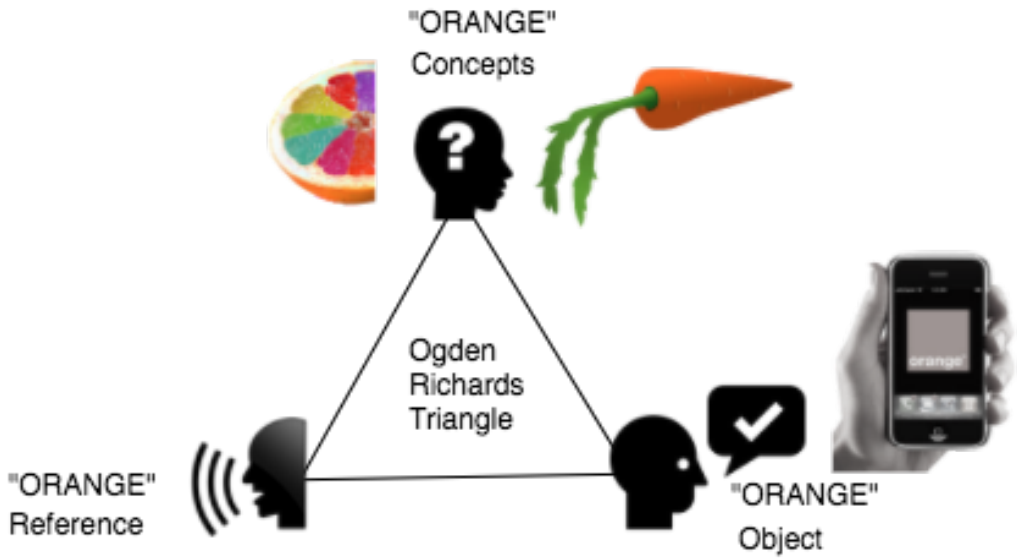


Figure 10-1 Semiotic Triangle (adapted from Ogden and Richards)

This example give us multiple ways to "unpack" the reference that was used:

- A specific citrus fruit (whatever colour it may be),
- The colour of any matching object
- A mobile telephone company.

The label itself holds many meanings and must be contextualised in order to connect the term to the object. A similar process appears to be in play with the term Web Observatory.

Chomsky's early work on transformational grammars also highlights two linguistic levels: *Deep Structure* containing the elemental parts of the concept.

e.g., 'Observatory' and 'Web'

We also process *Surface Structures* in which in the concept of 'Web' is generally unpacked as either [44] a collection of data-**on**-the-web (content) or data-**about**-the-web (metadata). Whilst alternative surface structures are often thought to be subtly different:

e.g., 'John hits the ball' vs 'the ball is hit by John'

in the WO case the surface structure ambiguity is perfect. No additional surface information (nuance) is available to extend the deep structure of 'Web' and 'Observatory' and so alternative interpretations of 'Web Observatory' have identical surface structure. The term in English is thus

[44] One survey participant suggested that Web Observatories related to the study of spiders

structurally ambiguous (perhaps *artfully* so) and through this device casts a wider net for potential participants than either alternative alone.

This becomes an important question for the characterisation of the Web Observatory and is related to the characterisation of Web Science itself as a recently emerging discipline positioned within a cluster of related disciplines such as network science, internet science and computational social science. There may even be *disbenefit* to disambiguate the terms since to exert pressure to abandon either data-ON-the-Web or data-ABOUT-the-Web completely would potentially be to abandon a significant portion of the scope/applicability of WO and this undoubtedly influences those groups who might participate in a larger W^3O.

Where this broader scope of definition is welcomed, I observed the duality to be seen as 'flexibility' while, where less welcome, it appeared as 'uncertainty'. A blind poll (Ch4) asked participants to define/evaluate several alternative surface structure representations of the deep structure ('Web' + 'Looking') and a broad range of responses were received which appeared to be explained, I concluded, by the level of occupational contextualisation for such an idea. We believe then that the ideographic 'framing' (Bateson 1972, Goffman 1974) of concepts is an important counterpart to any objective reality.

Equally the use of the term "Observatory" sparked comment about the nature of WO as a 'passive repository' vs. an 'active probe' for the Web. Unlike astronomical observatories whose activities are presumed to have no effect on the galaxies they observe, WOs do potentially feedback into their own ecosystem either *explicitly* through the publication of activities/results or *implicitly* through the application of results (changes in policy/strategy/focus). What are the ethics of WOs introducing probes/changes onto the Web ecosystem in order to measure the reaction/effect? Our survey participants variously responded with enthusiastic comments about "knowledge" and "research" finding the idea "cool", whilst others feared "surveillance" or "spying" calling the idea "creepy". Thus issues around ethics, privacy/transparency and impact emerge as factors for which the analogy of an "Observatory" may be seen to be less robust.

In addition to noting that different contextualisations (Surface structures) exist I also coined the term "*airbrushing*" to describe the way in which participants smoothly, repeatedly and (apparently) unconsciously switched between and conflated these definitions of WO such as data-**on**-the-web vs. data-**about**-the-Web) and WO vs. W^3O. Participants broadly failed to make distinctions clear during interview often mixing the physical/notional and moving back-and-forth (airbrushing) between different surface structures when recounting their ideas/experiences of WO. This may suggest that participants are not prepared to exclude either content or metadata from the definition of WO. WO, to some extent, becomes a 'transparent' concept (Star &

Ruhleder 1994) here, or perhaps an interim (building block) concept with the ultimate focus resting on a collection or ecology of WOs (W^3O) as a shared networked system rather than on the standalone WO.

Thus, as Wendy Hall has quipped, the two worst things about Web Observatories are the word "Web" and the word "Observatory".

10.3.2 WO-as-a-boundary-object

In addition to factual definitions of WO, varied characterisations of WO as an application emerged from the research that are not necessarily implied by the nascent/exploratory state of the project.

e.g. “WO? It’s all about digital literacy!”

though little work has yet been produced associating these two ideas.

The WO is still at an early stage and exists in a partially fluid state for all those engaged in exploring the designs/concepts and in a space of agile building and experimentation. The developers/builders, for example, are those for whom WO may be framed as a "tool/instrument". In contrast, for end users I repeatedly observed varied conceptualisations in those encountering WO for the first time and who framed it as a solution/opportunity for a problem in their own eco-system.

It is, of course, perfectly acceptable that certain definitions can co-exist without tension or contradiction but the large range of applications suggested to me that WO was being conceptualised in a corollary of Maslow’s aphorism:

> "when-you-own-a-hammer-everything-starts-to-look-like-a-nail"

which suggests that the existing tool becomes the default for every application. In the case of our WO observations I noted from the positions taken by participants that:

> "when-you-are-surrounded-by-nails-everything-starts-to-look-like-a-hammer".

By this I suggest that in the face of organisational/social challenges any available technology (including WOs) may be recruited as a candidate solution. This desire to recruit new technologies to previously challenging problems appears to explain the wide range of uses to which participants (from a wide range of backgrounds) suggested that WO should be put, again confirming that occupational framing be considered a central element to understand WOs.

In order to support multiple framings within a shared (cross-sector) WO community, the conceptualisation of WO needs not only to vary (as with Chomsky surface structures) but also to **flex** (tack) back and forth between centrally-shared (deep?) structures around the tool and contextually-shared (surface) structures relating to its application. Based on the air-brushing and contextual switching observed I suggest this qualifies WO as a so-called boundary object supporting divergent (or even inconsistent) definitions in the contextual use of the object while supporting a functional shared definition.

In contrast to the individual (cognitive) surface structure idea, the boundary object (in the work of Susan Leigh-Star) is a shared concept/object often at the organisational (occupational) level for the ecosystem of participants working with it. They share a definition/understanding when discussing it centrally (at a basic functional) level and yet this definition loses cohesion (decomposes) when discussing/applying the idea outside of this central location. (Star & Griesmer 1989) has three requirements to define a boundary object:

- **Interpretive flexibility** (which we see in the "airbrushing" of terms and framing of WO)
- **Connection to the structure of work process needs** (which we see in the recruitment/alignment of WO to specific solution and the desire for particular features)
- **The dynamic between 'ill-structured' and tailored uses of the objects** (which we see in the process of mapping between the platform vs. application framing of WO and W^3O allowing, in Stars words, an "arrangement of objects allowing people to work together without consensus".

Specifically, airbrushing (Star calls it 'tacking') occurs when participants re-frame or recruit the concept to match their own ecosystem requirements and rules. To be clear - the original shared definition *is not considered wrong and is not permanently replaced by the local definition*, rather the meaning of the Boundary object "flexes" between local and global usage allowing the participant to discuss/share the object on a wider basis while adapting to narrower constraints within a narrower context.

The insight we gain from recognising WO-as-a-boundary is three-fold:

- We may ultimately need to focus less on definitions of what a WO is in authoritative terms and more on which particular contexts/frames are rewarded for engagement with a WO and participation in the broader WO system (stimulating engagement)
- That boundary object theory predicts that consensus is not required for cooperative work to occur - thus implying that convergent or normative definitions in other areas may be

superfluous vs education around standards and interoperations (Stimulating interoperations/automation).

- The theory of boundary objects predicts (through the treatment of standards, methods and the creation of residual categories) a process of growing divergence between the requirements of peripheral contextual systems and a centrally maintained object leading to eventual reclassification and greater differentiation of terms potentially resulting in two new boundary objects. (Potentially leading to greater diversity/coverage without loss of access/interoperation.)

At the time of writing it is too early in the development of WO to observe/predict the impact of this effect or predict specific residual categories but one might presume the effect on tightly-coupled sets of WOs rather than on a distributed, loosely-coupled W^3O. Whilst we might consider each WO as a boundary object in-of-itself (Star 1988, Star & Greisemer 1989) to the groups of users that access it, the increasing set of WO→W^3O (implying a set of boundary objects) may more closely resemble a boundary infrastructure (Bowker & Star 2000).

10.3.3 WO-as-boundary-infrastructure

As the notion of WO 'tacks' to the notion of W^3O, the emergent Web-of-Observatories can be conceptualised as a collection of boundary objects. (Bowker & Star 2000) consider this offers the possibility of a boundary *infrastructure*. In earlier work by (Star & Ruhleder 1994) the authors set out the key features of infrastructure, against which we compare the findings of this project to determine if W^3O (notionally a collection of WO boundary objects) is also an infrastructure for which the criteria are:

1. **Embeddedness** - the setting of infrastructure into social arrangements - which we see in the occupational contextualisation of WO by participants and builders
2. **Transparency** - whereby the infrastructure persists across usage/solutions - which we see in the varied applications/interest of groups WOs rather than the system itself (also noted above)
3. Learned as part of **membership** - which we see from the community focus around WST/WSTNet for Web Science and ODUG/ODI for open data systems
4. Links with **conventions** of practice - which we saw in the investigation of the virtual astronomical observatory and the primacy of local convention over technical optimisation
5. Embodiment of **standards** - which we see in the production of both technical and legal standards for WO interoperation and usage
6. Built on an **installed base** - which we see in the alignment of WO with larger research programmes and foci.
7. Becomes **visible upon breakdown** - which we saw reported during early pilot stage events in which technical limitations highlighted deficiencies and, as noted earlier, highlights issues around the appropriateness of WO use in locations with poor infrastructure capacity/reliability.
8. Is fixed in **modular increments**, not all at once or globally - which we see in the bottom-up strategy proposed by Tiropanis and Hall in the evolution vs. definition of WO structures and incremental take-up by other partners.

Whilst we see some evidence of boundary infrastructure traits in individual WOs, the characterisation of W^3O as boundary infrastructure is a better fit and more expressive in the overall model of a WO ecosystem/ecology. We observe WOs→W^3O and also boundary objects →boundary infrastructure. WO is thus distinct from W^3O not only in terms of features/function but also conceptually as a collection of boundary objects resulting in the eventual possibility of boundary infrastructure, reflecting varied conflicts and co-operations across W^3O vs. more parochial themes within the scope of a single WO.

10.3.4 WO-as-a-novel-solution

WO may be applied in novel ways even if its constituent parts overlap with those of other systems. Since its applications and data may be highly varied/dynamic I have avoided a cataloguing process which offers only a snapshot of WO data/services as they currently are. Instead the research has contributed a collection of D, N, A facets which may be extended as required and from which specific applications may be built in/across different industries and sectors. Functional 'genes' serve several objectives here:

- WO planners/designers may select elements from which to plan new development to construct WO systems
- Existing application owners may compare existing functionality with the WO vocabulary in order to plan/estimate differential development to adapt applications to participate in the WO eco-system
- In doing so WO/W^{3}O may be compared to other applications to assess 'novelty' from a technology and symbology point of view. In doing so we address the question: 'Is WO just a type of [X]?'

(Malone et al., 2009) makes much of developing a genome of repeating functional and process genes in the design/development of Collective Intelligence systems.

A feature score was developed above between WO, W^{3}O and other web-based systems under the lens of Alter's taxonomy of decision support systems. The scoring method (based on direct system observation and document review) found none of the other systems to offer more facets of the taxonomy than Observatories. The difference was significant for W^{3}O but only marginal for WO and given the lack of interrater reliability (IRR) checks for this analysis we may consider the results to lie in a range +/- several points. Several observations emerge from this part of the research:

- (Subject to IRR corrections) it is possible that WO as a standalone system may not be significantly distinctive/novel in relation to other existing system such as Web Analytics or Data repository applications.
- W^{3}O appears to show greater distinctiveness (even when adjusting for IRR) suggested by the collaborative, distributed nature of the approach. From this we may conclude that W^{3}O cannot logically be subsumed by other system types i.e., W^{3}O cannot simply be a type of [X] if it apparently exhibits more functionality than [X]. This higher level of functionality (and the differences between WO and W^{3}O suggest a topic worthy of academic interest in its own right. It also suggests a complementary question regarding

the nature of other systems in relation to a WO eco-system: 'Is [X] just a kind of (a source for) WO?' i.e., potentially part of (subsumed by) the W^3O construct.

Using a functional count alone may be insufficient to determine how novel/innovative the application of a new technology may be. On functional novelty (Djorgovski & Williams 2005) note of the astronomical virtual Observatory VO (which inspired WO) that:

> "..any of the individual functions envisioned for the VO can be accomplished using existing tools (e.g. Federating massive datasets, exploring them in a search for particular objects, outliers or correlations but in most cases such studies would be too time-consuming and impractical and many scientists would have to solve the same issues repeatedly…VO serves as an *enabler* of science with massive/complex datasets and as an efficiency amplifier. "

One might conclude therefore that W^3O rather than WO is the more distinctive/novel approach while, of course, recognising that first delivering WO is necessary but not sufficient to enable W^3O.

Thus, we may consider that it is in the application of the technology (the ways in which Actors *execute* their goals via Narratives) and not simply the structure or arrangement of the technology where WO offers it's distinctive and valuable contribution. I submit it is reasonable and reflective of the contribution of individual WOs (particularly in terms of longitudinal study, curated data, trust, provenance and reusable apps/tools) even before participating in W^3O.

10.3.5 WOs and Social Machines

WOs have been characterised by (Rowland-Campbell [45] 2014) as "Social Machines to observe other Social Machines". While there is a compact elegance to this characterisation, there are risks in using this as a formal definition:

1. While academia is still establishing/debating definitions of Social Machines elsewhere (risking defining one unknown in terms of another unknown)
2. If one fails to recognise that WO-as-a-social-machine is only one signification of WO amongst many - risking rendering the definition to be of limited use/application
3. If one conflates being *part* of a larger Social Machine (a second-order relationship) with intrinsically *being* a Social Machines (first order attribute) - risking allowing the scope for Social Machines to be unmanageable/undiscerning.

[45] http://intersticia.com.au/wp-content/uploads/2014/12/WebObservatory.pdf

To evaluate the idea of WO-as-a-social-machine we may apply definitions from (Berners-Lee & Fischetti 1999) and (Smart et al., 2014) requiring three elements to be fulfilled:

- That machine elements and human elements co-operate/collaborate in determining outcomes (i.e., humans do not simply use the tools but notionally the tools also use the humans)
- That this interaction takes place at scale via the Web (i.e., single actors fulfilling personal goals may be part of a collective Social Machine interaction (e.g., GWAPs) but cannot individually be in-of-themselves Social Machines.
- That some social constraint/goal be the target

Whilst the discourse positions WO as a Social Machine we may consider both the community W^3O and individual WO perspectives:

	W^3O	**WO**
Machine and Humans collaboration	No current human processing Activity confined to collection/collaboration via data sharing/apps	No current human processing Activity confined to collection May collaborate via data sharing/apps
Operation at scale	Requires participants to disclose (some) assets globally and share (some) assets with other parties in exchange for some agreement/consideration Must involve >1 node or source and operate via the Web for discovery	May choose to share or keep all assets private. Not required to use discovery protocols. May be restricted to a single source and could operate via a private IP connection
Addressing social constraints	Intended to address global users through sharing which may cross themes and sources	Intended to address local needs of user community - may be theme-centric or source-centric

Table 10-2 WO-as-a-Social-Machine

Ultimately WO nodes are necessary-but-not-sufficient for a viable W^3O and may, or may not be "social" (1) in the sense of scale, (2) in terms of objectives nor (3) (in terms of what is currently implemented) focussed around collaboration/communication between participants. Using Malone's definition of joint working (see below) as "collection" vs "collaboration" we can see WO currently endeavouring to address the joint collection of data or apps/services but not yet offering integral elements of collaboration on specific questions or outcomes. This is in contrast to, say, the Zooniverse platform with communication channels and shared objectives (via projects/experiments) or Malone's Co-Lab focussed on climate change solutions.

W^3O then is the *aspirational*, shared, contextualised evolution of individual WOs and we are perhaps seeing evidence of the early stages for the pre-requisites of W^3O.

At this early stage WO also partly represents an appeal for the resources and engagement required to fulfil the extended vision of the W^3O evidenced by the body of vital preparatory work by local teams in the creation of standards, templates, APIs, demonstrators and pilot projects. Research participants sounded caution around the lack/loss of access to important data for research and the need to respond to the "data deluge". In this sense the WO might be thought to be a *social movement* leading to a Social Machine exhibiting several key elements to fulfil the definition of a social movement as requiring:

The formation of some kind of collective identity;	which we see in the creation of WST and WSTNet
The development of a shared normative orientation;	which we see in the creation of the WO project and the stated objective to collect/re-share data openly
The sharing of a concern for change of the status quo	5. which we see from interview participants around the difficulty of addressing a growing data deluge 6. concern around transparency vs privacy 7. data monopolies in a few large corporates 8. inaccessibility to non-programmers 9. concern around quality/provenance of data
The occurrence of moments of practical action that are at least subjectively connected together across time addressing this concern for change.	which we see in the delivery of events/workshops and practical solutions such as WO templates, WO licenses, demonstrator apps and integration tools

Table 10-3 Social Machine or social movement

> "Thus we define a social movement as a form of [political] association between persons who have at least a minimal sense of themselves as connected to others in common purpose and who come together across an extended period of time to effect social change in the name of that purpose"
>
> **(James & Van Seeters 2014)**

W^{3}O aspires to become a Social Machine in which researchers collaborate at Web scale on problems which resist the efforts/capabilities of single organisations and to do this the need for such a facility must be communicated and resources/partners recruited and so the social movement elements remain an integral part of this effort.

10.3.6 WO-as-a-set-of-genes

Developing a faceted taxonomy as a superset of facets (in effect a morphospace[46]) which transcends the features of any individual system suggested the biological metaphor of a 'gene pool' from which WO organisms might be constructed. The DNA organisation by features/definition [D], narratives [N] and actors [A] closely mirrors the (Malone et al., 2009) What? How? Who/Why? approach which he concedes is itself borrowed from organisational theory. This approach is modular without being reductive since it fully embraces the interactions and sequences of the genes. It offers several opportunities:

1. For comparison and planning for designers and systems engineers using D-facets
2. For establishing terms/trust/sequence for business process engineers and commercial/legal professionals using N-facets
3. For investment modelling, service planning and adoption management using A-facets.

The idea of a morphospace allows the exploration not only of what can be done with existing genes but also the ability to review/extend/replace genes in the model over time offering flexibility and hopefully conferring longevity to the approach.

The ability to consider parts of the DNA model individually as well as in concert using different socio-technical perspectives (see DNA AND NDA) fosters not only multi-team collaboration but also multi-model (interdisciplinary) collaboration.

Comparing this to earlier work by (Malone et al., 2009) on the genome of collective intelligence systems we see an apparently similar model called "What", "How","Who" and "Why". Malone's model, whilst ostensibly similar to DNA has several differences:

- "Who" recognises only actors working independently in a Crowd (sic) or under orders in a Hierarchy setting. There is no element of individual agency (as this relates to collective intelligence) and neither influence vs. authority or 'net behaviours' considered. Non-human agents are not considered here although Malone has subsequently reported on collective intelligence in groups comprising non-human agents. There are no actors who seek to consume the outputs of the system vs. those who produce.
- Malone's says it is "impossible to do justice" in describing "Why" and steps back from the complexity of social worlds. He calls his model a "simple overview" offering

"Money/Love/Glory" model (essentially "money" vs. "not-money") as possible motivations and, as such, is substantially less nuanced than Reiss' model on which DNA is based.

- "What" reduces the potential complexity of system interactions to "Act" vs "Decide" and whilst these may be valid super-ordinated collections they offer little guidance to builders/designers of systems who might attempt to create systems that "Act"
- "How" is described here as a procedural qualifier for "What" and comprises "Act"+"Collection or Collaboration" and "Decide"+" Individual or Group".
- More theoretical concepts/describing critical social factors such as trust and convention have no explicit place in the Malone model e.g., trust is not easily categorised as Who, Why, How or What.

Thus, D-facets represent the Design/Delineation of the physical feature set comprising processing elements, interfaces and boundaries. A-facets represent the Agents (Actors)/Agency in the system or ecosystem which drive/inhibit behaviour and reasons to engage and exchange. N-facets define the narrative or notional exchange via the exchange and are those information elements which cannot be 'installed' on systems or made identical with Agents/participants.

e.g., Trust is neither a function running on a server nor the identity of an agent interacting via the ecosystem but rather is the result of a narrative or exchange of tokens giving rise to the desire/motivation to act or abstain.

10.3.7 WO-as-a-project

WO-as-a-project runs currently as what might be termed a "skunkworks" - not dissimilar from the development of the Web from the first node - where diffuse elements are contributed by a number of partners without central funding and management. While this affords the project flexibility and agility there are also limits to resources and capacity which result from a smaller, agile approach.

While the characterisation of available sources and potential nascent Observatories stood at more than 50 for the 2013 review, the overall number at the close of this project in 2016 had not increased substantially in number though significant improvement in functionality were noted along with the appearance of open source WO templates by the University of Southampton. There have been notable contributions by WSTNet members in co-ordinating events and SIGs, and in deploying template WOs for other members leading to new WO nodes at KAIST (S. Korea), WebSci Aus (with UniSA) and IIIT Bangalore.

I note the similarity in the types of issues faced by WO and the Virtual Observatory Alliance whilst observing the marked *dissimilarity* in the project co-ordination and funding models. While interviews reveal that WO moves forward through individual contributions related to disparate (un-coordinated) funding and projects, the framework for virtual observatories was centrally funded by NSF/NASA in the US alone in excess of $16million [47] with smaller, but also substantial figures, funding 79 European partner data centres via the Euro-VO project and EU funding.

In reviewing archive material for VAO I note in terms of lessons learned in his final report (Hanisch et al., 2015) has far less to say about things that might have been done differently from a technical perspective than about lessons learned from a social perspective. Socially-embedded factors such as funding, project coordination, management (costing approx. $2.5m), outreach, education, marketing and generally managing for collaboration/compliance figure largely in his report.

The aims/ambitions of WO→W^{3}O are broadly similar to that of the VAO and interviews confirm that substantial funding to align the VAO partners was seen as a primary success factor.

While suggesting the benefits of more funding for any project will hardly seem insightful, I submit that future funding models as a means to coordinate the community of existing resources may require additional focus as the WO project begins to mature and accelerate.

10.3.8 WO-as-a-paradigm

Based on ideas in "From Search to Observation" (Brown et al., 2013), one of the most interesting WO perspectives from the interviews was a discussion with [David] in the Academic tribe. We discussed whether, rather than evolving as a set of discrete **tools/systems** (WO-as-Web-Observatories) that we might instead be seeing WO-as-Web-Observation: the evolution or **paradigm shift** from browsing and consuming pages/app-oriented data to browsing and consuming dataset/data-feed-oriented sources.

In this scenario Web Observation is something that ultimately encompasses/subsumes all tools/sources and becomes an overarching concept like "Search" or even more fundamentally simply becomes "what the Web is now". It is interesting to note in this case that even failure to establish a large set of bespoke interoperating WO nodes in favour of recruiting existing systems/sources en masse to W^{3}O would potentially be an equally successful outcome for the Web Science community. The WO project currently recruits existing search

[47] reviewed down from an original commitment of $27million

infrastructure/standards through schema.org to create web-scale resource discovery but we might also consider the extension of this discovery mechanism to provide more direct access to the datasets themselves through standard APIs or tools provided by larger corporates such as Google or commercial information brokers such as Bloomberg or Thomson Reuters.

10.3.9 WO-as-an-innovation

In this section we will consider WO in terms of innovation/adoption theory (see Ch2) comparing current observations with examples from open data and the astronomical VO as proxies.

As we saw in Ch1 adoption, particularly for infrastructure technologies, may take longer and require more than the simple disclosure or even availability of the technology. There are numerous surprising examples of (apparently) contemporary technologies vs. their original discovery dates:

Innovation	***Year invented/discovered***
Carbon Nanotubes	1952
Fibre Optics	1840's
Lasers	1958
DNA	1869

Table 10-4 Invention vs Adoption

We must therefore consider other factors beyond knowledge of an innovation to expect its adoption:

10.3.9.1 Innovation Diffusion

In Ch2 Rogers model of innovation diffusion was introduced consisting of five stages of innovation adoption with key factors influencing the process below (Table 10-5), I compare relative positions of the completed astronomical VO project with the developing WO project according to Rogers' model:

Stages	*VO (by close of project) t+10yrs*	*WO (currently) t+3yrs*
Knowledge	Widely promoted/understood all levels Key white papers available	Specialist promotion/understanding Key white papers available
Perception of value	Highly coherent between users contextualised in one domain	Apparently diffuse between users for diverse domains
Decision	Funded by government	Funded by implementers Free rider problem / costs
Implementation	Large, centrally - funded teams	Smaller skunkworks - tactical funding
Confirmation (observed vs. experienced)	High profile shared collections Commercial support Strong Government support	Open data collection Commercial competition Weak Government support
Influence Factors		
Relative advantage	No better practical alternatives	Competes with Commercial offerings and other open data platforms
Compatibility	Central to problem /focus	Central to problem/focus

	Special data formats (social requirements) Central directory	Open data formats Open discovery markup / distributed directory
Complexity/Simplicity	Multiple interfaces incl. non-technical with comprehensive supporting material Focus on political acceptability over technical superiority	Currently Web portal, API and hard-wired apps Basic support material Focus on pragmatic technical solution
Trialability	Web-based, API and integrated tool suite Many demonstrators Academic Conferences	Web-based, API - tools not yet integrated Few demonstrators Hackathons/WO events
Observability	Eventually high	Currently low

Table 10-5 Adoption Stages/Influence Factors. Adapted from (Rogers 1995.)

Key areas of note are:

1. The differences in funding (the US VO alone had an initial budget of $27 million) and the relative lack of government support/funding for WO vs. alternative data-oriented incubators and initiatives. A fundamental difference here is the commercialisation of data science vs. astronomy (there is little or no commercial astronomy per se) and the very different competitive landscape that results.
2. The wider applicability of WO leading to a less well-defined conceptual solution (WO means very different solutions to different groups) and so broader contextual material and demonstrators may be required to underscore different business cases and account for diverse and even conflicting motivations

In astronomy, there is little advantage to withholding data in the long term (apart from an initial embargo → 'first publisher advantage'), and datasets and algorithms in this domain would rarely be considered proprietary or closed but instead, would be published for attribution/citation. The

new data economy, however, offers considerable advantages/business models for those who hoard data and license their data assets and analytical insights. This makes Big Data and (proprietary) data science techniques/algorithms the centrepiece of the data economy and constrains the willingness to release data openly.

Below (Fig 10-2) I offer the availability of public linked data sets as a proxy for what might constrain the value/adoption of WO datasets which also need to be prepared for use by the WO. Whilst there is growth (particularly recently), 1'000 fully open data sets in total (approx. 2.5% from a total of >40'000 datasets) would be considered very early adoption and far short of Rogers 15% requirement for the start of mainstream adoption. It should, however, be noted that even single data sets (such as DBpedia) can be extremely large (many millions of triples) and highly valuable, even in isolation.

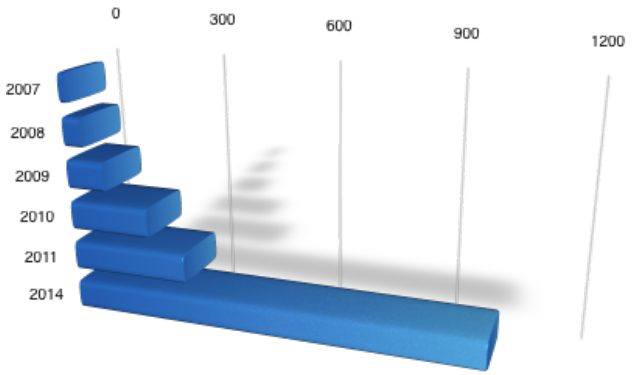


Figure 10-2 Visualisation of linked data sets. Adapted from
http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

Whilst WO does not require linked open data sets per se, analysis from the CKAN-based data.gov.uk indicates a low level of adoption for data sets which require *preparation* with meta-data/mark-up.

Openness Score	*Descriptions*	*Number of data sets*
None	No attributes	18’361
★	On the Web with an open license	828
★★	Machine readable format	1709
★★★	Non-proprietary format	14’571
★★★★	Uses RDF standards	None (!)
★★★★★	Available as linked RDF	127

Table 10-6 Openness ratings. Source data.gov.uk

Notes on potential errors

1. No figure is cited for 4-star data sets (possible omission)
2. The total accounted for here by CKAN is missing >5’000 data sets some 3’800 of which are ‘unpublished’

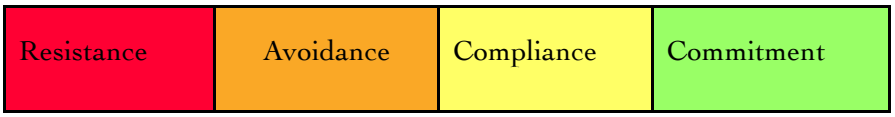
Whilst it should be stressed that the figures above are for government open data sets rather than directly for WO data sets (currently reporting 70 data sets) the principles of adoption are closely aligned and the general lack of appetite for preparing/curating data for others to use without a commercial benefit remains directly applicable.

Lessons from Rogers to develop WO adoption:

1. Closer, more explicit associations with both Big Data (commercial)/Open Data (Community) may attract better 'halo effects' in terms of visibility and resources. Additional communications/outreach programmes to associate WO with challenges outside of academic research may broaden appeal.
2. More examples of working demonstrators (which are beginning to appear through WUN) would help to emphasise the specific advantages of the WO approach to diverse user communities
3. Examples should be developed using the simplest possible variant of the required technologies especially in terms of formats and rules to reduce barriers to entry
4. Tools/Guides to convert existing data sets to discoverable/re-usable formats would drive more Trialability as would the creation/development of communities or groups with a natural need or tendency to want to share data with each other (such as specific government agencies, police forces and other special interest groups)
5. Once these groups and demonstrators are developed, the process of observing, analysing and then further promoting/advertising the benefits can be undertaken through case study research and dissemination.

10.3.9.2 Innovation resistance and persuasion

In terms of innovation adoption Klein says that there is always a level of *resistance*, which is defeated or persists giving rise to net mood or characterisation regarding the innovations along a continuum (negative to positive) of:



Which is dictated by a combination of factors

1. innovation values fit (Poor, Neutral, Good)
2. implementation climate (Strong, Weak)

	Innovation-Values Fit		
	Poor	Neutral	Good
Strong implementation climate	Employee opposition & resistance Compliant innovation use at best	Employee indifference Adequate innovation use	Employee enthusiasm Committed consistent & creative innovation use
Weak implementation climate	Employee relief Essentially no innovation use	Employee disregard Essentially no innovation use	Employee frustration and disappointment Sporadic and inadequate innovation use

Table 10-7 Innovations fit. Adapted from (Klein et al., 1996)

Lessons from innovation fit/Innovation resistance

In terms of sentiment, the case for building new WO's and WO applications may not yet have been sufficiently well made to capture the commitment and support of managers and teams widely particularly as these business cases may be distinct from those of the academic team who are currently building the templates/demonstrators. Also even if we allow the values fit to be 'Neutral' or even 'Good' and we see a lack of skills to confidently and effectively implement these systems in practice, the weak implementation climate will create substantial barriers to adoption. Technical teams, without the resource or skill to build and support these systems will remain firmly in the resistant (lower) part of the spectrum where only sporadic or inadequate use of the innovation is made.

- Virtualised environments and/or wizard-based automated installation/configuration procedures may help to bolster weaker implementation environments
- User training events with carefully prepared scenarios and tools will assist in emphasising the accessibility and ease-of-access to relevant data

10.3.9.3 Disruption, Competence and Value Chains

Additional perspectives on adoption come from (Bower & Christensen 1995) on disruptive innovation in which existing providers fail to respond to challenger brands due to a mismatch in cost/benefit models of the current market vs. the modified market represented by the disrupting innovation. Whilst this offers an explanatory model for the actions of the encumbent providers and historic examples of where disruption has succeeded it fails, however, to outline reliable methods to predict disruption or instructions on how to ensure disruption.

Two aspects seem relevant here:

- At this early stage in WO development it seems less relevant to ask if WO is *currently* disruptive – only if it is *potentially* disruptive and to reflect on who might be disrupted
- The inclusive paradigm of cooperation and interoperation required for W^3O to flourish would seem to indicate that intentionally disrupting the broader range of information providers/collaborators might be counter-productive – given that W^3O is not underpinned by the idea of gathering/storing data directly but rather through distributed collaborators.

(Tushman & Anderson 1986) argue that it is the competence-*enhancing* or competence-*destroying* nature of the innovation which determines the success/failure of the new technology. Where a new technology simply competes at being good/better without changing the perceived value of previous investments in existing technology, then defence for the incumbent is relatively easy whereas innovations which detract from investments made in existing technologies make disruption by the market entrant more likely.

Participants talking about WO certainly showed evidence of anticipated competence-enhancing features (either for themselves or their constituents/customers) though it is still unclear how easy achieving these features may be and to what extent the free/commercial tension may set competence-enhancing features against revenue-reducing features in commercial/budgetary considerations.

(Christensen & Rosenbloom 1995) observes that *value chain* effects may dominate here and that value creation/destruction must be considered not only within the incumbent firm but across the incumbent value chain. This is a view core to the (Adner 2012) model. It proposes that *entire ecosystems* of actors determine the ultimate adoption/success of innovations and that an eco-system view (using a so-called 'Wide Lens') should be used to ensure adoption/support/compliance from all the parties involved in the value chain. For W^3O, in particular, the wider adoption chain is critical and so considering W^3O through Adner's "wider lens" appears highly relevant.

Lessons learned from disruption, competence and value chains

WO may consider from these models that adopters will not only be focussed on the platform itself but as the research has shown also on outcomes and particular topics or data sets/resources and the competencies that relate to these perspectives. The project should consider the super-set of the perspectives to form part of the eco-systems of actors and objectives but may wish to engage with them more specifically in terms of communication, training and outcomes. No work has yet been undertaken to understand which parties, if any, suffer a *dis*benefit from the

adoption of WO. We have seen the significance of political/economic issues in the history of the virtual astronomical observatory and cases of how technical elegance/efficiency was dominated by organisational/political issues.

10.3.10 WO-as-Knowledge-Infrastructure

Considering WO as an example of a broader (socio-technical) approach for the capture/management of knowledge and, critically, how such systems may be socially- rather than technologically-shaped, we find relevant work in the study of CSCW (computer supported co-operative systems), SCOT (the social construction of technology) and the work of Pinch and Bijker (see Ch2) and in particular from Pollock, Williams and Edwards in addition to the authors previously covered earlier.

(Williams & Edge 1996) highlight the tensions around innovation & adoption between resistance to adoption (or entrenchment) which is fuelled by sunk costs in existing technologies, convention/standards (e.g. the QWERTY keyboard) vs. the pressure (they call it dynamism) of relentless technology supply seeking to create differentiated offerings and competitive advantage (see the review above of disruptive innovation). For WO, this may clarify the initial resistance seen to WOs in the absence of direct WO experience and signals the need for awareness not only of perceived benefits but also who may be *dis*-advantaged or *de*-incentivised by the adoption of WO technologies and infrastructures.

Williams specifically flags the need for social lenses through which to understand, not only the creation or structure of technologies, but specifically claims that the process is "shaped by a range of broader social economic cultural and political factors" and judges the deterministic linear model of technology supply and adoption as insufficient. Technologies, the authors claim, often emerge through a complex process of action and interaction between heterogeneous players with failure (destabilisation) rather than success (stabilisation) often being the observed results. This results not purely from the nature of the technology itself, but also as a result of the political alignment of stakeholders and net perceived benefits of adoption - characterised as turbulence within the "negotiability of technologies".

Williams refers to Fleck's idea of "innofusion" in which the adoption of technology is shaped through the process of "learning by struggling" which is related to the idea developed here of problematising/contextualising WO technology. In (Pollock & Williams 2010) the typical methods used to evaluate technological adoption are critiqued as being too focussed on single-site ethnographies rather than multi-group, multi-perspective experiments thus lending weight to the choice of multi-case study in this research. They highlight the typical approaches to study (impact

studies and implementation studies - reporting only on what was done) as not only often lacking objective criticality but also failing to dig deeper into questions of *why* the technology was chosen, *what* was expected and whether *expectations* were met, modified or otherwise impacted by the "struggle to get the technology to work in useful ways at the point of application". This viewpoint strongly informs the recommendation here for future longitudinal WO work to explore not only the process of adoption but also the 'struggle' to realise expected benefits as observed in the early WO Pilot work.

A key point made in this project highlights the difference between WO as a technical artefact and W^3O as a collaborative framework which echoes, and is supported by, (Edwards et al., 2013) who make a clear distinction between smoothly designed, 'coherent', often tightly-focussed (knowledge) *systems* and the distinctive messiness and social/political disharmony that can arise from knowledge *infrastructures* (a messiness also alluded to by Berners-Lee and Tiropanis & Hall Ch2).

Knowledge infrastructure is defined as "robust networks of people, artefacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" thus mirroring the conclusion that technological models of systems like WO are necessary-but-not-sufficient to capture the diversity of perspectives and interactions at Web scale.

"All infrastructures", Edwards claims, "embed social norms, relationships and ways of thinking, acting, and working" exhibiting "unique origins and goals" - thus supporting the focus on recruiting external WO sources/apps and the tribal/social perspective on WOs offered here.

Knowledge across technical, disciplinary (ontological) and tribal boundaries are often contested (Edwards et al., 2007) something referred to as "science friction" (Edwards et al., 2011) and the nature of data sets as freely interchangeable (fungible) commodities leading to (inexpertly) remixed datasets (Weinberger 2014) refers to meandering collections of knowledge as 'playlists' which risk losing a grounding in the all-important context of data collection and data curation.

Whilst undoubtedly scalable and powerful and potentially valuable, the authors point out the risk of unfocussed crowd-sourced data collection with no peer review which may sacrifice data quality for data availability (something they call 'data arbitrage'). We saw this effect in the WO hackathon chapter. They hint that a lack of peer review, certification and curation/context may lead to deeply divided (flawed) interpretations of non-persistent data by groups seeking to support their positions/objectives and making political ground through the medium of knowledge and knowledge infrastructures.

Comparing Infrastructure principles with findings for W^{3}O we see:

Infrastructure principles portray..	**Related WO findings suggest..**
the examination of early publishing/libraries (Edwards et al., 2013) asserting that "changes in knowledge infrastructures may reinforce or redistribute authority, influence, and power".	parallels to power being inherent in the technology which delivers and shapes information/data/media in Ch2 and our parallel with the work of Kittel.
the nature of knowledge infrastructure, not as a cohesive technical arrangement, but as a messy, deeply political construct	the Astronomical VO exemplar discussed in Ch 8 supports this view with social/political issues figuring more prominently in several accounts
that whilst systems need fixed global standards for realistic accessibility, users need local flexibility for realistic application	the repeated contextualisation of WO within an occupational context seen in Ch7 and Ch8 supports this notion
that the ascendency and advantage gained by one infrastructure group potentially acting at the expense of another - equating the adoption of new infrastructure to a power struggle/political act.	the resistance (or support) by participants with little/no experience of WO (in Ch4) suggests a hegemonic or political dimension
that new types of data worker may be required to bridge these cultural, ontological and procedural gaps	the role of Web Science in this space is alluded to but not developed in this project but does reflect the researchers own experience
the use of actor networks, boundary objects and trading zones may support this vital bridging work	boundary objects/infrastructure were identified (above) within the context of the project as potential models through which to understand WO
that long term preservation/curation is key to realise benefits from knowledge infrastructures	long term preservation/curation was identified (in Ch4 and Ch8) as a key distinguishing factor between WO and other ICT's

that an extensive adoption/maturation period may be required before the full value of WO→W^{3}O is 'metabolised'	differences in confidence/sentiment across age groups was identified (in Ch4) during one of the exercises.

Table 10-8 WO-as-knowledge-infrastructure

(Edwards et al., 2013) makes seven recommendations for Knowledge Infrastructures:

1. Create and nourish mechanisms for large-scale, long-term research.
2. Build interdisciplinary collaborations across natural and social sciences.
3. Develop comparative analysis techniques for studying large-scale, long-term data.
4. Create sustainable, shareable data archives.
5. Build better software for qualitative work.
6. Integrate qualitative work with statistical techniques and social network analysis.
7. Imagine new forms of cyberscholarship

Each of which harmonises closely with the stated objective of the Web Science community broadly and as more specific objectives for the development and continued improvement of the Web Observatory.

As suggested in the conclusions of this thesis, the WO differs importantly from W^{3}O through its inherent (and potentially diverse) community of participants, sources and objectives. It was suggested here that the diversity of intent and context even for a single system qualified WOs as potential boundary objects with the suggestion that W^{3}O be considered a boundary infrastructure. The definition used by Edwards et al equally evokes the messy, political, emergent idea of knowledge infrastructures rather than knowledge systems.

10.4 Conclusion

In this section we have evaluated the grounded theory which has emerged from the project using standard criteria including fit, relevance, workability, modifiability. The model successfully reflects both technical and social elements and does so from a *genome* perspective which retains high levels of flexibility and nuance to account for broad differences in the conceptualisation and application of WO by users. We reviewed several established models of innovation adoption and noted that adoption is also often subject to an array of social and technical factors providing a further fit between the DNA model and our understanding of how adoption may be managed using this approach. Observations/recommendations are made for consideration in accelerating adoption for WO systems through the creation of more nuanced materials and examples in order to reflect the complex and sometimes contradictory objectives of the wider WO eco-system of users.

In the final chapter, we will summarise the findings from the project, reflect on answers to the research questions and consider the robustness of the findings and possible future work.

Chapter 11: Conclusions

In Short..

A review of the research context is offered, and key findings from the project are presented, linking findings to the original research questions. I reflect on what has been learned, the contributions and the limits/caveats around the approach. Future pathways for the work are suggested, and a final note reflecting on the research process is added.

Project Summary

The project has assembled input from 30 organisations and more than 100 interview participants across 76 interview sessions. Content and ideas from 800 publications and 100 survey respondents were reviewed with the purpose of building an accurate, useful and extensible definition of Web Observatories. I build on existing work, which primarily offers technical and production-oriented[48] views and develops a new socio-technical model of WO recognising both production and consumption[49] perspectives. The definition takes the form of a faceted taxonomy of WO "genes" and its representation through a three-part model called DNA.

11.1 The Research Context

Whilst the idea of sharing research data through virtual observatories has been seen before in astronomy and physics to investigate physical phenomena, as more and more of our lives are conducted online, the impact and significance of Web phenomena increases. This creates the opportunity (and the need) for observatories relating to the usage and structure of the Web which is driven by:

- The ever-greater measurement (increased *datafication*) of our interactions
- the increased ability to analyse/store data (greater *resolution*)
- lower cost and more rewards through numerous business models (greater *motivation*)

thus creating a self-reinforcing "data deluge" beyond the capacity of any single system or group to manage.

[48] Relating to those who build WOs or offer WO services
[49] Relating to those who require/use WO services

Fortunately, the Web also offers a solution to this challenge through the possibility of sharing the task of capturing, analysing and interpreting these vast data resources (certainly *collectively* and perhaps *collaboratively*) through distributed catalogues and shared standards in the form of virtual Web Observatories.

Just as the notion of the Web emerges from a vast array of individual Web Servers and connected devices, so a Web of Observatories (W^3O) emerges from an array of individual Observatory servers (WO), data sources and analytical services. Thus a distinction is made between the individual physical WO "boxes" and an emergent W^3O service/community.

W^3O is a thus a superset of systems and resources requiring an inclusive/collaborative approach which may overlap Web *Science* Observatories with non-academic data repositories, analytical systems and collaboration environments that are grounded in other contexts, purposes or communities and which has a novel open, multi-source, multi-system, multi-purpose perspective.

In addition to the vital standards/processes that underpin the *ability* to participate, the potential drivers/rewards for adoption and participation in the wider community must be understood, developed and communicated in order to achieve a coherent W^3O structure. The corresponding barriers, difficulties and objections must be considered. Simply put, users/organisations must also have a *reason* to participate. Thus the DNA model which emerges here extends the idea of purely technical architecture with three perspectives:

- **Definition** - by defining and delimiting what WOs are in a physical/functional sense
- **Narratives** - by defining how WO interactions are played out or "performed" through a series of narrative exchanges
- **Agency** (Sociality) - what local meanings/ambitions WOs convey for users and how these are balanced by social forces thus driving net behaviours and reasons for action/engagement.

Different perspectives/ambitions for WO users drive both technical and social/political challenges and compromises which need to be overcome before W^3O can achieve significant adoption and reach its goal of engaging systems/people **at scale** (a requirement for W^3O to emerge as a true social machine). Until then, the Web Observatory project retains elements of:

- A fluid "start-up" innovation project
- An aspirational *social movement* championing a cause/vision for participants
- Elements of a latent *social machine* aiming to achieve transformational capability and impact at scale.

The specific path to impact is unclear. If the WO movement is successful:

1. WOs may emerge as a new distinct class of IT *system* orchestrated through a W^3O social machine or
2. "Observation" could alternatively emerge as new/evolved form of Web *interaction* underpinned by existing systems/services on the Web.

In either case, this research delivers a model from which new observatories may base their thinking and from which existing systems may adapt their operations/processes to align to a global Web of Observatories.

11.2 Document Review

Chapters 1-3 introduced WOs as a powerful/complex contemporary tool developed in response to assist in the analysis of a correspondingly vast and complex contemporary Web. The roots of Web complexity and the emergence of the WO within Web Science context are discussed and the literature around the Web, Web Science and Observatories are surveyed. A plan is set out to investigate the WO phenomenon, to construct from observations what is distinctive about it in relation to other classes of system and to investigate how mapping structure/variations within WOs might inform our understanding of how new WOs may be built and how they might collaborate/interoperate to form a W^3O eco-system.

Chapters 4-8 present a series of experiments and cases moving from early pilot work and 'straw man' models which assist in the iterative development of a substantive WO model grounded in broadly-based participant interviewing and interpretative phenomenological analysis.

Chapters 9-11 present the final three-part model known as DNA which represents a vocabulary of WO genes from which WOs may be construed and constructed. It gives examples of how WOs may be represented and compared through this lens. In this final chapter, the insights, conclusions and caveats around the work are reviewed suggesting future directions for further research.

11.3 Results/Findings

Whilst more detailed descriptions are captured in their respective chapters the findings from the research can be summarised as follows:

1. The WO is a small, early stage project with no centralised funding attempting to create viral engagement in collaborative research for Web Science. The path to adoption comprises the creation of specialist WO nodes plus recruiting pre-existing systems/sources integrated through cataloguing and discovery services. This work is inspired by earlier virtual astronomical observatory projects though notably by comparison this earlier project was heavily funded internationally over a ten-year period.
2. The work on observatories therefore notionally comprises two areas: the individual **WO** node and the emergent community of interoperating WO nodes/users which we have termed **W^3O**. Conflation/confusion between the nature and implications of these two perspectives was seen extensively throughout the project at all levels of WO experience.
3. Despite the availability of many candidate systems/sources to become WOs, few have so far chosen to do so, even within the Web Science community, despite an active/successful WO project, a growing literature and live demonstrator systems. This may partly result from a "fluid" conceptualisation of WO (what it is and what it means) which appears to vary substantially across participant groups, thus contributing to the potential issue of coordination/participation.
4. Knowledge of the WO itself seems secondary (or even transparent) to the way in which the WO is conceptualised or **framed** by each participant narrative. This framing appears largely occupational/social in nature. Even a "blind" questionnaire, run in the absence of any specific knowledge about WO, demonstrated the tendency for social/occupational framing.
5. Despite variations/adaptions rooted in occupational framing and contextual WO application, participants did recognise more fundamental structural principles/elements of WO whilst switching between more/less contextualised views. This suggests the possibility of WO acting as a flexible **boundary object** which has implications for how communication/adoption may be managed through roles which emerge around the creation/use of WO.
6. A body of conceptual elements for the structure/usage of WO were elicited using content analysis of documentary material and structured questionnaires and a proof of concept project was run to validate these concepts, to test data gathering techniques and to gather feedback from WO users "in the wild".

The findings were:

- Social/motivational elements and narratives were poorly represented in the documentary material indicating supplementary interviewing and observation were required to obtain good quality data.
- Direct questionnaires about detailed WO technology/design were ineffective in gathering/confirming technical features since few participants wanted/needed to understand the *technical* features of WO (transparency) in order to associate with the desired *outcome*. Technical models were more easily extracted from documentary sources.

7. In terms of feedback early (pre-production) users gave valuable feedback expressing concerns around practical/operational barriers to using the WO:
   - Concerns around network performance and accessibility to WO and its data,
   - Low confidence in the provenance, meaning and quality (re-usability) of third-party contributed datasets
   - Low confidence around technical (programming) skills deficits in leveraging the WO functionality.

   Whilst this not a critique of any live WO systems/deployments as they currently exist [50] it does offer insight into "points of potential pressure" for new WO systems generically and the types/levels of service and support that users will require. This also informs operational design/policy requirements for WOs. Key issues of network/compute resources, and the issue of 'dark data' for both content and algorithms are of primary concern.
8. Based on user interviews and a lack of data on motivation/demand in the document corpus to act as a proxy for motivations, an existing model for motivations by Reiss was adapted as a substitute for direct enquiry and validated against data obtained by Open Government data requestors. In the broader data gathering process, it became apparent that rather more information is available on **data supply** than for **data demand** and more generally the

[50] Particularly since this feedback was gathered during an event conducted as a "hacking" exercise with predefined, largely undocumented resources, low-speed internet connections and firewall policies "unsympathetic" to this kind of research.

consumption perspective for WOs (why data was required) was identified as underrepresented in the existing literature vs. the *production* perspective (what data was being offered).

9. The conceptual elements were structured and presented as a faceted taxonomy but this format, whilst flexible/extensible, offered poor visualisation and little ability to model/comprehend structures, flows and relationships between the facets. Thus additional visual models were developed to show structure, values and narratives resulting in the DNA model which offers perspectives on technical, social and narrative elements.
10. Extensive interviewing/observation was conducted corresponding to a broad review across three social classes of user that were termed *Tribes*. The attitudes and requirements of these tribes were gathered and compared with the straw models using iterative grounded theory coding approaches whilst deeper analysis using IPA was conducted with selected members of each tribe. This resulted in broadly shared tribal models and roles - an alignment which is unsurprising given that social learning and enculturation is thought to take place in such tribal groups. Given that tribal differences arise from culture, the alignment interoperation *between* tribes (both for processes and motivations) becomes an issue to be considered.
11. Whilst it is clear that standards for data and technology are vital for interaction/interoperation, no significant variation in technological (D) WO elements were discovered in the core conceptualisations of WO across the academic, business and community tribes. This is perhaps due to general homogeneity and shared "best practice" in modern technical platforms. Correspondingly it was observed that the meaning (signification) of WO from the Agent/Agency (A elements) and the Narratives (N elements) that are required by them were more variable/nuanced. Therefore, this variation is potentially more important to understand with regard to participation/adoption if we assume that technical alignment may be more straightforward than social/political alignments. This view is borne out by interviews and reports from the experiences of the astronomical VO team whose work inspired the WO.
12. Further analysis was conducted to identify **cross**-tribal commonalities as a basis for adoption management between Tribes and three such perspectives/roles were found to emerge from each of the three tribes. These were

    - **Curator** syndicate - broadly focussed on data itself generally or about specific topics
    - **Architect** syndicate - broadly focussed on structure/technology/processes
    - **Innovator** syndicate - broadly focussed on impact, outcome or the "advantage."

The existence of such syndicated roles explains both the "transparency" of WO technical details to non-WO Architects and the flexing of focus between the central WO artefact and the content (for

Curators) and the application of WO (for Innovators). This contributes an important insight in the cross-management of groups when encouraging innovation adoption since innovation resistance may vary by both tribal role and WO role. It suggests that WO *SIGs* [51] might be organised along these lines rather than solely according to generic WO lines.

11.4 Recommendations/Observations

In addition to the production of models underpinning technical/process design, much user feedback has been gathered during an extensive interview process. Key touch points for users and possible actions/opportunities for the WO project are reported here as indirect findings of the research.

A core objective of the WO project is participation, adoption and collaboration between isolated WOs delivering W^3O. This delivers network effects and synergistic opportunities which arise from the wider (global) scope of W^3O and its diverse sources of data, analytics and participation that exceed those of any single WO. W^3O is perceived by many participants as the distinguishing feature of Web Observatories rather than their private or stand-alone operation. W^3O cannot emerge without WO and hence both aspects need to be considered:

Enabling W^3O through WO adoption requires two core activity streams:

- Enabling **technical integration**/interoperation between WO nodes not simply between a node and its own users
- Enabling **engagement/adoption** by WO users with varied objectives/conceptualisations and not simply within a single tribal/cultural framework like business-for-profit OR academia-for-citation

Note that both of these points are substantially less problematic when viewed from the perspective of a single WO operating independently under their own rules and for a private/limited user community.

Addressing (1) requires:

1. Creating/adopting/extending standards for data and communication
2. Reference architectures and applications
3. API's for operation/workflow between WOs and/or non-WO platforms
4. Communicating/rehearsing/supporting the development.

[51] Special Interest Groups

Addressing (2) requires:

1. Creating/adopting/extending standards for licenses and terms of use whilst recognising that not all users will share data/services FOC and nor will all data be open.
2. Aligning the interest of users/funders/contributors through the creation of demonstrators and targeted applications around:
    i. Subject/Topic data or data curation per se (WO Curator perspective)
    ii. Technical platforms and platform technologies (WO Architect perspective)
    iii. Outcomes/Causes shared by communities/regions/organisations (e.g., WUN demonstrator WOs) (WO Innovator perspective).
3. The nature of innovation resistance (Ch2/Ch10) should be considered and addressed through the creation of (successful) case studies and support material for liability/licenses, training, development, integration and automation to remove the perceived barriers of difficulty and cost of getting-up-to-speed.
4. Unequal access to network/compute capacity should be considered for those users with limited resources if we are to avoid observatories-only-for-the-well-funded. Some form of "WO lite" access might be considered offering batch access, caching servers or smaller/conflated datasets for those with poor network access[52].
5. Unequal access to technical/programming skills should be considered. Simplified access and training should be targeted as a priority for those with limited technical/programming skills to reflect the growing expectation of "drag-and-drop" interfaces for data analysis such as those seen in highly successful products like Tableau and WO variants such as Recorded Future, COSMOS or Quid.com
6. Developing opportunities for W^3O in the area of licensing for synthetic/composite data sets which currently make sharing and re-use of non-open resources or those with heterogeneous licenses very challenging for all groups. Engagement with open licensing standards such as ODRL may be beneficial.
7. **Focusing on trust in the data and services that are offered is perhaps the most significant opportunity for W^3O.** The perceptions around fake news and the perceived vulnerability from manipulation/contamination of data have wide-ranging consequences for business, academia and government. W^3O may succeed through the establishment of a trusted and transparent platform in this space. This may emerge through the development of suitable provenance and data citation services and in doing so will create

[52] Based on internet/web access projects for developing countries such as BRCK and LibraryBox

distinctiveness from other open or commercial systems which may be more easily subverted without a secure system of record.

11.5 Addressing the aims of the Research

In this section, we will consider to what extent the findings of the research address the original research objectives. At the outset of the project three research questions were identified to explore the nature of WO, the conceptualisation/contextualisation of WO and the potential benefits to adoption of substantive model based on current WOs:

RQ1 - *Which perspectives can help us to clarify the structure and nature of WO not only as a purely technical artefact but also as an assemblage of users + technologies in a social context?*

Addressing RQ1

To do this several supplementary questions were considered:

- What are the social and technical elements of WO?
- Are there other types of element?
- Is the WO predominantly a context-free tool or is it attempting to address 'social constraints' at scale in the form of what Berners-Lee called a 'social machine'?

A three-axis model (including a "boxology" notation to capture narratives) with more than 100 elements has been developed along with a faceted taxonomy of WO concepts under the banner DNA. The three axes are influenced by work from (Matthewman 2011), (De Roure 2012[53]) and (Monge & Contractor 2003) suggesting technology offers/requires multiple perspectives:

- Investigating the technological/physical elements of WO: what (or who) are the components of the WO systems that are "functional" and their associated structure. **WO-as-an-Artefact**
- Investigating the "performed" elements of WO: the abstract narratives and notional exchanges of value between users/systems. **WO-as-a-Platform/Community** with an associated series of actions
- Investigating the meaning of WO to users: who (or what) imparts meaning/motivations for using WOs. **WO-as-a-Solution** (incl. knowledge)

[53] http://www.slideshare.net/dder/web-observatories

The resulting DNA model (Ch9) is thus more grounded in the WO discourse than simply capturing technical vs. social elements or functional vs. application elements. In addition to the social vs. technical perspective, additional narrative elements were found to emerge internally through the interaction/performance between the WO and its Actors or externally as eco-system inputs or emergent properties/outputs from WO. These are captured in the E^5 model which add both structure and a phased flow to the five classes of narrative comprising **Eco-System**, **Encounter**, **Enhancement**, **Execution** and **Emergent** phases/properties.

This creates a bounding box expressing a vocabulary of genes and perspectives from which WO's may be construed and constructed (i.e., what they *may* have) rather than a normative list of the elements which WOs *must* have.

As commented above, the WO (and particularly the W^3O) are *potentially* social machines. This is partly because the requirement for operation "at scale" has yet to be fulfilled and also, at the time of writing, few Web Observatories have yet developed mechanisms to actively support communication/collaboration between human and non-human agents, which would underpin the mechanisms of a fully-developed social machine. Such mechanisms are observable in related systems (such as citizen science platforms and hybrid crowd/machine-learning classification platforms), but these do not necessarily fulfil other criteria to be considered Web Observatories.

RQ2 - *If we consider WO as being socially-embedded in the processes and ambitions of different groups who use it is there evidence to suggest that WOs might be perceived/operated differently across social groups?*

Addressing RQ2

The supplementary questions were:

- What are the implications of differences for engagement/interaction with WO between groups?
- Is WO innovative (novel) technically and/or socially with respect to other technologies and approaches? How does this affect adoption?

A set of parallel observations across three distinct social tribes (Academia, Business and Community) engaging with more than 100 interview participants formed the basis of enumerating the elements (particularly narrative and motivational elements) of WO as it is signified in different groups. Additional questionnaire work confirmed the existence of social/occupational framing as an important consideration. Particular attention was paid to developing narrative/contextual models for selected groups in an IPA analytical exercise and a comparison between groups was performed. Via such narratives, we may gain better understanding which

motivations/measurements drive behaviours (including adoption) in different groups whilst the multiple sometimes conflicting elements provide important insight into engagement and innovation resistance. Socially shared perspectives were identified *within* the tribes and whilst these confirm an understanding of typical resonance/dissonance between public/private sector groups, important shared perspectives/roles emerged *between* tribes with implied opportunities for communication/engagement between diverse WO groups.

The notional affordances of WO and W^3O were examined within the context of an existing taxonomy (for decision support) which afforded the opportunity to compare a range of technical systems. WO as a standalone system whilst highly functional in its own right is less distinctive than W^3O which exceeded all other classes of system that were compared in terms of affordances and hence could not thought to be a variant (sub-class) of them. It is important to note that these are proposed/notional affordances (from the vocabulary/taxonomy of affordances) and hence if such functionality were not implemented the distinctiveness would not be established.

RQ3 - *What benefits can a socio-technical model of WOs offer in terms of insight into the creation of new observatories, innovative applications and the encouragement of participation by existing systems and data sources?*

Addressing RQ3

The supplementary question which developed was:

- What would a substantive model of WO look like and how might it be leveraged?

A set of visualisations/templates for each axis in DNA has been developed, and these can be used either *absolutely* to plan new WO systems/services or *differentially* to extend/enhance existing system to join the W^3O eco-system. A suggested approach has been developed not only for analysing WOs in terms of DNA elements individually but also in terms of narratives and cause/effect models which allows for multiple perspectives on social/technical primacy - this has been termed *DNA-AND-NDA*[54].

An analysis of commonalities between participants across tribes suggests the existence of syndicated interests which may be used to organise resources and focus efforts on promoting adoption of WO/W^3O. Observations of WO users in vivo have also suggested potential blocking issues (negative factors for adoption as discussed in Ch10) for WO that are referred to in the recommendations section and are predominantly non-architectural issues but do relate to overall design (especially Ux Design) and supporting eco-system resources.

[54] Indicating different causal arrangements/models of the DNA genes

11.6 Limitations

The limitations of this project fall into two broad area: limits of scope/capacity and limits of available data/claims to knowledge.

In all projects, there are limits to resources including time, budget, human/cognitive resources and space to report back. Hence some items that had been planned were not completed (e.g., [DataCo] beta service user interviews) and some that were completed could not be included due to space restrictions (secondary vignettes for the ABC interview chapters). Where new phenomena are under study it can be particularly difficult to obtain large data sets about a 'rare thing' without resorting to related items or proxies which therefore raises questions about claims to knowledge: are the claims valid for the 'rare thing' or only to the related items?

11.6.1 Limits of scope

The C-Tribe.

In the original conceptualisation, communities scale from "communities of one" (covering data on the Web in PDS' - personal data stores) through to communities of shared interest on to larger communities of governance including government. While interviews were conducted with academics on personal data stores, limited time/access prevented a more detailed exploration of personal data stores and personal WO solutions.

DataCo customer community

A series of interviews with [DataCo] users of [DataNames] had been foreseen and approved. Project timetables and access to customers prevented the completion of sub-set of interviews leaving a more production-oriented view of [DataNames] than a balanced production/consumption view.

WO-as-a-Social Machine.

While there is discussion in the community of the relatedness of WOs to Social Machines, there are two immediate problems to address.

- Taking Social Machines as the primary perspective on WOs presents the risk of analysing one new/undefined socio-technical system in terms of another, given Social Machine research is new/ongoing with the literature yet to achieve broadly agreed definitions.
- Few WOs are apparently yet equipped/instrumented to enable 'social' behaviour directly. While other social machines in, say, citizen science (Zooniverse, EyeWire) offer valuable

usage/interaction datasets, these are arguably not Web Observatories whereas similar collaboration/communication tools and other measures to instrument WOs and WO usage are, as yet not widely apparent. While Social Machines have not been ignored per se - they have not been placed as a central focus potentially missing insights/perspectives.

11.6.2 Limits of Data/Claims

WOs vs. WO Cousins.

Given the small set of systems explicitly identifying as WOs the research has drawn on insights from related and apparently similar 'observatory-like systems'. The inclusion/exclusion of specific examples is not a clear-cut process and risks either over-generalising through the inclusion of an overly broad a set of systems or ignoring key features through too narrow a definition. I submit that the inclusion of features and insights around data management, analytics, data sharing, collaboration and the motivations for sharing/engaging may be validly included in a characterisation of the space in which WO and W^3O are seeking to operate. I stop short however of offering statistical characterisations or predictive models based on proxy systems, neither of which, in my view, would be meaningful given the small sample size, a lack of a method to objectively include/exclude systems and no inter-rater reliability check on the selection process.

Creating/Testing W^3O interactions.

Interoperating WOs are a key objective yet only by substantially reducing the scope of interviews could I have employed action research techniques in order to build interacting Observatories. Without this, it is (in my view) too early to evaluate interacting WOs "in the wild" beyond fundamental catalogue sharing mechanisms though we can observe early (but valuable) work on discovery, licenses, ethics and demonstrators. As noted above, observing sharing/interaction through usage data, collaboration features or other 'instrumentation' is not yet possible. It has therefore been difficult to offer conclusions about sharing across WOs other than through proxy systems.

11.7 Contribution

This research has contributed:

- The first taxonomy of WOs comprising a faceted hierarchy of WO elements
- A flexible 'socio-technical' DNA model of WOs comprising a body of network patterns/processes and visual templates which may be useful in underpinning future business cases and designs for WOs. These are:
    - The WO **design** model allowing the representation of function elements and standards that enable interoperation
    - The WO value exchange **narrative** model allowing the representation of patterns/sequences of national values that are exchanged
    - The WO **agency** model allowing the representation of a broad agency/structure model giving rise to net behaviours that are embedded in roles/contexts
- A set of emergent roles which operate cross-tribe and are therefore important to an understanding of engagement and collaboration certainly for WOs but also potentially for a wider set of collaborative systems and social machines
- An early-stage methodological approach for WO analysis called **DNA-AND-NDA** which allows the representation of complex combinations of social and technical factors. This approach may be applicable to a wider study of social machines and an examination of significance/causal factors
- Important narratives captured users in different social groups (tribes) illustrating examples of (un)aligned thinking around WO will inform design thinking around WOs and similar systems
- Recommendations/observations arising from extensive user interviews and participant observation guiding the development of WO and the WO project.

These elements combine to support the goals of building new Web Observatories (WO) and adapting existing sources/systems leading to a global network of interoperating Web Observatories (W^3O).

The supplementary value for potential WO practitioners lies in that during the emergence of new disciplines and approaches there are potential overlaps, definition/scope problems and confusion to address. Without distinctiveness, it may be difficult for a new approach to attract focus, funding [55] and resources from leadership teams and potential users without which a project's

[55] The relatively low funding for WO compared to the VAO was covered in earlier chapters.

success may be undermined. Whilst WOs naturally share common features with other types of technical system, I argue that it not subsumed by them. It is not only sufficiently distinct to be considered in its own right but also offers affordances that make it a potential source of "remixed" data and insights that will be unique for participants in a global W^3O eco-system.

Three axes: the **D** (defining and delimiting), the (**A**) (exploring agency and ambition) and the 'performed' bridge (**N**) (examining narratives and notional exchange) form a final DNA model:

1. The classification/disambiguation of current systems to allow for more specific work on extending existing systems
2. The design of future systems against a reference model (that will be refined/improved over time) reducing time and cost to be compliant with participant systems
3. A framework for understanding the incentives, motivations and cultural/political aspects of WO reducing the time to understand the implied requirements/rules of "trading" with other WO participants.

Whilst the results/insights for this project are firmly grounded in a study of WOs, the highly fluid conceptualisation of WOs and the inherent merging of diverse groups, sources and systems within the W^3O concept may enable the principles, models and research methods developed here to achieve a wider adoption in the study of social systems on the Web and Web Science more broadly in the study/analysis of social machines, shared platforms in CSCW (Computer Supported Collaborative Work) and underpinning the study of socio-technical information systems.

11.8 Future Work

Flowing from the earlier section the following are particularly highlighted as areas for additional work on the examination of WOs and, in particular, W^3O.

Longitudinal / confirmatory work

WO is a young concept which, the research has shown, "flexes" based on the perspective of the WO user. Future (repeated periodic) work around the WO definition would need to work with that flexibility by:

- Evaluating additional Observatory-like systems against the current template to confirm accuracy/fit
- Revisiting the standard template definition of the WO taxonomy to observe if new technologies lead to new affordances

- Examining the ecosystem through the emergence (dominance?) of certain templates, hosted solutions both as open platforms and commercial offerings

Evaluating "DNA AND NDA" method with interdisciplinary teams

The ability to flexibly re-order DNA analytical perspectives to establish candidate causal factors according to differing socio-technical models could usefully be developed and tested through parallel projects or "hackathon" events. Groups may be invited to interpret the same DNA data from different causal perspectives or in which teams are invited to assemble the D's, N's and A's from different teams into a single model. The communication and collaboration requirements for WO systems and what could be learned from instrumenting/observing this collaboration would also be highlighted through such research. DNA-AND-NDA might feasibly be developed into a broader interdisciplinary Web Science approach or methodology.

Inter WO processing and the emergence of W^3O

The sharing of data and services between WOs has been outlined in the literature but not studied in vivo. The study of projects running across multiple WO instances and the practicalities of networking security and performance will be usefully showcased informing design processes and giving citable case studies to encourage further participation in research with and about WOs.

Models around proof and liability for synthetic/collaborative data

The practicalities of documenting and expressing permissions and obligations (POE) around datasets are a key challenge for WOs and work in this area must continue (Wilson et al., 2014). Systems and standards must be tested with commercial partners and other WOs if the ethical/legal basis of shared data and the liability around service provision are to be understood. Similarly, even where the licenses for individual items are established/understood, the impact of combining data into novel synthetic forms must be researched where the underlying components have differing (or even incompatible) licenses. This is an area vital to WO maturity and may require engaging with open standards and technologies such as Block Chain, ODRL and PROV.

Personal Data, Privacy and the Web Observatory

Studying the use of personal data systems through live examples has proven challenging. Corporate privacy solutions such as http://www.trustlayers.com and http://www.privitar.com are starting to appear around how companies handle personal data. Government programmes such as midata in the UK mandates businesses to make personal data available to individuals (who may, however, choose to ignore this service). These are primarily examples of privacy projects and not examples of individuals actively organising and marshalling their own data in so-called

Personal Data Systems (or personal WOs) of which there are few active examples outside academia (Southampton's INDX PDS and MIT's OpenPDS).

We are also seeing the introduction of what might be called the "Personal Web Observatory" or “Nano Observatory" based on ultra-low cost hardware such as Raspberry PI, Arduino and Micro:Bit which gives a more flexible ecosystem (though less ubiquitous coverage) than using the smartphone as a WO platform. It remains to be seen to what extent this physical implementation may constitute an Observatory per se or rather an Observatory source/sensor in an IoT sense.

11.9 Final Remarks

Web Science and approaches like the Web Observatory reflect the need to understand and engage with an increasingly pervasive and complex Web of data, processes and relationships that span all sectors of modern society. The vision of a global deployment of dedicated Web Observatories is not inevitable and the distinctive functionality/focus of WOs may yet be subsumed by other services, systems or providers. Even the latter outcome speaks, not to the failure of the WO concept but rather, to the broader appeal for such a capability amongst various groups/providers as something useful or even vital.

What remains clear is that the Web affects us and shapes us even as we shape and change the Web. As more and more aspects of our lives and interactions are made digital and are mediated by this global digital platform, the more the operation (what we do), the ethics (what is acceptable to do) and even the existence of the Web (via apps and embedded devices) become transparent. The distinction between ‘society’ and 'society-on-the-Web' begins to blur.

To some extent to understand what happens on the Web is to begin to understand ourselves and to do that we must ensure the willingness and capability to share data and insights ethically as we live *with* the Web, *on* the Web and *through* the Web.

Web Observatories and a global Web of Observatories serve that ideal and I believe that their time has come. I hope that academia, business and government will take up the call with sufficient resources and support to enable the broadest level of participation.

> "There is only one thing stronger than all the armies of the world: and that is an idea whose time has come."
>
> **Victor Hugo**

I.C.B. 2017

Bibliography

Abell, D.F., 1980. Defining the business, Prentice Hall.

Accomazzi, A. & Dave, R., 2011. Semantic Interlinking of Resources in the Virtual Observatory Era. arXiv.org, astro-ph.IM.

Adner, R., 2012. The Wide Lens, Portfolio Publishing.

Aguilar, F., 1967. Scanning the business environment. New York, Macmillan

Ajzen, I. & Fishbein, M., 1977. Attitude-behaviour relations: A theoretical analysis and review of empirical research. Psychological Bulletin; Psychological Bulletin.

Alderfer, C.P., 1968. An empirical test of a new theory of human needs. Organizational Behaviour and Human Performance, 4(2), pp.142–175.

Allen, W., 2010. Mutual Improvement, Or, A Scheme for the Self-adjustment of the Social Machine (1846). Kessinger Publishing.

Alter, S., 1978. Development patterns for decision support systems. MIS Quarterly, 2(3), p.33.

Altshuller, G., 1996. And suddenly the inventor appeared: TRIZ, the theory of inventive problem solving. Worcester, MA: Technical Innovation Centre, 1996. - 173 pp.

Anderson, C., 2010. The Long Tail. 2004, Wired Magazine.

Anderson, J. & Markides, C., 2007. Strategic innovation at the Base of the Economic Pyramid. Harvard Business Online.

Ashby, W.R., 2015. An Introduction to Cybernetics (1956) - Scholar's Choice Edition, reprinted (2015)

Assiter, A., 1984. Althusser and structuralism. British Journal of Sociology. London School of Economics. 35 (2): 272–296. doi:10.2307/590235. JSTOR 590235.

Ausubel, J.H., 2011. THE NEW INTERNATIONAL DEEP CARBON OBSERVATORY. 11th Gas Workshop Abstracts.

Barabasi, A.-L., 1999. Emergence of Scaling in Random Networks. Science (New York, N.Y.), 286(5439), pp.509–512.

Bibliography

Bateson, G., 1972. Steps to an Ecology of Mind - New York: Ballantine.

Bazeley, P., 2013. Qualitative Data Analysis. Sage Publications Ltd

Bazeley, P., 2007. Qualitative Data Analysis with NVivo. Qualitative Data Analysis with NVivo.

Beeston, G.P., Urrutia, M.L., Halcrow, C., Xioa, X., Liu, L., Wang, J., Kim, J., Park, K., 2014. Humour reactions in crisis: a proximal analysis of Chinese posts on sina weibo in reaction to the salt panic of march 2011. In New York, New York, USA: International World Wide Web Conferences Steering Committee Request Permissions, pp. 1043–1048.

Benkler, Y., 2006. The Wealth of Networks, Yale University Press.

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. Scientific American.

Berners-Lee, T., 1998a. A roadmap to the Semantic Web. Available at http://www.w3.org, History

Berners-Lee, T., 1989. Information Management: A Proposal. Available at http://www.w3.org, History.

Berners-Lee, T., 1998b. Semantic web road map. Available at http://wwwW3.org, History.

Berners-Lee, T., 2013. The Web Index Report 2013. webindex.org, pp.1–23.

Berners-Lee, T. & Fischetti, M., 1999. Weaving the Web, Texere Publishing.

Berners-Lee, T., Weitzner, D. J., Hall, W., O'Hara, K., Shadbolt, N. & Hendler, J., 2006a. 'A Framework for Web Science', Foundations and Trends in Web Science 1(1), 1–130.

Berners-Lee, T., Hall, W., Hendler, J.,Shadbolt, N., Weitzner, D., 2006b. Creating a science of the Web. Science (New York, N.Y.), 313(5788), pp.769–771.

Bower, J.L. & Christensen, C.M., 1995. Disruptive technologies: catching the wave. cbred.uwf.edu.

Bowker, G.C. & Star, S.L., 2000. Sorting Things Out, MIT Press.

Bowker, G.C., Baker, K., Millerand, F., Ribes, D., 2010. Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In J. Hunsinger, L. Klastrup, & M. Allen, eds. International Handbook of Internet Research. Dordrecht: Springer Netherlands, pp. 97–117.

Brin, S. & Page, L., 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW7: Proceedings of the seventh international conference on World Wide Web 7. Elsevier Science Publishers B. V.

Brown, IC., Hall, W. and Harris, L., 2013. “From Search to Observation” in Proceedings of the 22nd International Conference on World Wide Web Companion 2013, pp. 1317–1320. Rio de Janeiro, Brazil. May 2013.

Brown, IC., Hall, W. and Harris, L., 2014. “Towards a taxonomy for Web Observatories”. At WOW2014 Web Observatory Workshop, Seoul, S. Korea. Apr 2014.

Brown, IC., Hall, W. and Harris, L., 2015. “DNA: Towards a method for analysing Social Machines & Web Observatories”. At WOW2015 Web Observatory Workshop, Oxford, UK. Jun 2015.

Brown, IC., Hall, W. and Harris, L., 2015. “DNA: From Search to Observation revisited” in WebSci15: Proceedings of the ACM Web Science conference. Oxford, UK. Jun 2015

Brown, IC., Harris, L. and Hall, W., 2013. “Enabling Web 3.0 adoption in the digital economy” at Digital Economy (DE2013) Conference, Salford, UK. Nov 2013.

Brown, IC., Hall, W., 2015. “Web Observatories: Research Briefing 2015”. Published by Web Science Trust for US Air Force Office of Scientific research (AFOSR): Award No. FA9550-15-1-0020). Southampton, UK. Dec 2015.

Brown, IC., Hall, W., 2016. “Web Observatories: Research Briefing 2016”. Published by Web Science Trust for US Air Force Office of Scientific research (AFOSR): Award No. FA9550-15-1-0020). Southampton, UK. Dec 2016

Brown, IC., 2013. “W.O.L.F: The Web Observatory Linkage Framework” in Digital Economy Web Science Doctoral Training Centre (DTC) research exhibition. Southampton, UK. Sept 2013. Available at https://www.scribd.com/collections/4174578/Web-Science-Posters-2013.

Brown, IC., 2016. “Going Open: [DataNames]: a narrative analysis”. Internal consultancy report commissioned by [DataCo]. London, UK. Jan 2016.

Brown, J., 1985. An Introduction to the Uses of Facet Theory. In Facet Theory. New York, NY: Springer New York, pp. 17–57.

Bryant, A., 2002. Re-grounding grounded theory. Journal of Information Technology Theory

Bryman, A., 2012. Social Research Methods, Oxford University Press.

Budavári, T., Malik, T., Szalay, AS., 2003. SkyQuery -- A Prototype Distributed Query Web Service for the Virtual Observatory. Astronomical Data Analysis Software and Systems (ADASS) XIII, 295, p.31.

Bibliography

Budavári, T., Dobos, L. & Szalay, A., 2013. SkyQuery: Federating Astronomy Archives. Computing in Science & Engineering, (99), p.1.

Buneman, O.P., 2009. Curated databases. In ECDL'09: Proceedings of the 13th European conference on Research and advanced technology for digital libraries. Springer-Verlag.

Buneman, O.P., Choi, B., Fan, W., Hutchison, R., Mann, R., Viglas, SD., 2005. Vectorizing and querying large XML repositories. Proceedings 21st Annual conference (ICDE) on Data Engineering.

Burnap, P., Rana, O., Willams, M., Housely, W., Edwards, A., Morgan, J., Sloan, L., Conjero, J., 2014. COSMOS: Towards an integrated and scalable service for analysing social media on demand, International Journal of Parallel, Emergent and Distributed Systems.

Bush, V., 1945. As we may think. In The Atlantic Monthly 1945.

Canter, D., 1983. The potential of facet theory for applied social psychology. Quality & Quantity.

Carr, N., 2011. The shallows: What the Internet is doing to our brains, New York : W.W. Norton.

Cerf, Vinton G. and Robert E. Kahn, 1974. A Protocol for Packet Network Intercommunication, IEEE Transactions on Communications (COM-22), May, 1974, pp. 637-648.

Charmaz, K., 2014. Constructing Grounded Theory, SAGE.

Chomsky, N., 1957. Syntactic Structures. Mouton.

Christensen, C.M. & Rosenbloom, R.S., 1995. Explaining the attacker's advantage: Technological paradigms, organizational dynamics, and the value network. Research Policy, 24(2), pp.233–257.

Chua, T.S., Luan, H., Sun, M., Yang, S., 2012. NExT: NUS-Tsinghua Center for Extreme Search of User-Generated Content. Multi Media, IEEE, 19(3), pp.81–87.

Cohen, T., 2013. Average Is Over: Powering America Beyond the Age of the Great Stagnation, Dutton

Collan, M. & Tetard, F., 2007. Lazy User Theory of Solution Selection. Proceedings or the CELDA 2007 Conference. Algarve, Portugal, 7–9, December, 2007. pp. 273–278.

Contractor, N., 2009. The emergence of multidimensional networks. Journal of Computer-Mediated Communication.

Contractor, N.S. & Monge, P.R., 2003. Using multi-theoretical multi-level (mtml) models to study adversarial networks. Dynamic social network modelling and analysis.

Contractor, N.S., Wasserman, S. & Faust, K., 2006. Testing Multitheoretical, Multilevel Hypotheses About Organizational Networks: An Analytic Framework and Empirical Example. Academy of management review, 31(3), pp.681–703.

Corbin, J. & Strauss, A., 2007. Basics of Qualitative Research, SAGE.

Creswell, J.W., 2003. Research Design, SAGE Publications, Incorporated.

Demarest, M., "The Information Triad: A Model Of Past, Current And Future Information Technology Utilization In The Firm", DSSResources.COM, 06/29/2007.

Diakopoulos,N., Luther, K., Medynskiy, Y., Essa, I., The Evolution of Authorship in a Remix Society. In Proceedings of Hypertext and Hypermedia. Manchester, UK, September 2007

Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F., 2016. BotOrNot: A System to Evaluate Social Bots. arXiv.org 2016

Davis, F.D., Bagozzi, R.P. & Warshaw, P.R., 1989. User acceptance of computer technology: a comparison of two theoretical models. Management science, 35(8), pp.982–1003.

De Roure, D., Goble, C. & Stevens, R., 2007. Designing the myExperiment virtual research environment for the social sharing of workflows. e-Science and Grid Computing.

De Roure, D., Hooper, C., Meredith-Lobay, M., Page, K., Tarte, S., Cruickshank, D. and De Roure, C. 2013. Observing Social Machines Part 1: what to observe? SOCM2013: The Theory and Practice of Social Machines, Rio de Janeiro, Brazil, International World Wide Web Conferences Steering Committee, pp. 901-904.

De Roure, D., Jennings, N.R. & Shadbolt, N.R., 2005. The Semantic Grid: Past, Present, and Future. Proceedings of the IEEE, 93(3), pp.669–681.

Deleuze, G. & Guattari, F., 1983. Deleuze: Anti-Oedipus, trans - Google Scholar, Robert Hurley.

Denscombe, M., 2008. Communities of Practice: A Research Paradigm for the Mixed Methods Approach. Journal of Mixed Methods Research, 2(3), pp.270–283.

Difranzo, D., Erickson, J., Gloria, M., McGuiness, DL., 2014. Building Web Observatories for Health Web Science. In Proceedings of 6th Annual ACM Web Science Conference.

Difranzo, D., Erickson, J., Gloria, M., Luciano, J., McGuiness, DL., Hendler, J., 2014. The web observatory extension: facilitating web science collaboration through semantic markup. In Proceedings of 6th International World Wide Web Conferences.

Bibliography

Djorgovski, S.G. & Williams, R., 2005. Virtual Observatory: From Concept to Implementation. arXiv.org, astro-ph.

Donath, J., 2014. The Social Machine, MIT Press.

Eckert, P., 2006. Communities of Practice. (EK Brown, RE Asher, & J. MY Simpson, Eds.) Encyclopedia of language & linguistics)

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S., 2013. Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. Ann Arbor: Deep Blue. http://hdl.handle.net/2027.42/97552.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007). *Understanding Infrastructure: Dynamics, Tensions, and Design*. Ann Arbor: Deep Blue. http://hdl.handle.net/2027.42/49353.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Social Studies of Science, 41*(5), 667-690.

Ellul, J., 1954. The technological society. (Pylyshyn, Z., Bannon, L. Eds.) Perspectives on the computer revolution (1989). Ablex Publishing

Engelbart, D.C., 2001. Augmenting human intellect: a conceptual framework (1962), PACKER.

Evans., M., O'Hara, K., Tiropanis, T., Webber, C., 2013. Crime applications and Social Machines: crowdsourcing sensitive data. In WWW '13 Companion: Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, pp. 891–896.

Evenstad, S., 2011. An interpretative phenomenological Survey (IPA) of burnout among three ICT employees in Norway, Masters Thesis, Norwegian University of Science and Technology, Trondheim.

Feenberg, A., 1990. The critical theory of technology. Capitalism Nature Socialism, 1(5), pp.17–45

Feenberg, A., 2002. Transforming Technology: A Critical Theory Revisited. Oxford: Oxford University Press

Fichman, M., 1993. Science, Technology & Society: A Historical Perspective in Kirkpatrick, G., 2008. Technology and Social Power. Palgrave Macmillan (NY) p6

Gallen, C., 2013. Some Considerations for a Web Observatory. pp.1–6. WebSci '13 BWOW Workshop.

Gao, M., Singh, V.K. & Jain, R., 2012. Eventshop: from heterogeneous web streams to personalized situation detection and control. In WebSci '12: Proceedings of the 3rd Annual ACM Web Science Conference.

Gitlin, Todd. 1980. The Whole World Is Watching: Mass Media in the Making and Unmaking of the New Left. Berkeley, CA, Los Angeles, CA & London, U.K.: University of California Press.

Glaser, B.G. & Strauss, A.L., 1967. The Discovery of Grounded Theory. Strategies for Qualitative Research. Barney G. Glaser and Anselm L. Strauss. (1. Publ.) - Chicago: Aldine (1967). X, 271 S. 8°,

Gloria, M., Difranzo, D., Navarro, M., Hendler, J., 2013. The performativity of data: reconceptualizing the web of data, WebSci '13 Proceedings of the 5th Annual ACM Web Science Conference pp 109-117.

Gloria, M. & McGuinness, D.L., 2014. Building Web Observatories for Health Web Science. WebSci '14 Proceedings of the 6th Annual ACM Web Science Conference.

Goffman, E., 1959. The presentation of self in everyday life. Doubleday (NY)

Goffman, E., 1974. Frame Analysis: An Essay on the Organization of Experience. New York, NY: Harper & Row

Goldratt, E., & Cox, J., 1984 The Goal. Gower Publishing.

Grudin, J. 1990. The Computer Reaches Out: The historical continuity of user interface design. Paper presented at the Proceedings of CHI '90, ACM SIGCHI Conference, Seattle, Wash., USA.

Gubanov, M., Stonebraker, M., Bruckner, D., 2014. Text and structured data fusion in data tamer at scale. In 2014 IEEE 30th International Conference on Data Engineering (ICDE). IEEE, pp. 1258–1261.

Guttman, R. & Greenbaum, C.W., 1998. Facet theory: Its development and current status. European psychologist, 3(1), pp.13–36.

Hackett, P.M.W., 2014. Facet Theory and the Mapping Sentence, Palgrave Pivot.

Halcrow, C., 2014. Scaling and co-locating commonly used humour tags in Weibo. WWW 2014 Web Observatory Workshop.

Halford, S., Pope, C., Carr, L., 2010. A manifesto for Web Science. In Proceedings of WebSci10: Extending the Frontiers of Society On-Line, United States. 26 - 27 Apr 2010., pp. 1-6.

Hall, W. & Brown, I., 2015 (2016). Web Observatory: a report for the Airforce Office of Scientific Research. Award No. FA9550-15-1-0020

Hall, W., De Roure, D., Shadbolt, N., 2009. The evolution of the Web and implications for eResearch. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 367(1890), pp.991–1001.

Hall, W., & Tiropanis, T., 2012 Web evolution and Web Science [in special issue: The WEB we live in] Computer Networks, 56, (18), pp. 3859-3865

Hall, W., Tiropanis, T., Tinati, R., Booth, P., Gaskell, P. 2013. The Southampton University Web Observatory. In Workshop on Building Web Observatories (BWOW) at the International Web Science 13 Conference, pp. 1–4, 2013

Hall, W., Tiropanis, T., Tinati, R., Wang, X., Luczak-Rösch, M., & Simperl, E., 2014. The Web Science Observatory - The Challenges of Analytics over Distributed Linked Data Infrastructures ECRIM News, (96), pp. 29-30.

Hanisch, R.J., Berriman, G., Lazio, T. Bunn, S., 2015. The Virtual Astronomical Observatory: Re-engineering access to astronomical data. Astronomy and …, 11, pp.190–209.

Hazen, R.M., Hemley, R.J. & Bertka, C.M., 2010. The Deep Carbon Observatory: Unanswered Questions in Deep Carbon Science. 2010 GSA Denver Annual.

Healy, K., 2015. The Performativity of Networks. European Journal of Sociology, 56(02), pp.175–205. 2015

Hendler, J., 2013. Broad Data: Exploring the Emerging Web of Data. BIG DATA MARCH 2013 DOI: 10.1089/big.2013.1506

Hendler, J., 2011. The Semantic Web 10th year update. In WIMS '11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics

Article No. 1 Hendler, J., 2009. Web 3.0 Emerging. Computer, 42(1), pp.111–113.

Hendler, J. & Berners-Lee, T., 2010. From the Semantic Web to Social Machines: A research challenge for AI on the World Wide Web. Artificial Intelligence, 174(2), pp.156–161.

Hendler, J. & Golbeck, J., 2008. Metcalfe's Law, Web 2.0, and the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web archive Volume 6 Issue 1, February, 2008

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., Weitzner, D., 2008. Web science: an interdisciplinary approach to understanding the web. Commun. ACM (), 51(7), pp.60–69.

Hey, T. & Trefethen, A.E., 2002. The UK e-Science Core Programme and the Grid. Future Generation Computer Systems, 18(8), pp.1017–1031.

Holsapple, C.W., Whinston, A.B. & Benamati, J.H., 1996. Decision support systems. West Publishing Co.

James, P., van Seeters, P., 2014. Globalization and Politics, Vol. 2: Global Social Movements and Global Civil Society. London: Sage Publications.

Janis, I., 2008. Groupthink. IEEE Engineering Management Review, 36(1), pp.36–36.

Jones, P., Bradbury, L. & LeBoutillier, S., 2015. Introducing Social Theory, John Wiley & Sons.

Jones, M., Alony, I., 2011. Guiding the Use of Grounded Theory in Doctoral Studies – An Example from the Australian Film Industry IJDS , Volume 6.

Kahle, B., Prelinger, R. & Jackson, M.E., 2001. Public Access to Digital Material. D-Lib Magazine, 7(10).

Kahneman, D. & Tversky, A., 1984. Choices, values, and frames. American Psychologist, 39(4), pp.341–350.

Kauffman, S.A., 1993. The Origins of Order: Self-Organization and Selection in Evolution, Oxford University Press.

Keahey, K., 2012. Virtual Observatories: A Facility for Online Data Analysis. Available at: http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/keahey.pdf

Kirkpatrick, G., 2008. Technology and Social Power. Palgrave Macmillan (NY).

Kittler, F.A., 1999. Gramophone, Film, Typewriter, Stanford University Press.

Klein, K.J. & Sorra, J.S., 1996. The Challenge of Innovation Implementation. Academy of management review.

Kranzberg, M., 1986. Technology and History: "Kranzberg's Laws", Technology and Culture, Vol. 27, No. 3, pp. 544–560.

Krasner-Khait, B., 2001. Survivor: The History of the Library, History Magazine Oct/Nov 2001 (Moorshead, Quebec)

Kuutti, K., 1996. Activity Theory as a Potential Framework for Human Computer Interaction Research. In Nardi, B. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction. Cambridge, Massachusetts: The MIT Press.

Kwasnick, B., ”The Role of Classification in Knowledge Representation and Discovery." Available [Online]. https://www.ideals.illinois.edu/bitstream/handle/2142/8263/librarytrendsv48i1d_opt.pdf. [Accessed: 10-Jan-2014].

Latour, B., 2005. Reassembling the Social - An Introduction to Actor-Network-Theory. Oxford University Press.

Lessig, L., 2004. Free Culture. The Penguin Press, New York.

Lessig, L., 2008. Remix. The Penguin Press, New York.

Levine, R., 2009. The cluetrain manifesto, New York : Basic Books.

Levy, S. & Guttman, L., 1985. A Faceted Cross-Cultural Analysis of Some Core Social Values. In Facet Theory. New York, NY: Springer New York, pp. 205–221.

Licklider, J., 1968 The Computer as a communications device. In *Science and Technology,* April 1968.

Malone, T., Laubacher, R. & Dellarocas, C., 2009. Harnessing Crowds: Mapping the Genome of Collective Intelligence. pp.1–20. MIT Center for Collective Intelligence.

Malone, T.W., Crowston, K. & Herman, G.A., 2003. Organizing Business Knowledge: The MIT Process Handbook. Organizing Business Knowledge: The MIT Process Handbook.

Mandelbrot, B.B., 1977. Fractals : form, chance, and dimension, San Francisco : W. H. Freeman.

Mandelbrot, B.B., 1982. The fractal geometry of nature, San Francisco : W.H. Freeman.

Mann, R., Baxter, R., Carroll, R., Wen, Q., Buneman, P., Choi, B., Fan W., Hutchison, R., Viglas, S., 2003. Xml in the Virtual Observatory. Large Telescopes and Virtual Observatory: Visions for the Future, 8, p.37.

Martin, U. & Pease, A., 2013. Mathematical practice, crowdsourcing, and Social Machines. In CICM'13: Proceedings of the 2013 international conference on Intelligent Computer Mathematics. Springer-Verlag.

Markides, C.C., 1998. Strategic innovation in established companies. Sloan Management Review.

Markides, C. & Crainer, S., 2010. INNOVATING GLOBALLY. Business Strategy Review, 21(1), pp.24–27.

Maslow, A. & Herzeberg, A., 1954. Hierarchy of needs. AH Maslow.

Matthewman, S., 2011. Technology and Social Theory. Palgrave Macmillan

May, C.R., Mair, F., Finch, T., MacFarlaine, A., Dowrick, C., Treweek, S., Rapley, T., Ballini, L., Ong B., Rogers, A., Murray, E., Elwyn, G., Legare, F., Gunn, J., Montori, V., 2009. Development of a theory of implementation and integration: Normalization Process Theory. Implementation Science, 4(1), p.29.

McKelvey, K. & Menczer, F., 2013. Interoperability of Social Media Observatories. pp.1–3. WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web

McNiff, J., 2013. Action Research, Routledge.

McNiff, J. & Whitehead, J., 2011. All You Need to Know About Action Research, SAGE.

Meira, S., Buregio, V., Nascimento, L., Figueiredo, E., Neto, M., Encarnacao, B., Garcia, V., 2011. The Emerging Web of Social Machines. IEEE International Computer Software and Applications Conference. Proceedings, 24(17), pp.26–27.

Mitteroecker, P. & Huttegger, S.M. Biol Theory (2009) 4: 54. doi:10.1162/biot.2009.4.1.54

Monge, P.R. & Contractor, N.S., 2001. Emergence of communication networks. The new handbook of organizational communication.

Morshead,R. W., "Taxonomy of Educational Objectives Handbook II: Affective Domain," Studies in Philosophy and Education, vol. 4, no. 1, pp. 164–170, 1965.

Nickerson, R., Muntermann, J., Varshney, U., Isaac, H., 2009. Taxonomy Development in information systems: developing a taxonomy of mobile applications. European Conference in Information Systems.

Novak, J. & Canas, A., 2008. The Theory Underlying Concept Maps and How to Construct and Use Them. pp.1–36.

Nyquist, H., 1928. Abridgment of certain topics in telegraph transmission theory. Journal of the A.I.E.E., 47(3), pp.214–217.

Ogburn, William F. "Cultural Lag as Theory." 1957. Sociology & Social Research 41.3 (Jan. 1957): 167-174.

Ogden, C., Richards, I., 1924. The Meaning of Meaning. Philosophical Review_ 33:222.

O'Hara, K. and Brown, IC., 2014. "Responsible use of data" in HC245 written submissions to Commons Committee on Science & Technology. The Stationary Office, Nov 2014.

O'Hara, K., Contractor,N., Hall,W., Hendler,J., Shadbolt N., 2013. Web Science: Understanding the Emergence of Macro-Level Features on the World Wide Web Vol. 10 of Foundations and Trends in Web Science Now Publishers

O'Hara, K., Social Machine Politics Are Here to Stay., 2013. IEEE Internet Computing, 17(2), pp.87–90.

O'Hara, K., Sackley, A., Brown, IC., Tinati, R., Tiropanis, T. and Wang, X., 2014. "Security and Legitimacy in a Web Observatory." 2nd International workshop on Building Web Observatories (B-WOW). Bloomington, USA. Jun 2014.

Open Data Institute (2016) *Open enterprise: how three big businesses create value with open innovation*. London, UK. Cited as co-researcher. London, UK.

O'Reilly, T. 2005. What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Page, K.R. and De Roure, D. 2013. Trajectories through Social Machines. Building Web Observatories Workshop, ACM Web Science 2013, Paris, 2013.

Papadaki, E., Whitmarsh, A. & Walls, E., 2014. Some challenges for the web Observatory vision. In the 2014 ACM conference. New York, New York, USA: ACM Press, pp. 245–246.

Pariser, E., 2012. The Filter Bubble, Penguin UK.

Paukkeri, M., 2012. Learning a taxonomy from a set of text documents. Applied Soft Computing, 12(3), pp.1138–1148.

Pearson, J. & Shim, J.P., 1994. An empirical investigation into decision support systems capabilities: A proposed taxonomy. Information & Management, 27(1), pp.45–57.

Pinch, T.J. & Bijker, W.E., 1987. The Social Construction of Facts and Artifacts. MIT Press

Pope, C., 2014. Computers, Cyborgs, Webs and… medical sociology? preview.medicalsociologyonline.org. Accessed: April 2016

Pollock, N. & Williams, R., 2010. e-Infrastructures: How Do We Know and Understand Them? Strategic Ethnography and the Biography of Artefacts. Computer Supported Cooperative Work (CSCW), 19(6), pp.521–556.

Pongpaichet, S., Singh, V., Gao, M., Jain, R., 2013. EventShop: Recognizing Situations in Web Data Streams. pp.1–9. WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web Pages 1359-1368

Power, D.J., 2001. Supporting Decision-Makers: An Expanded Framework. pp.1–6.

Power, D.J. & Sharda, R., 2007. Model-driven decision support systems: Concepts and research directions. Decision Support Systems, 43(3), pp.1044–1061.

Price, S., Hall, W., Earl, G., Tiropanis, T., Tinati, R., Wang, X., Gandolfi, E., Gatewood, J., Boateng, R., Denemark, D., Groflin, A., Loader, B., Schmidt, M., Billings, M., Spanakis, G., Suleman H., Tsoi, K., Wessel, B., Xu, J., Birkin, M., 2017. Worldwide Universities Network (WUN) Web Observatory. Submitted to International World Wide Web Conferences Steering Committee.

Proudfoot, J., Klein, B., Barak, A., Carlberg, P., Cuijpers, P., Lange, A., Ritterband, L., Andersson, G., 2011. Establishing Guidelines for Executing and Reporting Internet Intervention Research. Cognitive Behaviour Therapy, 40(2), pp.82–97.

Quinn, P. & Hanisch, B., 2004. The International Virtual Observatory Alliance. Optimizing scientific return for astronomy through information technologies: 24-25 June, 2004, Glasgow, Scotland, United Kingdom, 5493, p.137.

Ranganathan, S. R., “Prolegomena to Library Classification,” 3rd Edition, Asia Publishing House, 1967.

Ram, S., 1987. A Model of Innovation Resistance by S. Ram. Advances in Consumer Research.

Ram, S. & Jung, H.S., 1994. Innovativeness in product usage: a comparison of early adopters and early majority. Psychology and Marketing.

Reid, K., Flowers, P. & Larkin, M., 2005. Exploring lived experience. Psychologist.

Reiss, S., 2004. Multifaceted Nature of Intrinsic Motivation: The Theory of 16 Basic Desires. Review of General Psychology, 8(3), pp.179–193.

Ritchey, T., 2012. Outline for a morphology of modelling methods. Acta Morphologica Generalis: On-Line Journal of the Swedish Morphological Society

Rogers, E., 1995. Diffusion of Innovations, 4th Edition, Simon & Schuster.

Roland-Campbell, A., 2014. The Web Observatory: A Social Machine to observe Social Machines. Available from: http://intersticia.com.au/wpcontent/uploads/2014/12/WebObservatory.pdf

Rousch, W., 2005. Social Machines | MIT Technology Review. Available at: http://www.technologyreview.com/featuredstory/404466/social-machines/.

Sackley, A. & Communities, H., 2014. Democratic futures: Crowdsourcing incident data. howardleague.org/local_justice_working_papers/. Accessed: April 2016.

de Saussure, F., 1986. Course in general linguistics (3rd ed.). (R. Harris, Trans.). Chicago: Open Court Publishing Company. (Original work published 1972). p. 9-10, 15.

Schumpeter, J.A., 1935. The Analysis of Economic Change. The Review of Economics and Statistics, 17(4), p.2.

Schwartz, B., 2009. The Paradox of Choice, Harper Collins.

Shadbolt, N.R., Smith, D., Simperl, E., Van Kleek, M., Yang, Y., Hall, W., 2013. Towards a classification framework for Social Machines. WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web, pp. 905-912

Shadbolt, N. & Berners-Lee, T., 2008. Web Science Emerges. Scientific American, 299(4), pp.76–81

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal, The, 27(3), pp.379–423.

Shao, C., Ciampaglia, G., Flammini, A., Menczer, F., 2016. Hoaxy: A Platform for Tracking Online Misinformation. arXiv.org, cs.SI, pp.745–750.

Shirkey, C., 2008. Here Comes Everybody. Penguin Press (NY)

Schneiderman, B., 2007. Web science: a provocative invitation to computer science. 50(6), pp.25–27. Available at: http://dl.acm.org/ft_gateway.cfm?id=1247022&type=html.

Shye, S., 1999. Facet theory. Encyclopedia of Statistical Sciences.

Simon, J., 2010. The entanglement of trust and knowledge on the Web. Ethics and Information Technology, 12(4).

Sismondo, S., 2011. An Introduction to Science and Technology Studies, John Wiley & Sons.

Smart, Paul R, Simperl, Elena and Shadbolt, Nigel (2014) A Taxonomic Framework for Social Machines. In, Miorandi, Daniele, Maltese, Vincenzo, Rovatsos, Michael, Nijholt, Anton and Stewart, James (eds.) Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society. Berlin, Germany, Springer.

Smith, J.A., 1996. "Beyond the divide between cognition and discourse: using interpretative phenomenological analysis in health psychology", Psychology and Health, Vol. 11 No. 2,pp. 261-271.

Smith, J.A., Flowers, P. and Larkin, M., 2009. Interpretative Phenomenological Analysis: Theory, Method and Research, Sage, Los Angeles, CA.

Smith, M.R. & Marx, L., 1994. Does Technology Drive History, MIT Press (MA).

Spaniol, M., Bencur, A., Viharos, Z., Weikum, G., 2012. Big Web Analytics: Toward a Virtual Web Observatory. Ercim News, 89, pp.23–24. Available at: http://pubman.mpdl.mpg.de

Spiteri, L., 1998. A Simplified Model for Facet Analysis. pp.1–30. Canadian Journal of Information and Library Science v23 (April-July 1998).

Sprague, R.H., Jr, 1980. A framework for the development of decision support systems. MIS Quarterly, 4(4), p.1.

Star, S.L. & Griesemer, J.R.,1989. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." Social Studies of Science, 19, 3, 387-420

Star, S.L., 1993. "Cooperation Without Consensus in Scientific Problem Solving: Dynamics of Closure in Open Systems". In Easterbrook, S. (Ed.), Computer-Supported Collaborative Work (CSCW): Cooperation or Conflict?, Springer-Verlag, UK,

Star, S.L. & Ruhleder, K., 1994. Steps Towards an Ecology of Infrastructure: Complex Problems in Design and Access for Large-Scale Collaborative Systems. CSCW, pp.253–264.

Sterman, J.D., 2000. Business dynamics: systems thinking and modelling for a complex world. McGraw-Hill

Stonebraker, M., Bruckner, D., Ilyas, I., Beskales, G., 2013. Data Curation at Scale: The Data Tamer System. CIDR.

Tarte, S.M., De Roure, D. & Willcox, P., 2014. Working out the plot: the role of stories in Social Machines, International World Wide Web Conferences Steering Committee

Thomas, G. & James, D., 2006. Reinventing grounded theory: some questions about theory, ground and discovery. British Educational Research Journal, 32(6), pp.767–795.

Tinati, R. & Carr, L., 2012. Understanding Social Machines. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom. IEEE, pp. 975–976.

Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N. and Hendler, J. 2013. The Web Science Observatory, IEEE Intelligent Systems 28(2) pp 100–104.

Tiropanis, T., Rowland-Campbell, A. & Hall, W. (2014) Government as a social machine in an ecosystem. At SOCM 2014: 2nd International Workshop on the Theory and Practice of Social Machines. 07 Apr 2014. , pp. 903-904

Tiropanis, T., Wang, X., Tinati, R. & Hall, W., 2014. Building a connected Web Observatory: architecture and challenges At 2nd International Workshop on Building Web Observatories (B-WOW14), ACM Web Science Conference 2014, United States. 23 - 26 Jun 2014. 10 pp.

Tiropanis, T., Hall, W., Hendler, J., deLarrinaga, C., 2014. The Web Observatory: A Middle Layer for Broad Data. 2(3), pp.129–133. Available at: http://online.liebertpub.com/doi/abs/10.1089/big.2014.0035.

Trochim, W.M.K. & Donnelly, J.P., 2008. Research Methods Knowledge Base

Tushman, M.L. & Anderson, P., 1986. Technological Discontinuities and Organizational Environments. Administrative science quarterly, 31(3), p.439.

Tversky, A. & Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science (New York, N.Y.), pp.453–45.

Van Kleek, M., Smith, D., Tinati, R., O'Hara, K., 2014. 7 billion home telescopes: observing Social Machines through personal data stores. In WWW Companion '14: Proceedings of the companion publication of the 23rd international conference on World wide web companion.

Van Kleek, Max, Smith, Daniel Alexander, Hall, Wendy and Shadbolt, Nigel R. (2013) "The Crowd Keeps Me in Shape": Social Psychology and the Present and Future of Health Social Machines At SOCM2013: The Theory and Practice of Social Machines, Brazil. , pp. 927-932

VanScoy, A. & Evenstad, S.B., 2015. Interpretative phenomenological analysis for LIS research. Journal of documentation, 71(2), pp.338–357.

VanScoy, A. 2013, Fully engaged practice and emotional connection: aspects of the practitioner perspective of reference and information service, Library & Information Science Research, Vol. 35 No. 4, pp. 272-278

Vickery, B. C. & Oficinas, A. Y., 1960. Faceted classification: a guide to construction and use of special schemes. Aslib, 1960.

Vygotsky, L., 1978. Mind in society. Harvard University Press

Walker, J., Taylor, J. & Carr, L., 2015. From public sector information catalogue to productive data: defining a national information infrastructure. At Building Web Observatories Workshop 2015, United Kingdom.

Wang, H., 2013. Semantically-enabled Knowledge Discovery in the Deep Carbon Observatory. AGU Fall Meeting

Wang, X., Tinati, R., Mayer, W., Rowland-Campbell, A., Tiropanis, T., Brown, IC., Hall, W., O'Hara, K., Stumptner, M., and Koronios, A., 2015. “Building a Web Observatory for the South Australian Government: Supporting an Age Friendly Population”. WebSci15: Proceedings of the ACM Web Science conference. Oxford, UK. Jun 2015

Weinberger, D., 2014. Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room. Basic Books, Inc. New York, NY, USA

Weiser, M., 1993. Some Computer Science Issues in Ubiquitous Computing. Communications of the ACM (July 1993).

Whitworth, B., & De Moor, A. (Eds.). (2009). Handbook of Research on Socio-Technical Design and Social Networking Systems. Hershey, PA: IGI.

Wiener, N., 1949. Cybernetics, Or Control and Communication in the Animal and the Machine. John Wiley & Sons Ltd.

Williams, R. & Edge, D., 1996. The social shaping of technology. Research Policy.

Woolley, A., Chabris, C., Pentland, A., Hashmi, N., Malone, T., 2010. Evidence for a collective intelligence factor in the performance of human groups. Science (New York, N.Y.), 330(6004), pp.686–688.

Yin, R.K., 2008. Case Study Research, 4th Ed. SAGE publications, Inc

Yip, M. & Webber, C., 2012. Structural analysis of online criminal social networks. In 2012 IEEE International Conference on Intelligence and Security Informatics (ISI 2012). IEEE, pp. 60–65.

Zwicky, F., 1957. Morphological Research and Invention. In Morphological Astronomy. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 6–30.

Zwicky, F., 1966. Discovery, Invention, Research Through the Morphological Approach. Toronto: Macmillan Company

Glossary of Terms

Big Data - A *relative* term giving perspective on the Volume, Velocity and Variety of a particular data stream in comparison to a traditional system's ability to handle these traits. Thus where sufficient capability/capacity exists "Big Data" may be considered to be "Data". May be contrasted with "broad data" from multiple sources/perspectives. Not to be confused with "a lot of data"

Boundary Object (Boundary Infrastructure) - (Collection of) Information artefacts(s), such as models, tools and applications, used in different ways by different communities. Boundary objects are plastic, interpreted differently across communities but with enough immutable content to maintain integrity.

Boxology - A boxology is a representation of an organized structure as a graph of labelled nodes ("boxes") and connections between them (as lines or arrows). The concept is useful because many problems in systems design are reducible to modular "black boxes" and connections or flow channels between them.

CERN – Particle physics laboratory hosting a large computing facility, which is primarily used to store and analyse data from experiments, as well as simulate events. Researchers need remote access to these facilities, so the lab has historically been a major wide area network hub. CERN is also the birthplace of the World Wide Web

CKAN - Comprehensive Knowledge Archive Network. Open source data platform portal from Open Knowledge Foundation

Dark Data (original)– typically a business term indicating internally held corporate data that is not leveraged for business benefit

Dark Data (proposed)– a broader Web term indicating data from decisions/actions a may be taken without the benefit of provenance, quality, context or ethical use.

DBpedia (from "DB" for "database") is a project aiming to extract structured content from the information created in the Wikipedia project.

Frame (after Goffman) - Typically a perspective or lens through which personal experiences are organized giving structure to individual perception of the events of experience.

Gisting - looking for the main idea or most important point in a written or spoken text

Glossary

GWAP – Luis von Ahn first proposed the idea of "human algorithm games", or games-with-a-purpose (GWAPs), in order to harness human time and energy for addressing problems that computers cannot yet tackle on their own.

Halo effect - a cognitive bias in which an observer's overall impression of a person, company, brand, or product influences the observer's feelings and thoughts about that entity's character or properties

HTML – Hypertext Markup Language. Standard for formatting Web pages.

IoT - The **Internet of things** is the inter-networking of physical devices, vehicles, buildings, and other items—embedded with electronics, software, sensors, actuators, and network connectivity that enable these objects to collect and exchange data

Internet Archive provides free public access to collections of digitized materials, including websites, movies/videos, moving images, and nearly three million public-domain books

iSurvey – Southampton university web-based survey platform

JSON – Java script object notation. Structured format for encoding data.

Knowledge Graph (e.g. Google Knowledge Graph) a technology used to store complex structured and unstructured information used by a computer systems and particularly for disambiguation of terms in information (Web) searching.

Laminations – multiple layers comprising a (Goffman) frame

LoC – Library of Congress. Classification format

Metcalfe's Law - states that the value of a telecommunications network is proportional to the square of the number of connected users of the system (n^2).

Morphospace - A representation of all the structure/arrangement of features that a (biological) system/entity might achieve

nVivo – Commerical qualitative research platform for analysis (esp coding) of documents/interviews

Observation - (In the sense of “monitoring”). Check the progress or quality of (something) over a period of time; keep under systematic review: (Source: Oxford English Dictionary)

Observatory - A room or building housing an astronomical telescope *or other scientific equipment* for the study of natural phenomena. (Source: Oxford English Dictionary)

ODUG – Open Data User Group

Point solution - Solving one particular problem without regard to related issues. Point solutions are widely used to fix a problem or implement a new service quickly

Schema- In psychology and cognitive science, a schema describes an organized pattern of thought or behaviour that organizes categories of information and the relationships among them. Such schemas underpin what we believe to be fundamentally "true" about concepts.

Skunkworks – A project developed by a small and loosely structured group of people who research and develop a project primarily for the sake of radical innovation.

Social Graph - In the Web context this is a graph that depicts a model or representation of a social network, where the word graph has been taken from graph theory.

Social Machine - In contrast to a deterministic Turing machine and comprising both human and technical components. Typically, the expression of a solution/reaction to social process via the combination of human resource and medium of distributed technologies such as the Internet, mobile devices and the Web.

Socrata - Commercial open data portal platform

Southbeach - Software package implementing Triz notation

SWOT – Strengths, Weaknesses, Opportunities and Threats. Analytical method

Syndicate - In the context of this project, the idea of a syndicate is a grouping of individuals or groups of groups who behave with a level of focus (not exclusive) on a particular perspective on the WO (specifically the around the data/content, around the system/design or around the utility/innovation). Syndicates are not mutually exclusive with respect to other syndicates nor with Tribes.

Tribe - In the context of this project, the idea of tribe is a grouping of individuals or grouping of groups (Actors) who act for a specific reasons (this may be do with utility/preferences, reward/punishment) typically within an occupational context. Tribal membership will generally be mutually exclusive with other tribes.

Triz - Russian analytical methodology and notation (after Altshuller).

VAO Virtual Astronomical Observatory. International project to collect and share astronomical data which inspired the later Web Observatory

Web Observatory (WO) One of potentially many *local* software/hardware systems controlled and operated by a named group comprising methods, tools, processes and technologies operated under an Observatory paradigm for the study of Web phenomena (esp. Social Machines).

(World Wide) Web Observatory (W^3O) or "Web of Observatories" - A singular, emergent, decentralised (global) service arising from the sharing of data and interoperation of services between sub-sets of individual regional interoperating WOs.

Web Science – An emerging interdisciplinary field concerned with the study of large-scale socio-technical systems, such as the World Wide Web (Wikipedia).

Web Observation - Use of a system to consider, capture, interpret and analyse data (typically *via* and/or *about* the Web) and *not* referring to a more general notion of "looking at things" or attempting a philosophical or ontological definition of visual perception.

Appendices

1. Appendix A – Contains a guide to Southbeach (TRIZ) notation
2. Appendix B – Contains analysis of survey material
3. Appendix C – Contains the WO ‘straw man’ model
4. Appendix D – Contains the D, N and A facets comprising the DNA Taxonomy
5. Appendix E – Contains extracts/analysis from IPA interviews

Excluded Materials

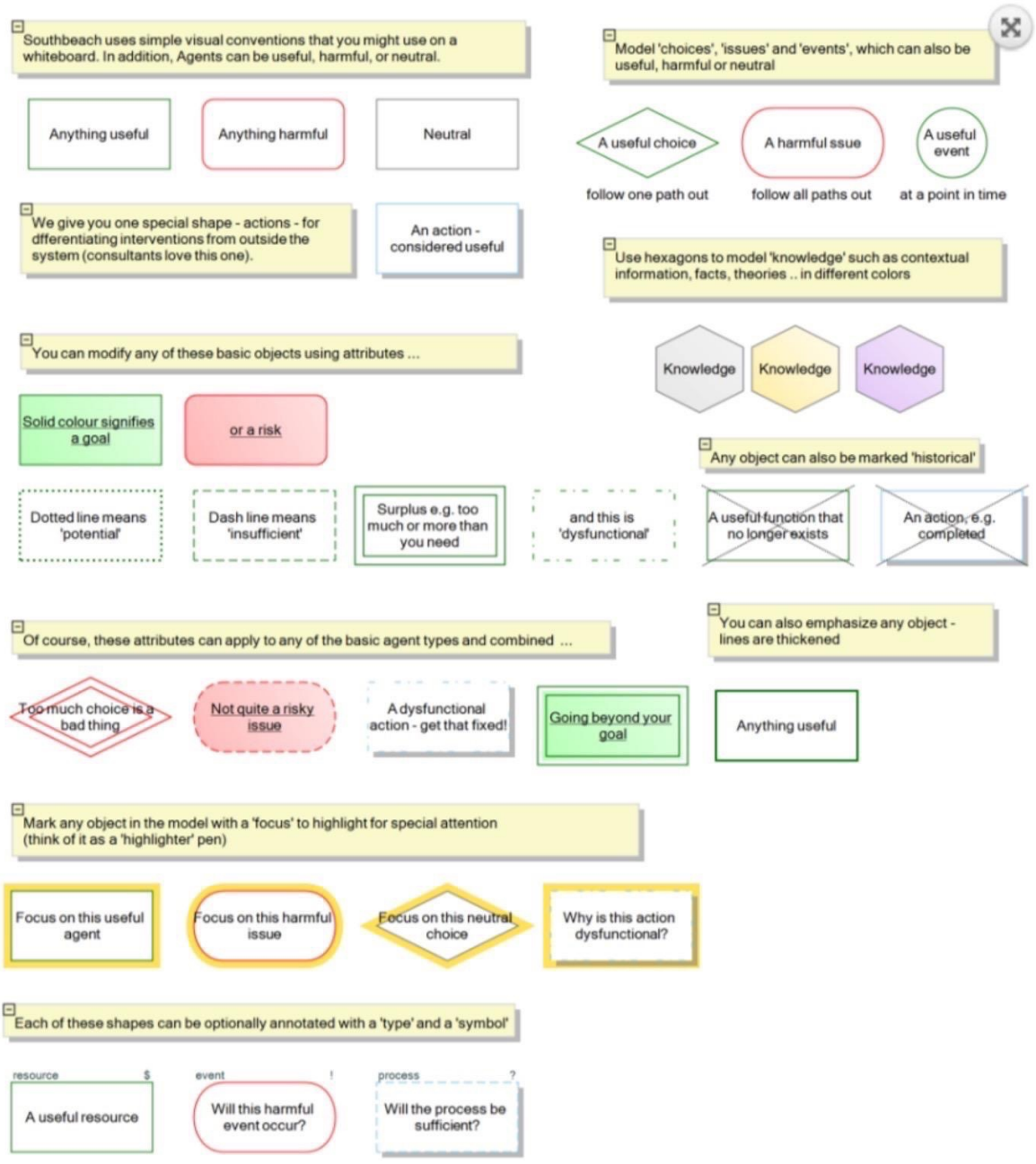
The following are excluded for reasons of confidentiality

1. Full interview transcripts
2. Interview recordings
3. IPA analysis tables

Appendix A TRIZ (Southbeach) Notation

A full document for the Southbeach implementation of TRIZ can be found at http://www.bptrends.com/

The following figures are taken from the southbeachinc.com quick guide and are reproduced with kind permission of Southbeach Inc. (15/3/17)



Appendix A


A Southbeach model consists of agents and effects
Burning fossil fuels
produces
Release of carbon dioxide
Green=useful, red=harmful, grey=neutral
Big fast cars are harmful - I am an environmentalist
Big fast cars are useful - I am a Top Gear fan
Big fast cars
from some perspective
produces
produces
Speed
Pollution
Agents are decomposed into their useful and harmful effects
Economic meltdown
produces
Re-examination of required regulatroy framework
counteracts
Future investment earnings growth rates
creates
produces
Business failure
Variations add additional meaning
Highlighter used for 'focus'
Potentially useful - not yet acting
Goal
Insufficient harm
No longer a risk
A surplus of whatever this is
useful external action/intervention
Dysfunction
Agents are actors in the system - some agents make choices or force issues/questions
When we do this
Should we try to avoid the choice?
produces
counteracts
Choice
We must face a harmful choice
produces
produces
Can we prevent the side effect?
prevents
This path is insufficient to meet our goals
This path is sufficient, but has a nasty side effect
produces
bad side effect
produces
produces
produces
Event
Before we can proceed - this issue must be resolved
produces
Goal
Agents can be 'tagged' to add semantic nuance
process
Manufacturing
creates
resource
Waste
uses
trend
Green Agenda
creates
service
Re-cycling

To illustrate the different effects that Southbeach provides, examples in this quickguide have been taken from the 'Green' agen

Appendix B Survey Findings

Findings by Group

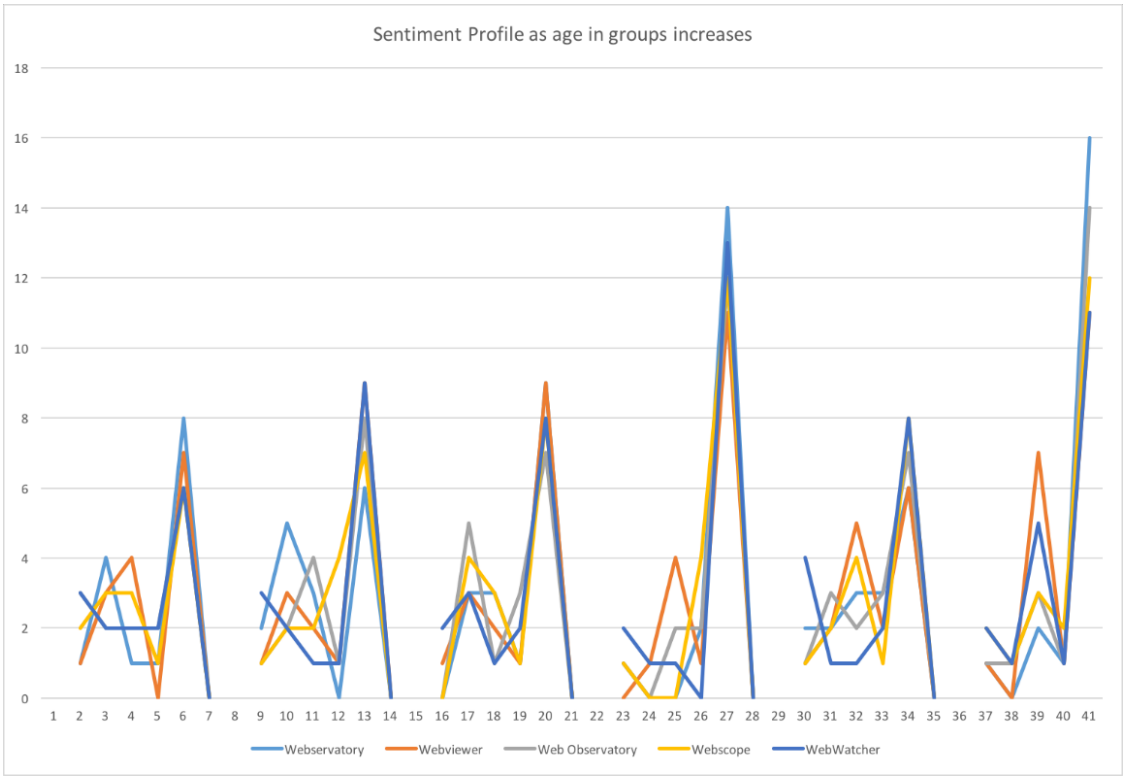
The following shows the results by Age-group depicting

1. Recognition (Don't knows/Non-answers vs. Answers)

2. Sentiment (Rated from very positive to very negative)

3. Rating the terms in order.

Overall Findings



Overall profiles by Age

Appendix B

Please rate how you feel about the following names

Answer	Very Negative	Quite Negative	Neutral	Quite Positive	Very Positive
Webservatory	59 (59.00%)	8 (8.00%)	12 (12.00%)	14 (14.00%)	7 (7.00%)
WebViewer	53 (53.00%)	6 (6.00%)	24 (24.00%)	12 (12.00%)	5 (5.00%)
Web Observatory	54 (54.00%)	12 (12.00%)	14 (14.00%)	13 (13.00%)	7 (7.00%)
WebScope	53 (53.00%)	13 (13.00%)	15 (15.00%)	12 (12.00%)	7 (7.00%)
WebWatcher	55 (55.00%)	8 (8.00%)	11 (11.00%)	10 (10.00%)	16 (16.00%)

Total unique respondents 100

Overall Rankings

Sentiment by Age

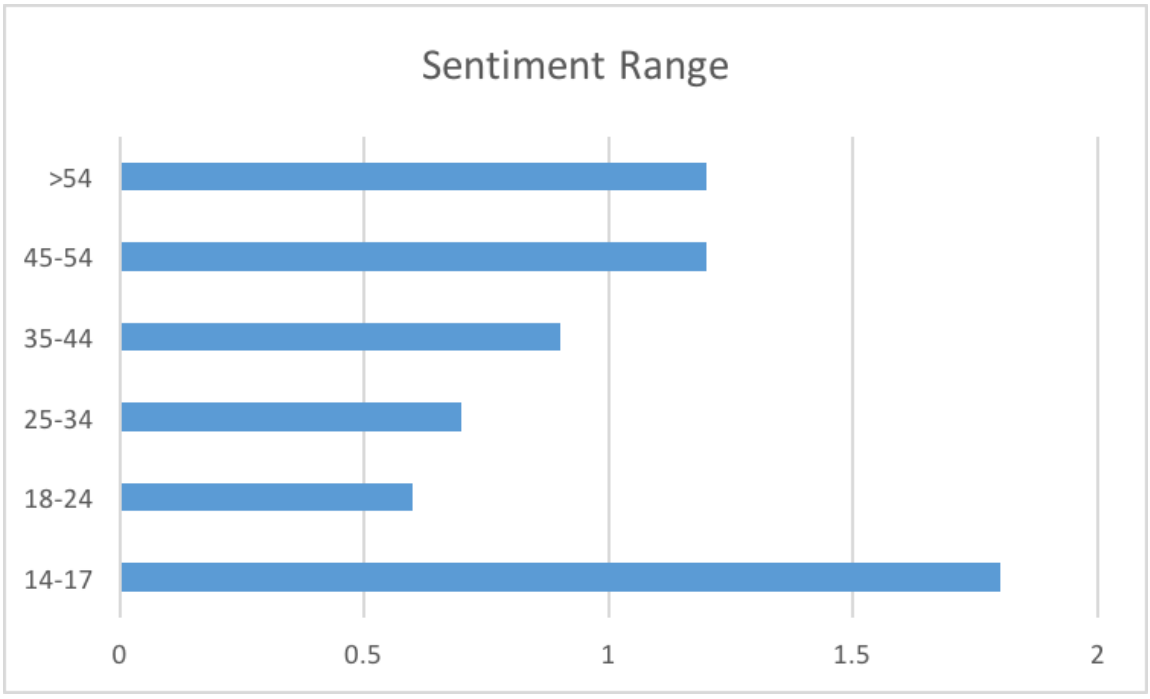
Please rank the names from your most positive impression at the top to your most negative impression at the bottom

Mean indicates the average ranking each item received. Because "1" is the highest ranking, the item with the lowest mean is the one that was ranked most highly.

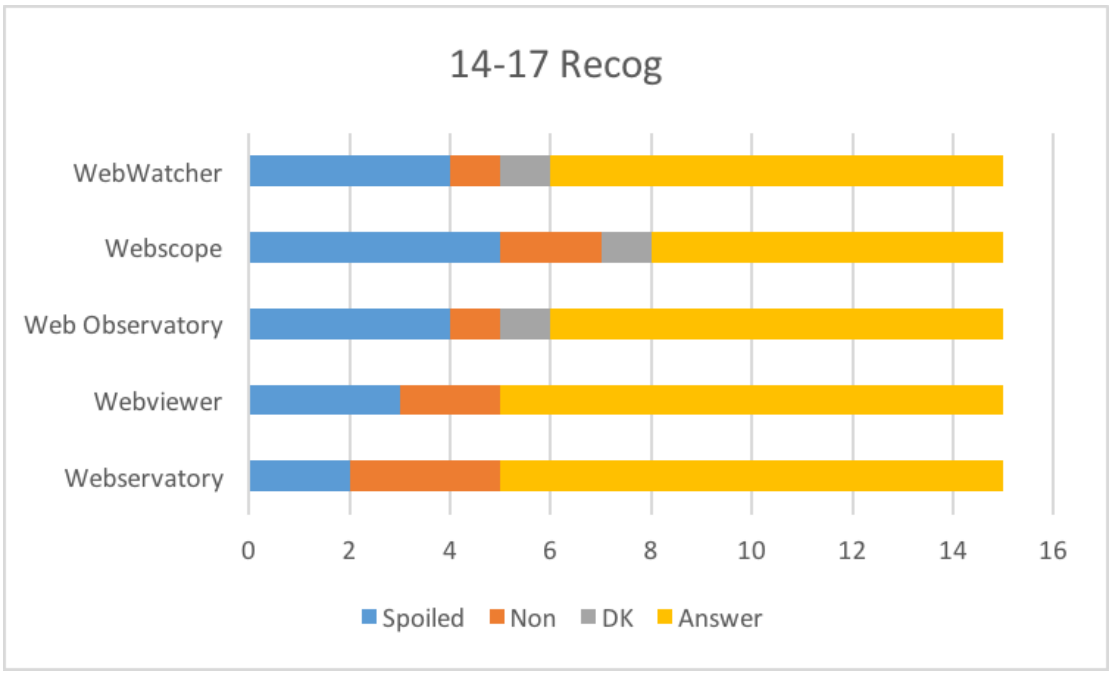
#	Answer	1	2	3	4	5	Mean
1	WebViewer	23	24	13	31	9	2.8
2	Webservatory	24	17	23	16	20	2.9
3	Web Observatory	11	22	32	21	14	3.1
4	WebScope	16	21	21	24	18	3.1
5	WebWatcher	26	16	11	8	39	3.2

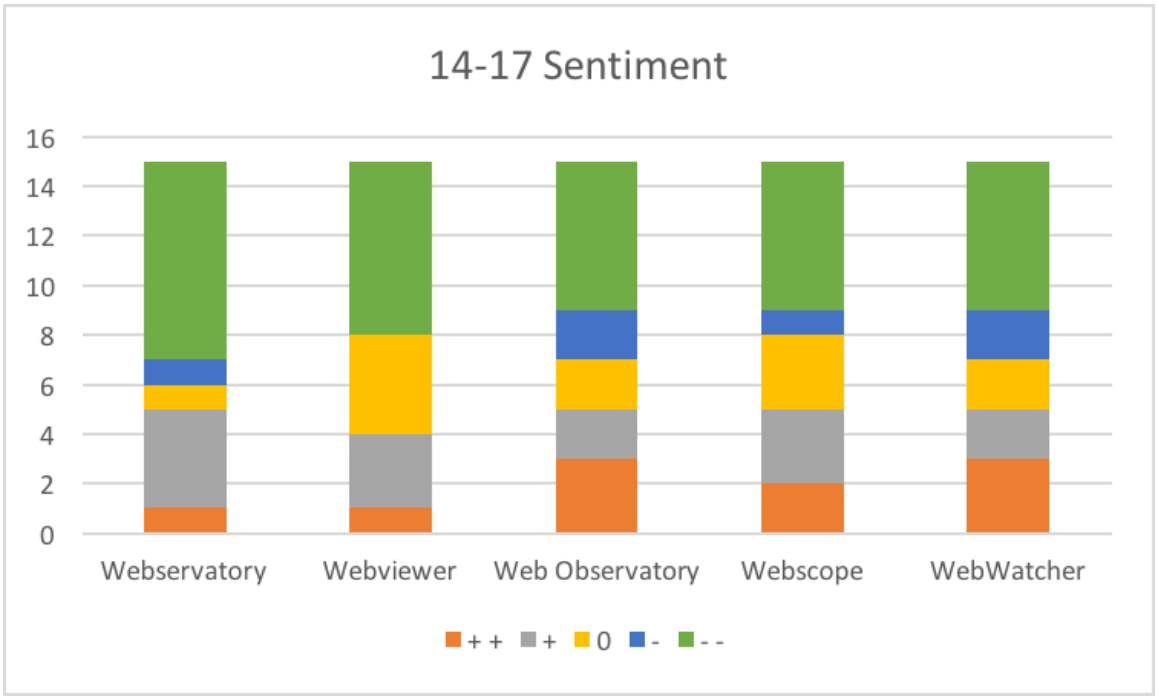
Total unique respondents 100

Overall sentiment interval = 0.4



Overall Sentiment interval by group





Please rank the names from your most positive impression at the top to your most negative impression at the bottom

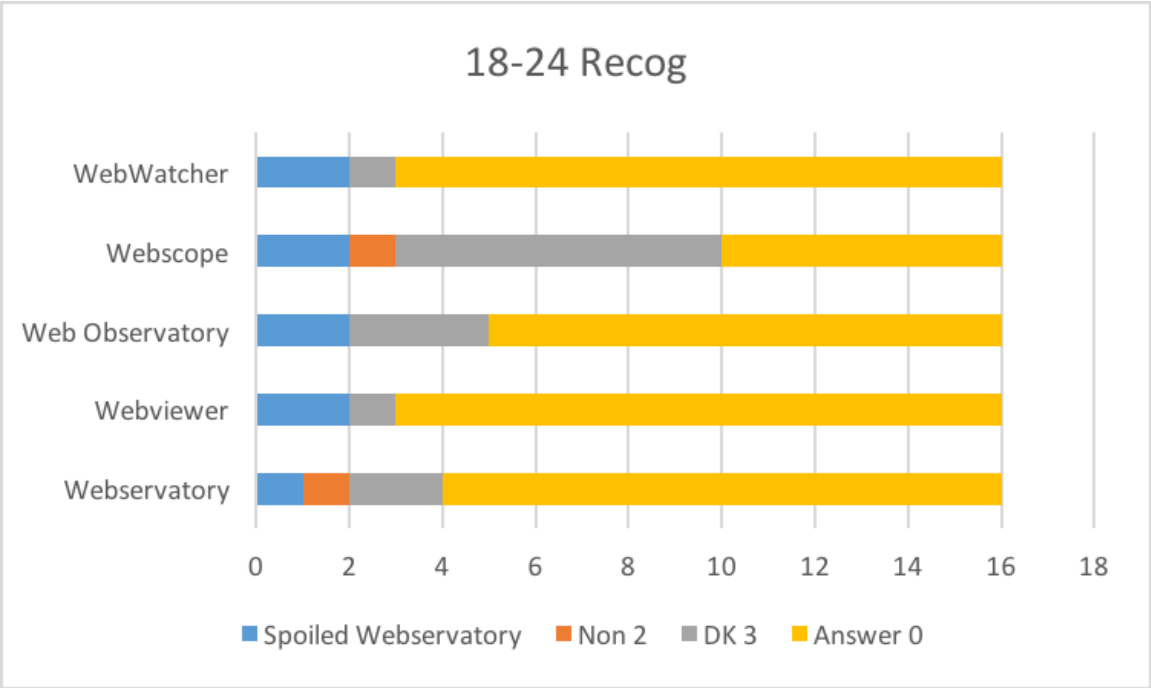
Mean indicates the average ranking each item received. Because "1" is the highest ranking, the item with the lowest mean is the one that was ranked most highly.

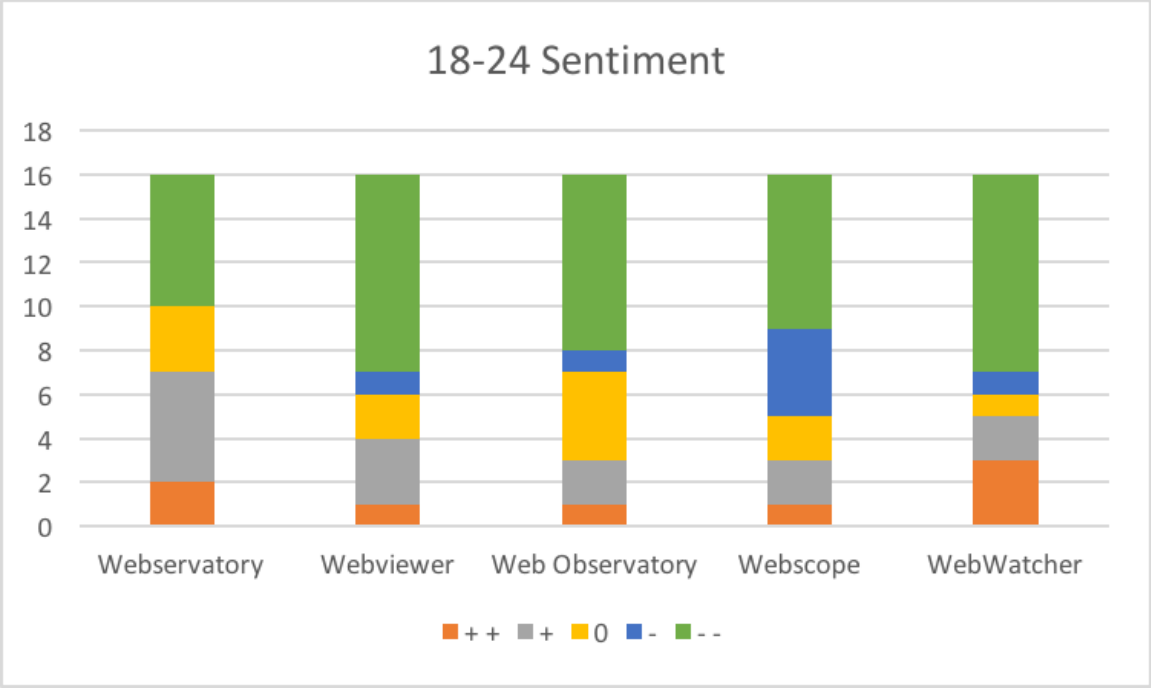
#	Answer	1	2	3	4	5	Mean
1	Webservatory	8	3	3	0	1	1.9
2	WebViewer	2	6	1	5	1	2.8
3	Web Observatory	0	3	8	4	0	3.1
4	WebScope	1	3	2	5	4	3.5
5	WebWatcher	4	0	1	1	9	3.7

Total unique respondents 15

Rating interval = 1.8

Appendix B





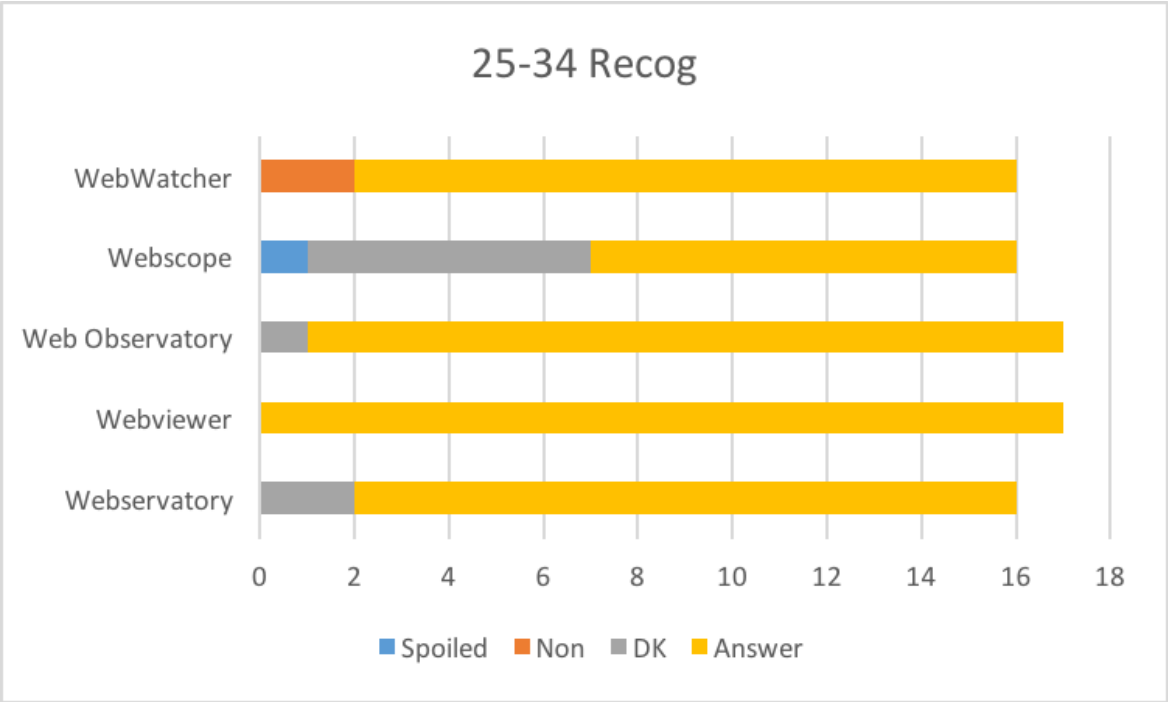
Please rank the names from your most positive impression at the top to your most negative impression at the bottom

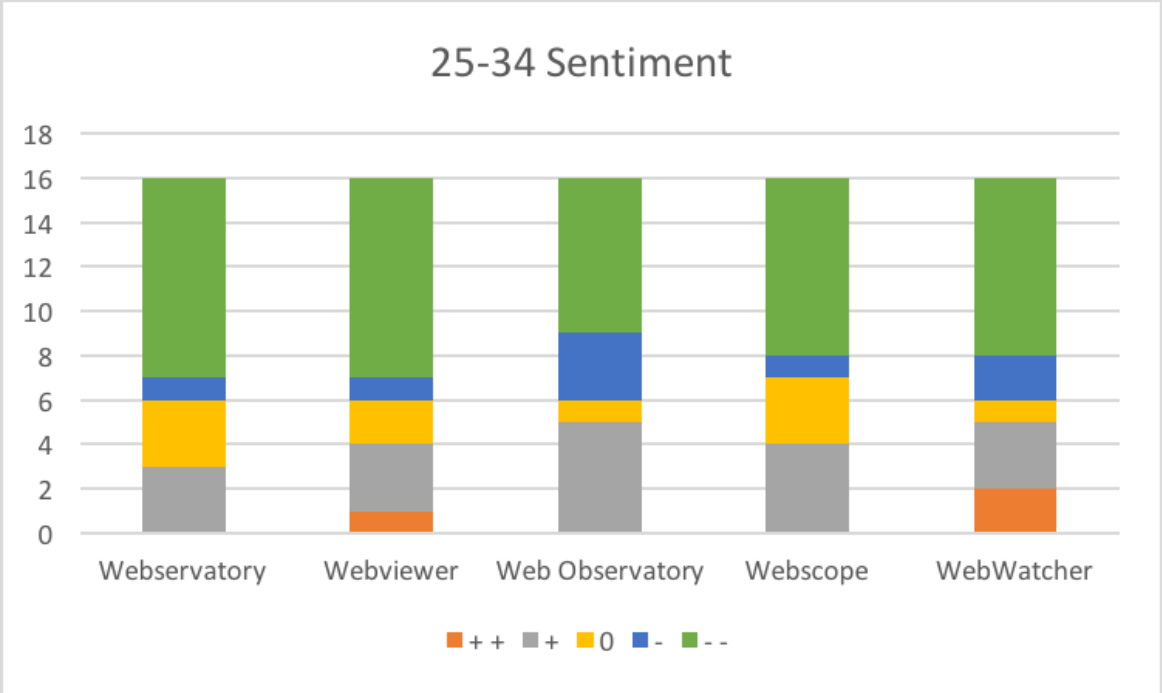
Mean indicates the average ranking each item received. Because "1" is the highest ranking, the item with the lowest mean is the one that was ranked most highly.

#	Answer	1	2	3	4	5	Mean
1	WebScope	5	3	3	2	3	2.7
2	Webservatory	4	2	5	2	3	2.9
3	Web Observatory	3	4	2	4	3	3.0
4	WebViewer	2	2	5	6	1	3.1
5	WebWatcher	2	5	1	2	6	3.3

Total unique respondents 16

Rating interval = 0.6





Please rank the names from your most positive impression at the top to your most negative impression at the bottom

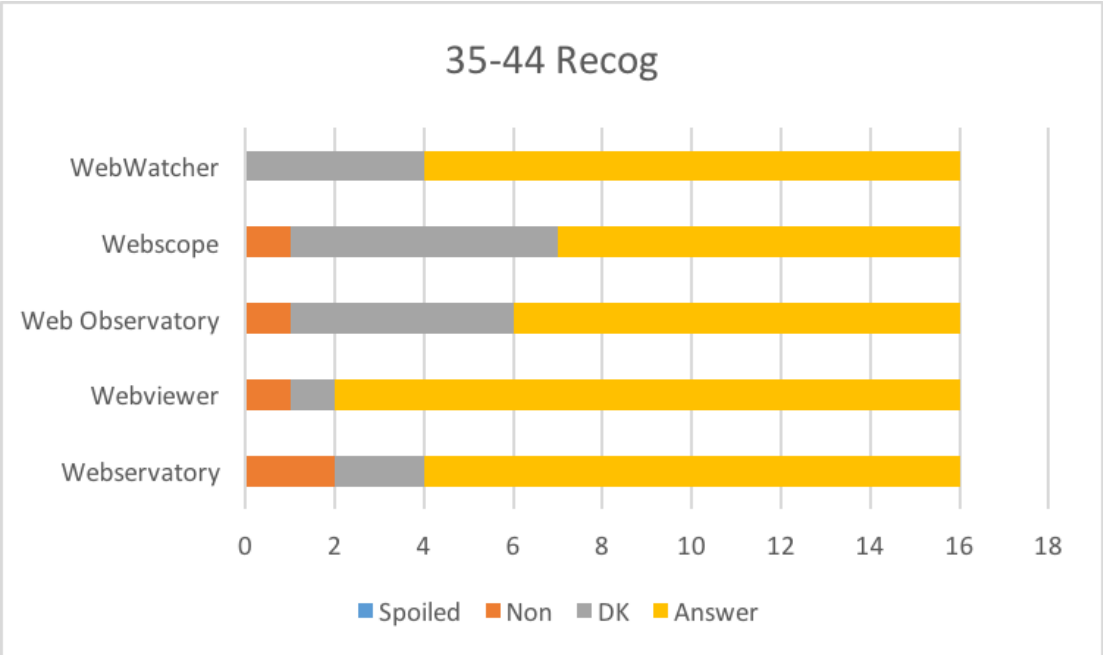
Mean indicates the average ranking each item received. Because "1" is the highest ranking, the item with the lowest mean is the one that was ranked most highly.

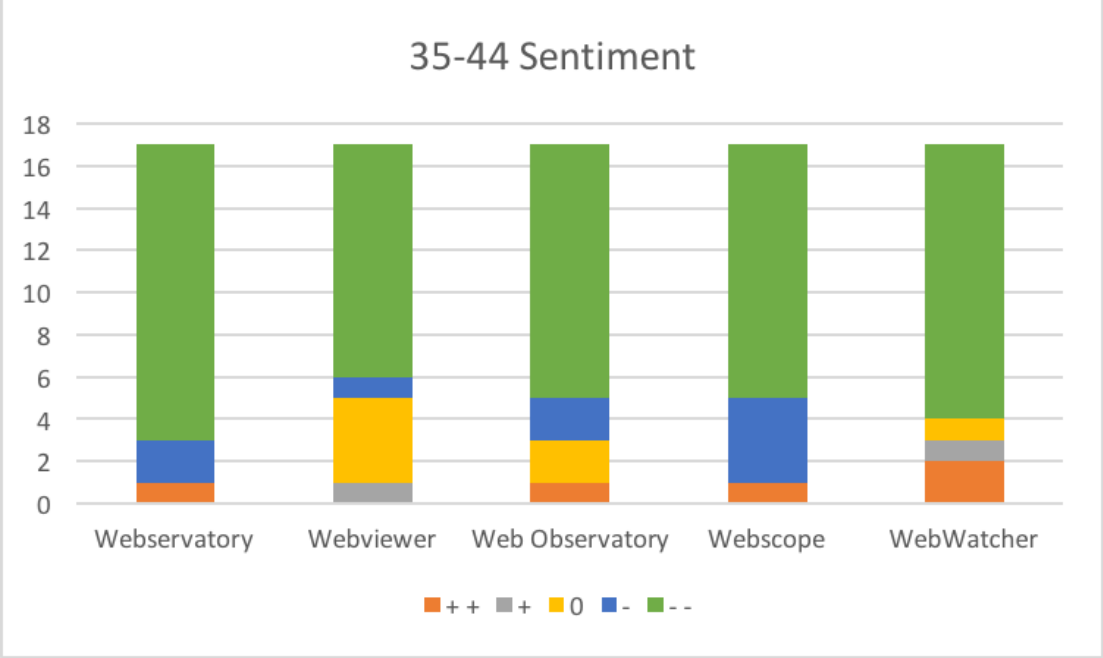
#	Answer	1	2	3	4	5	Mean
1	Webservatory	3	5	3	4	1	2.7
2	WebViewer	3	5	2	6	0	2.7
3	Web Observatory	4	2	5	2	3	2.9
4	WebWatcher	5	0	3	1	7	3.3
5	WebScope	1	4	3	3	5	3.4

Total unique respondents 16

Rating interval = 0.7

Appendix B





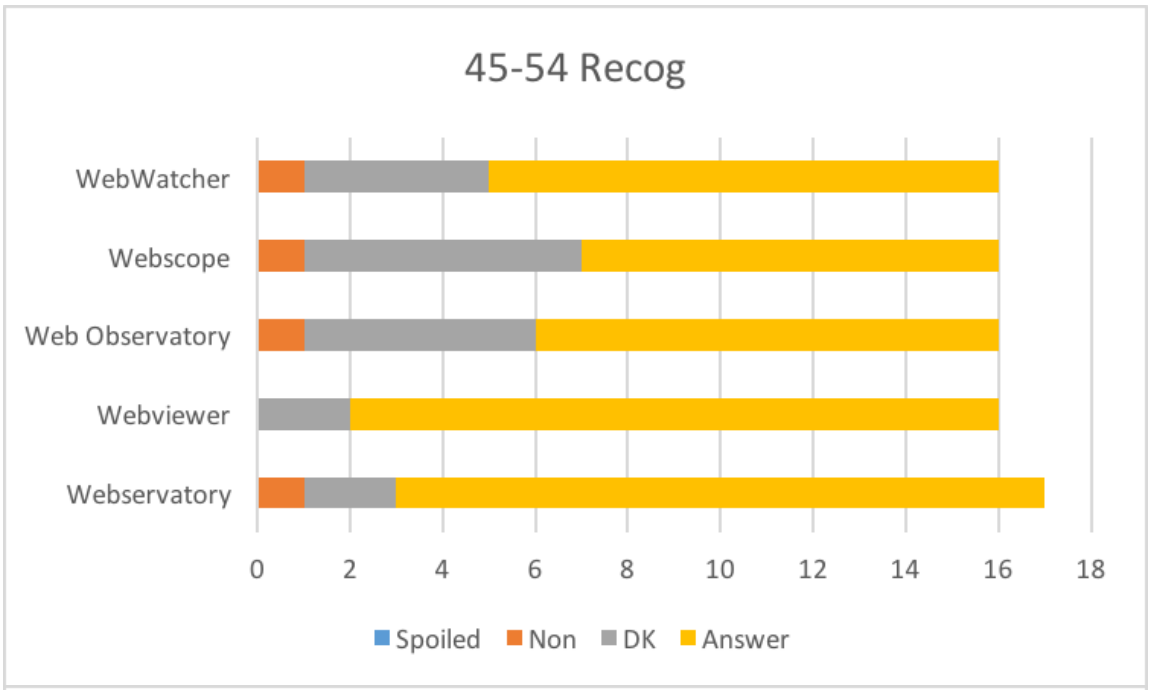
Please rank the names from your most positive impression at the top to your most negative impression at the bottom

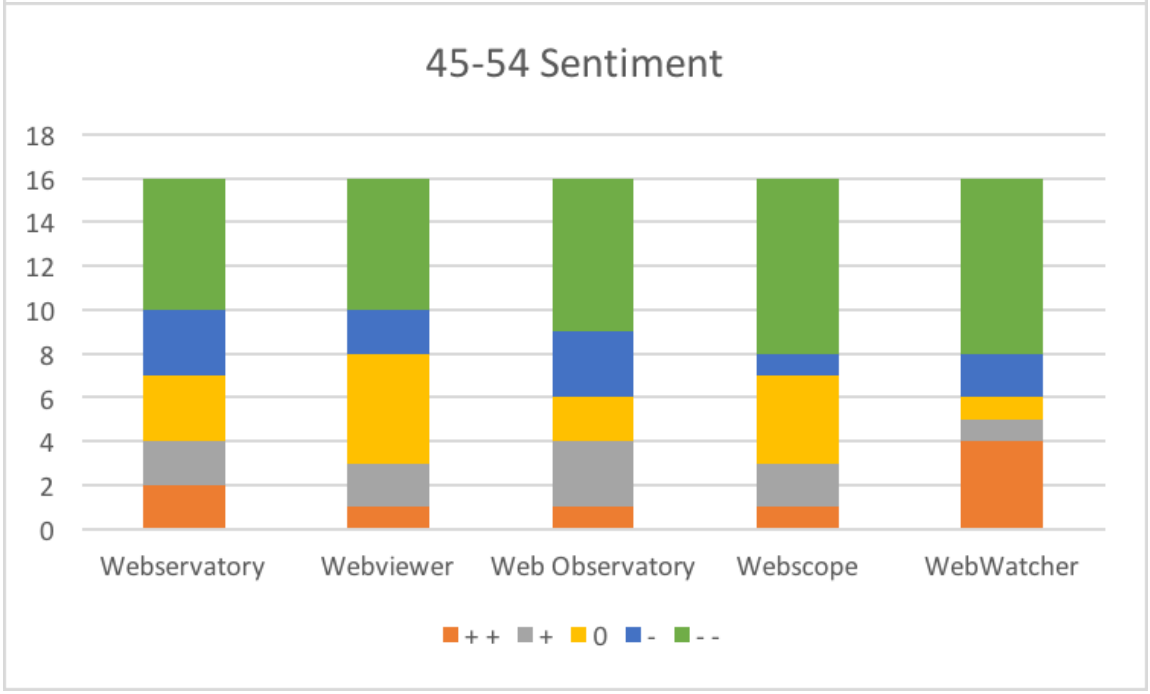
Mean indicates the average ranking each item received. Because "1" is the highest ranking, the item with the lowest mean is the one that was ranked most highly.

#	Answer	1	2	3	4	5	Mean
1	Web Observatory	2	7	5	2	1	2.6
2	WebScope	3	4	5	4	1	2.8
3	WebViewer	6	3	0	3	5	2.9
4	WebWatcher	3	3	3	2	6	3.3
5	Webservatory	3	0	4	6	4	3.5

Total unique respondents 17

Rating interval = 0.9





Please rank the names from your most positive impression at the top to your most negative impression at the bottom

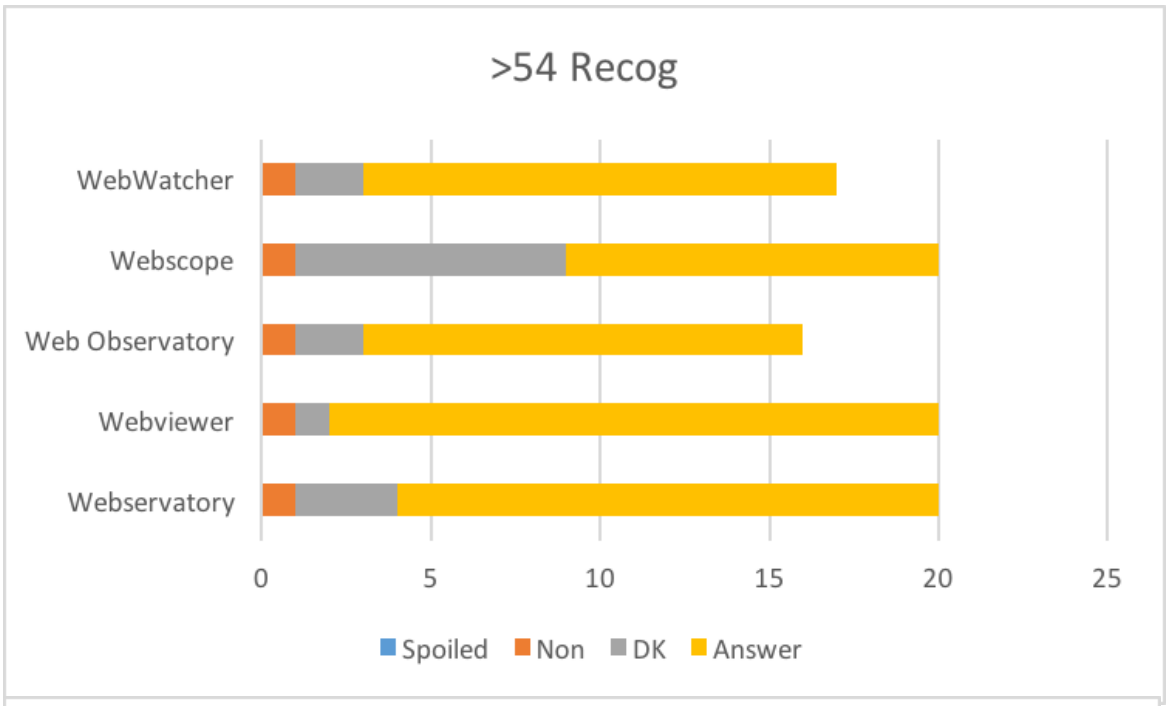
Mean indicates the average ranking each item received. Because "1" is the highest ranking, the item with the lowest mean is the one that was ranked most highly.

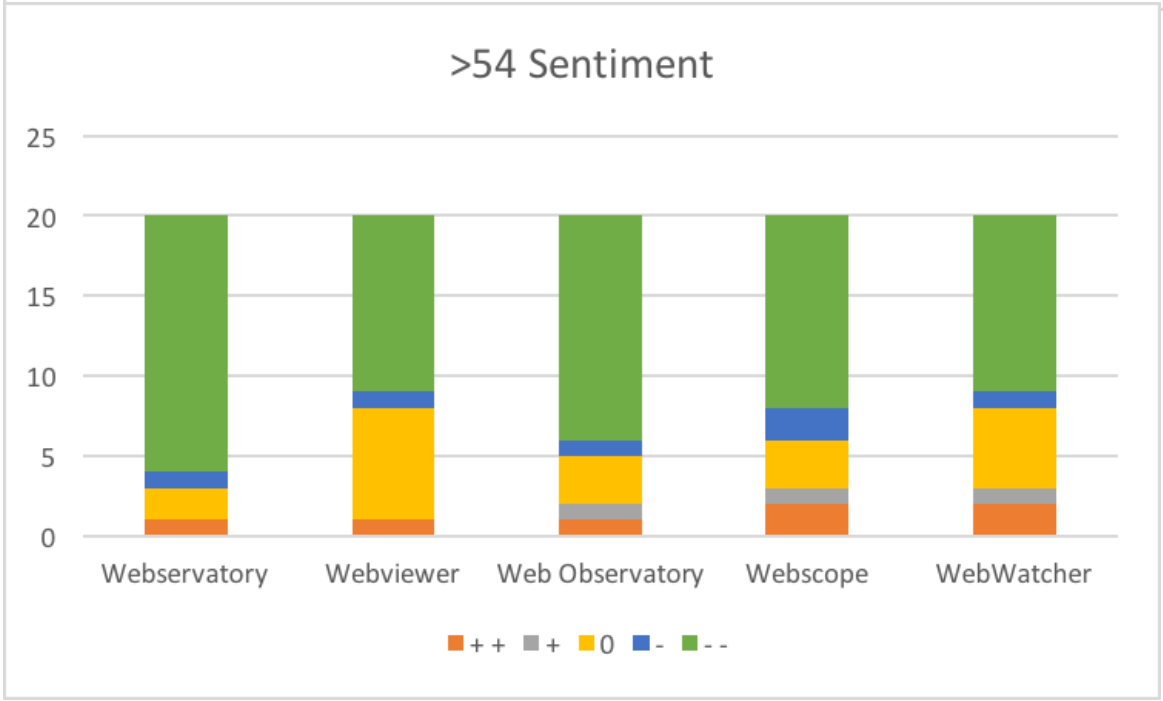
#	Answer	1	2	3	4	5	Mean
1	WebViewer	4	6	2	4	0	2.4
2	Webservatory	4	5	3	1	3	2.6
3	WebWatcher	4	3	2	1	6	3.1
4	WebScope	3	1	3	6	3	3.3
5	Web Observatory	1	1	6	4	4	3.6

Total unique respondents 16

Rating interval = 1.2

Appendix B





Please rank the names from your most positive impression at the top to your most negative impression at the bottom

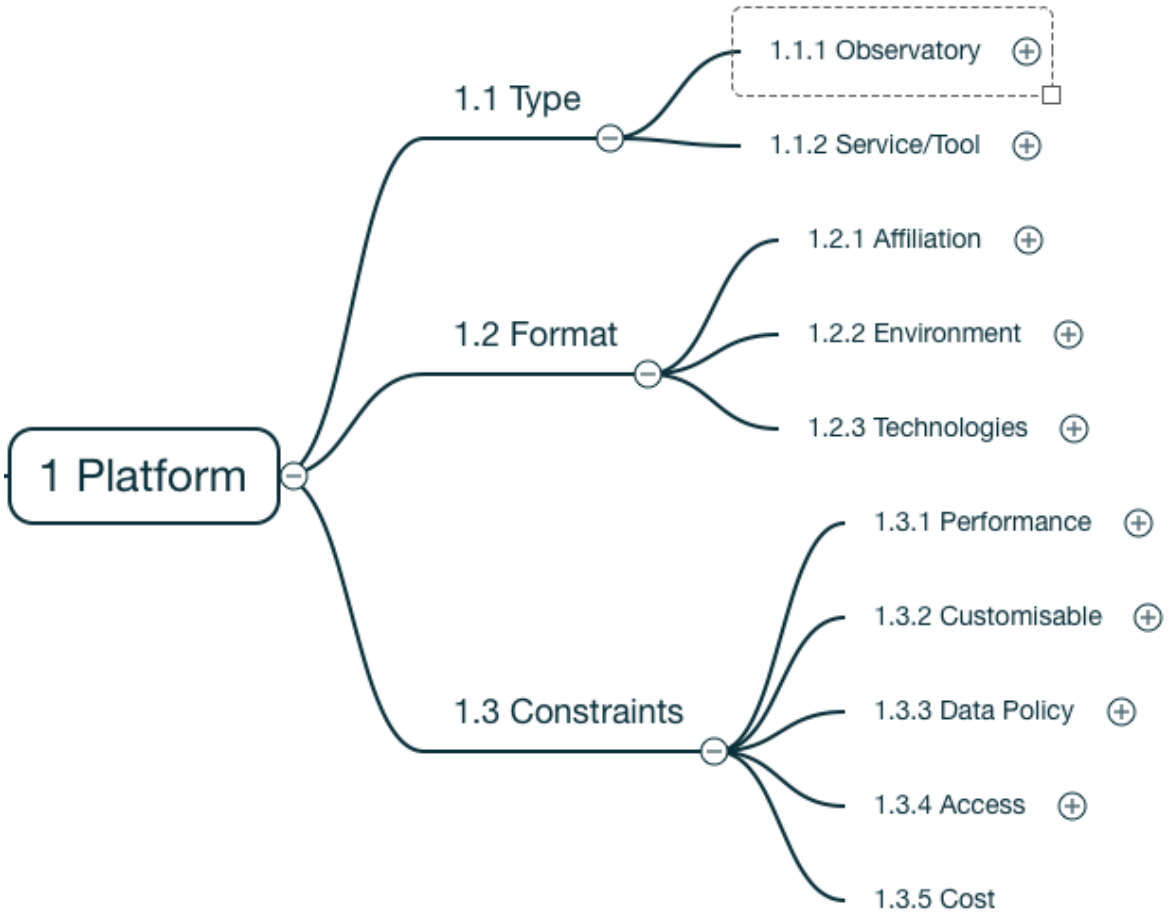
Mean indicates the average ranking each item received. Because "1" is the highest ranking, the item with the lowest mean is the one that was ranked most highly.

#	Answer	1	2	3	4	5	Mean
1	WebWatcher	8	5	1	1	5	2.5
2	WebScope	3	6	5	4	2	2.8
3	WebViewer	6	2	3	7	2	2.9
4	Web Observatory	1	5	6	5	3	3.2
5	Webservatory	2	2	5	3	8	3.7

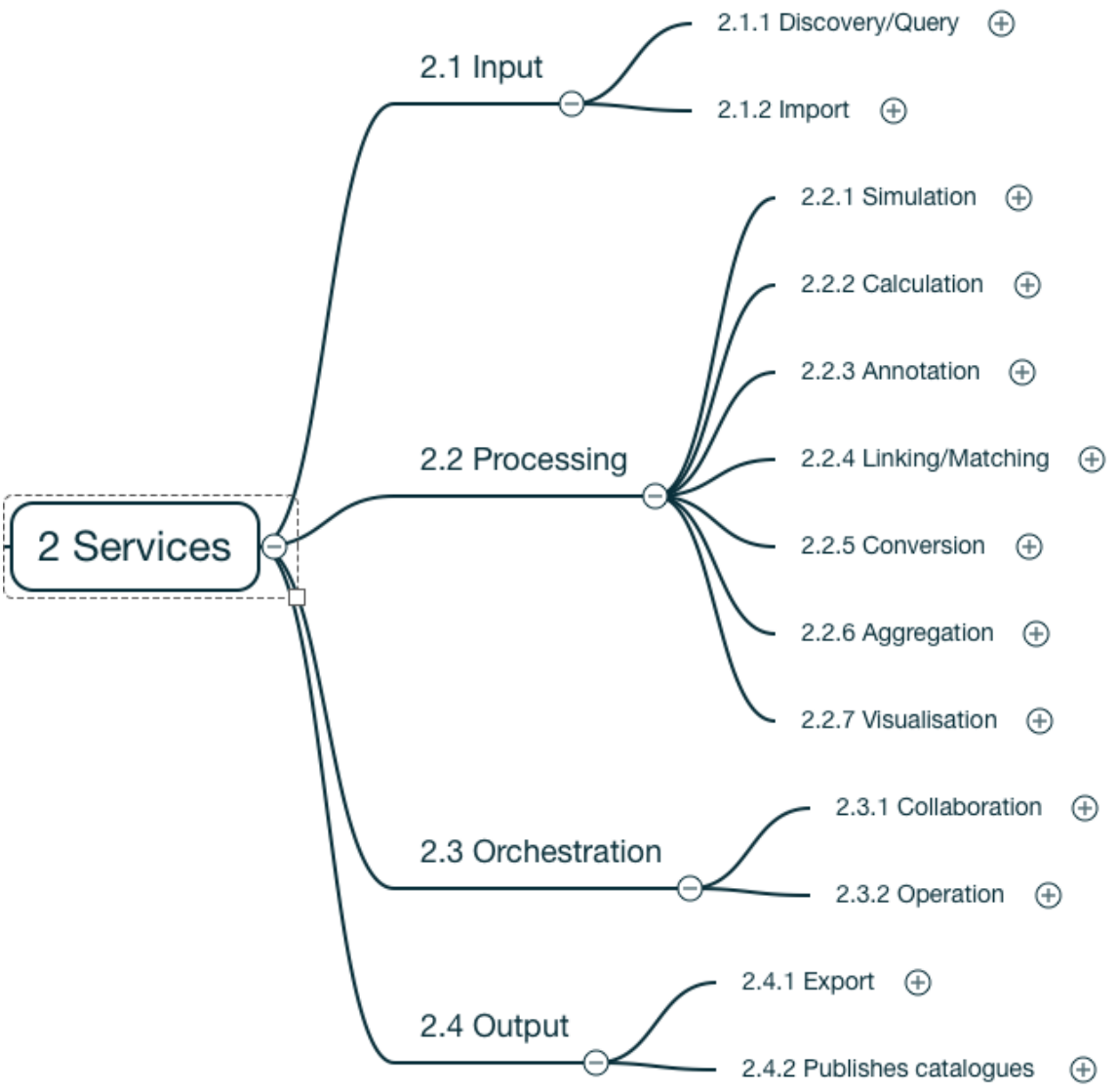
Total unique respondents 20

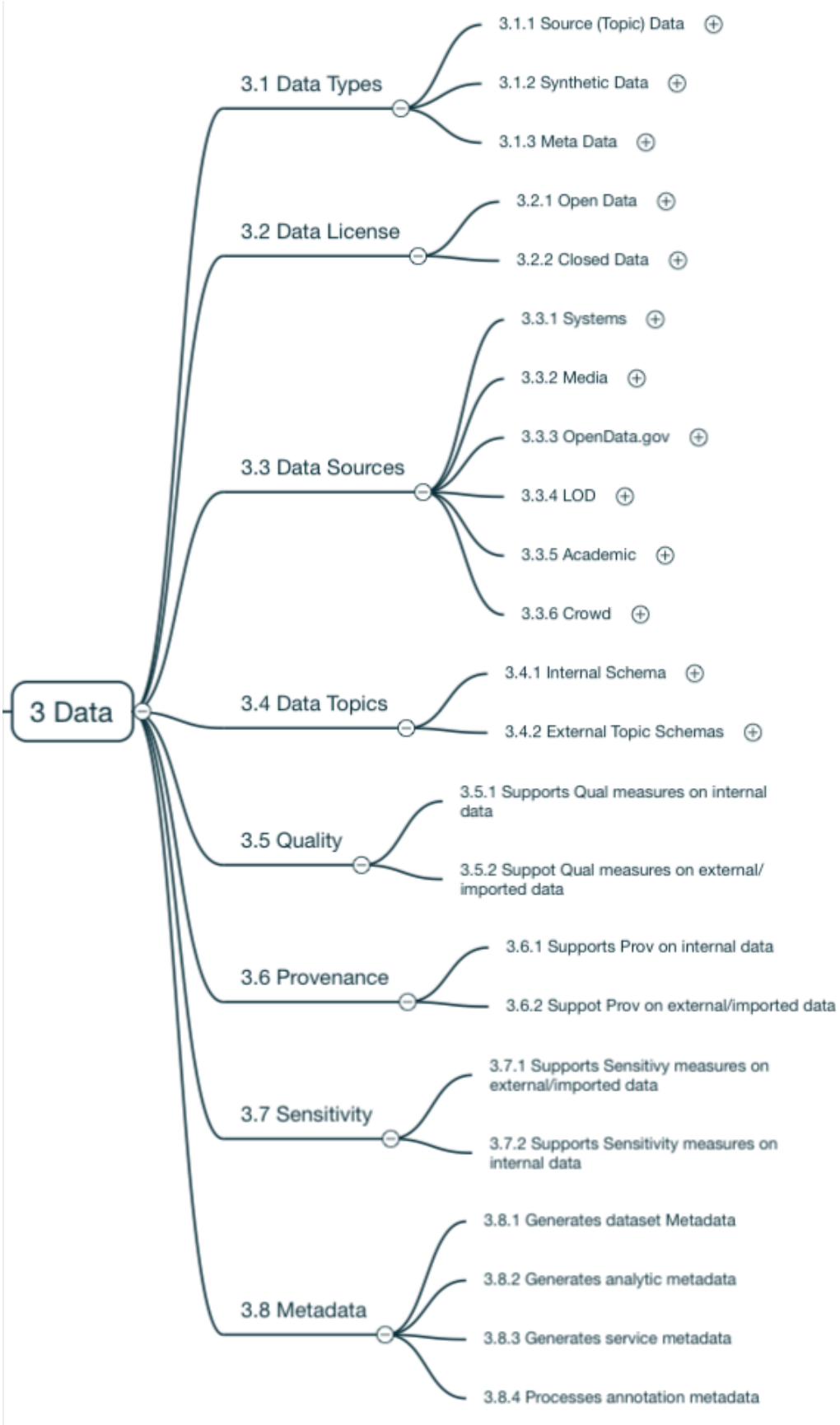
Rating interval = 1.2

Appendix C Seed Model

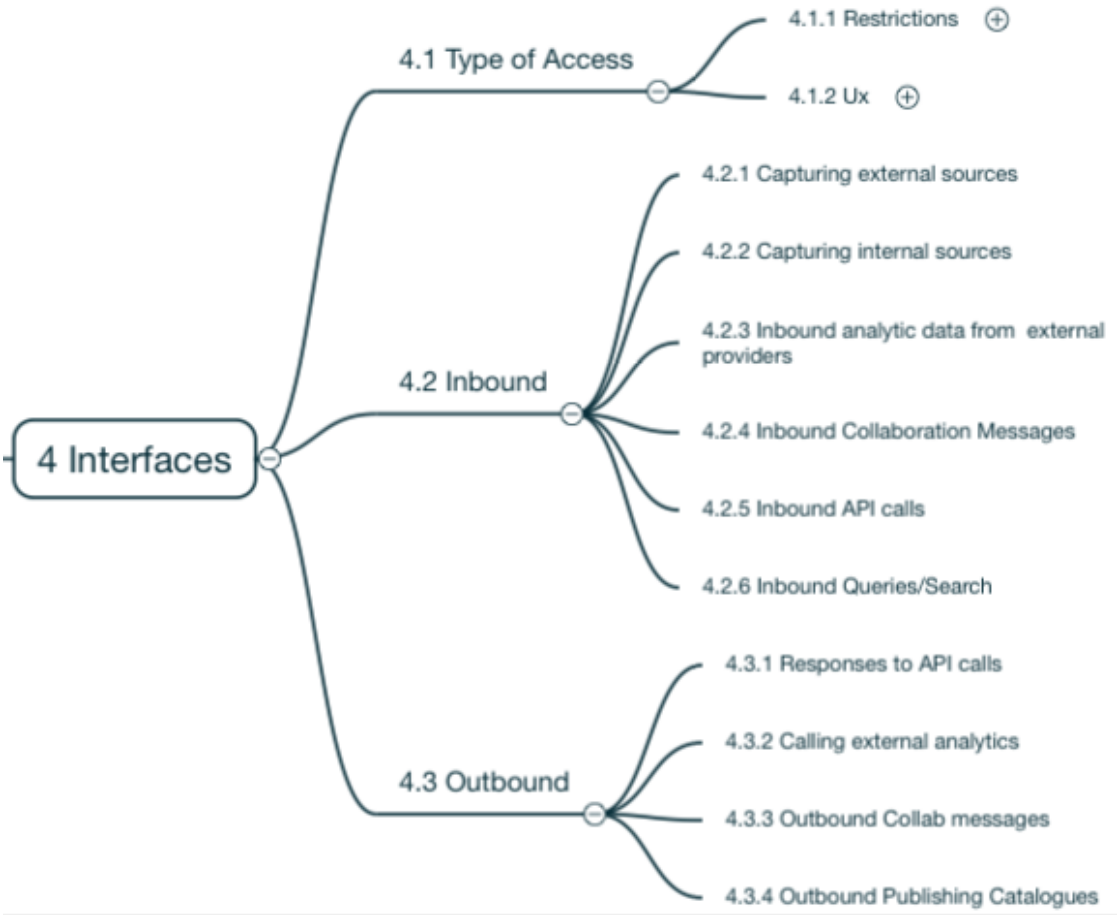


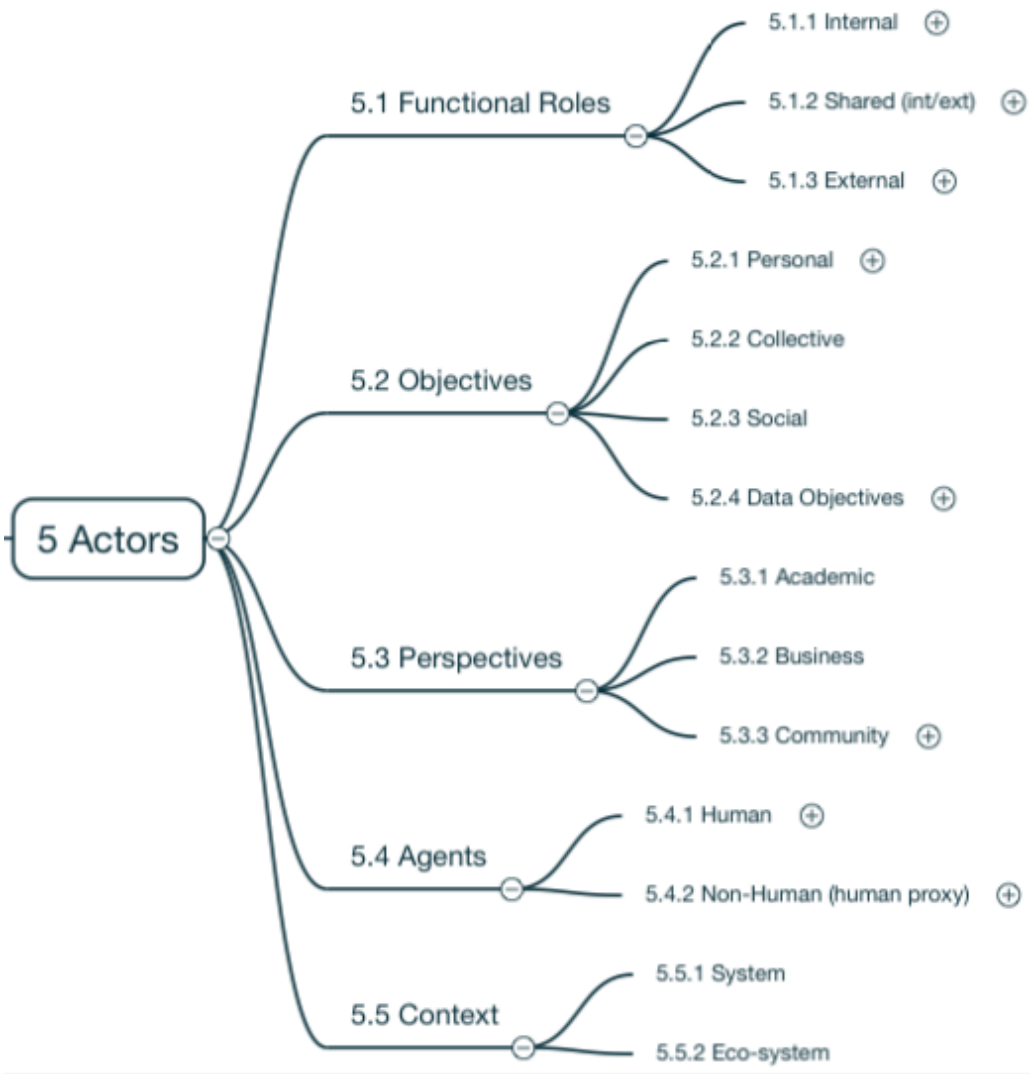
Appendix C


2 Services
2.1 Input
2.1.1 Discovery/Query
2.1.2 Import
2.2 Processing
2.2.1 Simulation
2.2.2 Calculation
2.2.3 Annotation
2.2.4 Linking/Matching
2.2.5 Conversion
2.2.6 Aggregation
2.2.7 Visualisation
2.3 Orchestration
2.3.1 Collaboration
2.3.2 Operation
2.4 Output
2.4.1 Export
2.4.2 Publishes catalogues


3 Data
3.1 Data Types
3.1.1 Source (Topic) Data
3.1.2 Synthetic Data
3.1.3 Meta Data
3.2 Data License
3.2.1 Open Data
3.2.2 Closed Data
3.3 Data Sources
3.3.1 Systems
3.3.2 Media
3.3.3 OpenData.gov
3.3.4 LOD
3.3.5 Academic
3.3.6 Crowd
3.4 Data Topics
3.4.1 Internal Schema
3.4.2 External Topic Schemas
3.5 Quality
3.5.1 Supports Qual measures on internal data
3.5.2 Suppot Qual measures on external/ imported data
3.6 Provenance
3.6.1 Supports Prov on internal data
3.6.2 Suppot Prov on external/imported data
3.7 Sensitivity
3.7.1 Supports Sensitivy measures on external/imported data
3.7.2 Supports Sensitivity measures on internal data
3.8 Metadata
3.8.1 Generates dataset Metadata
3.8.2 Generates analytic metadata
3.8.3 Generates service metadata
3.8.4 Processes annotation metadata

Appendix C


4 Interfaces
4.1 Type of Access
4.1.1 Restrictions
4.1.2 Ux
4.2 Inbound
4.2.1 Capturing external sources
4.2.2 Capturing internal sources
4.2.3 Inbound analytic data from external providers
4.2.4 Inbound Collaboration Messages
4.2.5 Inbound API calls
4.2.6 Inbound Queries/Search
4.3 Outbound
4.3.1 Responses to API calls
4.3.2 Calling external analytics
4.3.3 Outbound Collab messages
4.3.4 Outbound Publishing Catalogues


5 Actors
5.1 Functional Roles
5.1.1 Internal
5.1.2 Shared (int/ext)
5.1.3 External
5.2 Objectives
5.2.1 Personal
5.2.2 Collective
5.2.3 Social
5.2.4 Data Objectives
5.3 Perspectives
5.3.1 Academic
5.3.2 Business
5.3.3 Community
5.4 Agents
5.4.1 Human
5.4.2 Non-Human (human proxy)
5.5 Context
5.5.1 System
5.5.2 Eco-system

Appendix D Taxonomy

DNA / D-Facets

External data sources	1	Data source(s) provided from outside the system scope
External observatories	2	Data source(s) provided externally via another system/Observatory
Data capture API	3	Interface through which data is acquired
Discovery API	4	Interface through which existence/format/terms of data/services is determined both internally and externally
Internal data sources	5	System accesses data sources within the scope of its own control/border
Raw data	6	System stores externally acquired data in its original format/layout
Ext 3rd-party request	7	3rd party requesting a service provided by this WO
External data linking	8	Reference/link to data not stored on this WO
Analytics + visualisation	9	Creation of visual representation and calculated results
Processing + service API	10	Interface handling inbound/outbound service requests
Ext 3rd-party processing	11	3rd party providing service to this WO
Processed + synthetic data	12	Storage of transformed, simulated or combined data distinct from raw data
Machine intelligence	13	Use of ML to analyse data directly and/or human intelligence relating to data
Provenance, quality + trust	14	Supplementary information and processes relating to confidence in the source/quality of data assets
Data transformation	15	Process by which data is converted to different formats, different levels of granularity or resolution for use or storage

Human intelligence	16	Use of experts/crowd to analyse (categorise) data
Metadata	17	Processing and storage of information about other data assets separate from the data asset itself
Collaboration + annotation	18	Processing of enabling users to access, share, comment, markup, discuss and contextualise/link data assets with research
Permissions + licenses	19	Allocation and tracking of rights and responsibilities with assets
Service API	20	Interface handling the access to WO resources to user-facing applications
3rd Party application/Ux	21	Application accessing WO for users managed by group outside internal scope
Workflow + automation	22	Process of routing, orchestrating, scheduling tasks/requests/messages between human/machine components for the WO
Configuration + control	23	Process of calibrating/operating the WO from an operational perspective
Service + publishing	24	Process regulating which services may be consumed by internal/external applications and which data items will be published
Ux API	25	Interface through which user application accesses the WO
Internal Application/Ux	26	Internal scope user application for WO
Publishing API	27	Interface for publishing data
Published data sources	28	Set of data assets made available externally by this WO
Published discovery data	29	Set of metadata or catalogues updates made available to allow the discovery of assets on this WO

DNA / N-Facets

Group	Factor / Process	WO usage/exchange is affected by the …
ecosystem	Corp Structure	..Which may affect how/where organisations (not only commercial organisations) are able to participate in terms of authority, jurisdiction, charter, stakeholder impact.
	Convenience	.. extent to which access is made ubiquitous and usable for non-specialists affecting WO usage/exchanges
	Community	.. extent to which there are groups (communities of interest) and to what extent they have access to engage/share via WO
	Celebrity	..extent to which the group and/or the research issue enjoys public scrutiny and which sets expectations around priority, importance and inclusion of the said entity.
	Cost	..extent to which access to WO resource (people, technology, data and service is made cheap) and which → emergent/cost-benefit element.
	Charitable	.. extent to which applications are found for non-profit outcomes and the desirability/sustainability of these
	Compulsion	.. extent to which regulation/legal frameworks force action/inaction. The arduous nature of regulation, the potential penalties of non-compliance and the expected enforceability
	Commercial	.. extent to which chargeable applications are found and the profitability/sustainability of these. Describes the extant tendencies around market-share, intellectual property and control which may affect

		what firms are prepared to share and under which conditions
	Collegiality	..extent to which there are professional groups (communities of employment/engagement) and the extent to which they will tend to collaborate/compete around data/resources
Encounter	Confidentiality	..extent to which the material being accessed is sensitive, and there restricts its audience
	Conflict (resolution)	..extent to which terms/costs/ownership/access may be contended and resolved
	Charging Model	..extent to which tariffs may be applied to WO data/services based on usage, terms, group membership, location
	Connection	..extent to which users/systems use cheap + ubiquitous standards methods/standards to make an inbound/outbound technical connection to WO resource+services
	Canonical Source	..extent to which it is feasible to establish a preferred (de facto) source/version amongst alternatives across one/many WOs
	Communication	..extent to which different types of dialogue are managed/orchestrated ranging from discovery of sources, the disclosure of metadata, establishing connection/permission and the grant of license
	Clarification	..extent to which a multi-step process is needed to ascertain meaning (implies initial response is unclear/unambiguous)
	Certification	..extent to which the user requires (and the operator grants) some confirmation of quality, authenticity or provenance. Implies value for the user and liability for the operator

	Collecting / Charting	..extent to which WO records meta-data about its own resources and operations
	Conspicuity	..extent to which the level of awareness or level of disclosure about data capture and observations has different effects on the observed systems/users themselves
	Consignment	..extent to which data may be added to or linked with data in the WO as a data deposit.
Enhance	Commentary	..extent to which researcher-generated or machine generated annotations are added to WO resources
	Calibration	..extent to which data/service is adjusted in line with a known effect (error or bias)
	Capture/Charge/ Crowdsource	..extent to which data is provided to WO incrementally (Capture), through the upload of an existing dataset (Charging) or through direct user input (Crowdsourcing)
	Collections	..extent to which incremental values update one (or more) larger longitudinal datasets
	Computation	..extent to which derived/calculated values are required to support services or synthetic datasets
	Contingent processing	..the extent to which WO resources may undergo contextual processes based on a classification of source/usage/format
	Conversion	.. extent to which data may be stored (1) once in a global/harmonised format (2) in one of many formats (3) redundantly in multiple pre-converted formats (4) converted to a required format on-the-fly
	Correlation	.. extent to which WO supports correlation metrics across compatible datasets

	Classification	.. extent to which meta-data is added to resources grouping them according to a formal schema and/or folksonomy allowing discovery of potentially relevant sources across multiple WOs
	Co-creation	..extent to which humans and machines across different WOs may be involved in the creation and curation of resources
	Construction	..extent to which datasets may be aggregated, synthesised, updated from one or more individual sources/resources
Execute	Consumption	..the extent to which the connection/usage of one/more resources is required and measured in the deployment of a larger or aggregate resource
	Conflation/ Compression	..the extent to which the frequency/volume of data updates is managed to achieve different levels of data granularity/resolution though techniques such as arithmetic averaging, periodic sampling, aggregation/difference reporting and interpolation
	Circulation	..the extent to which (updated) resources are actively "pushed" or notified to interested parties to sync common understanding
	Collaboration	..the extent to which WO supports the sharing of ideas/messages/tasks/responsibilities between parties in pursuit of diverse goals in addition to the exchange of data. Does not require formal agreement but may informally imply roles, responsibilities and deadlines
	Constrain	.. the extent to which WO recognises partial restrictions on the scope/bandwidth of resources that are permitted for a specific user/system
	Contextualise	..extent to which an accurate output/rendering of resources (data/visualisations etc.) may depend on

		meta-data relating the requesting user or system or sources
	Curation	..extent to which local resources (rather than linked/referenced) may require (semi) automated or manual processes of selection, deletion, correction annotation and (re)classification
	Cataloguing	..extent to which references are discovered/updated for ext. linked resources
	Choreography	..the extent to which resources that are assembled from a set of int/ext components → resulting synthetic output(s) may need refreshing/synchronising in specific sequences and/or 'staging' at specific points in time
	Compartmentalisation	.. the extent to which WO supports the creation of tiered services based on community membership, license, confidentiality, jurisdiction or other framework
	Citation	.. the extent to which WO supports the disclosure of credit/ownership of resources and additionally the recognition/measurement of the re-use of resources by third parties
Emergence	Confidence	..the extent to which data/services operate within known bounds/limitations
	Credibility	..the extent to which reputation of specific sources or systems are recognised/rated
	Convention	.. the extent to which specific patterns of usage, behaviour and operation may become de facto rather de jure standards expressing the wishes, style and preferences of the community of users/providers (distinct from tech standards)
	Consensus/ Convergence	.. the extent to which understanding of facts, processes may converge over time as a result of discussion/collaboration on WO

	Commercialisation	.. the extent to which certain sources, services may be fee-based or evolve free→freemium→premium over time to address cost of service and other commercial objectives
	Collective action	.. the extent to which WO acts as focus/platform for groups to act collectively around issues of common interest
	Catalysing	.. the extent to which use of WO (as a meme) promotes interest in WO directly or in hosted research issues
	Credit	.. the extent to which the act of discovery, participation or sharing within the ecosystem results in reputational effects - distinct from (resource) citation
	Cascades	.. the extent to which patterns of usage and interoperation between systems and users create emergent patterns/insights (relates to Charting and Conspicuity)
	Consistency	.. the extent to which standards for sources, services and processes may emerge over time
	Confirmation	.. the extent to which WO may assist in reaching/testing research conclusions
	Conclusions	.. the extent to which WO supports decision modelling
	Coherence	.. the extent to which WO supports identification/management/resolution of contradictory inputs/outputs
	Cost Benefit	.. the extent to which users conclude the time/resource/funding model for WO is in line with the value of the outputs / effects
	Cohesion	.. the extent to which the various (trusted) distributed WO resources are aligned to support overall workflow (albeit manually driven)

	Conformity (vs. Subversion)	.. the extent to which operations are facilitated (blocked) by (in)compatible standards
	Complexity	.. the extent to which the interaction of increasing numbers of distributed resources affects other factors (such as Coherence/Cohesion)
	Contracts	.. the extent to which formal agreements and licenses result from participation on WO (distinct from collaboration)
	Consequence	.. the extent to which positive/negative results accrue from the identification, attribution and accountability around data/services under known terms of usage
	Culture	.. the extent to which behaviours specific to WOs, WO projects informally appear as a result of the development of a Social Machine as in vivo practice vs. formal contracts

DNA / A-Facets

Human Agents	The portion of the WO embedded in heuristics, motivations and actions of human agent
System	The internal context of the WO implementation
Owners	WO Investors, owners, managers and policy makers
Human Substrate	The human input required to install, maintain and operate WO systems
ecosystem	The external/broad context of the WO implementation
Human Computation	The human cognitive effort typically to classify, annotate and curate data
Service Partners	Humans providing parts of the WO functionality/service
Service Users	Humans consuming WO resources/services
Regulators	Humans setting external rules, measures, legislation
Competitors	Humans disrupting WO legally - according to rules
Hackers	Humans disrupting/subverting WO - typically illegally/criminally
Non-Human Agent	Portion of the WO embedded in mechanistic, algorithmic and AI/ML processes and structures
System	Internal context of WO implementation
Algorithms	Encoded logic dictating the actions/models for the system (e.g. calculations, features)
System policies	Encoded (typically static) configuration defining limits/boundaries for the system. e.g. permissions, quotas
ecosystem	The external/broad context of the WO implementation
Technical substrate	The technological layers supporting the operation of the WO and it connection to user (e.g. Networks)

Ext Client systems	Systems mediating OUTBOUND resources/services from WO
Ext Service	System delivering INBOUND resources/services to WO
Bots	Autonomous software agents typically attempting to subvert systems (typically in concert with Hackers)
Individual Agency	The behaviours/actions of human agents on WO/ W^3O for the purpose of maximising personal utility/benefit
Psychological theory	Models of human individual behaviour/cognitions informing our understanding of individual agency
Values	The meaning and relative importance given to objects/concepts available on or resulting from WO
Resources	The tools/facilities/abilities with which action may be taken e.g. data, skills
Emotions	The sentiment triggered by potential/actual behaviours e.g. fear, desire
Individual Choices/behaviours	Particular course of action/decision chosen as a result of what the individual believes a thing is/means, how valuable it/outcome is relative to other choices, the ease of enacting the behaviours and the level of confidence in the consequences and outcomes.
ecosystem	Introduction of other players/agents and their resources/behaviours into the WO model
Agent-based systems	Introduction of non-human "players" into the WO ecosystem (these may or may not be recognisable by other agents as non-human)
Collective Agency	Choices/Behaviours modified by the knowledge of and/or impact of the actions/resources of others e.g. collaboration, competition, standardisation. Could be positive/negative competing effects

Emergent effects	Net behaviours arising from the interaction of collective agency - not necessarily either one of the competing effects. e.g. a compromise, an imposed/negotiated alternative.
History/Experience	Choices/behaviours modified by data/learning about previous outcomes.
Social Agency	The regulation of action to maximise overall group utility vs. individual benefit.
Social Theory	Models of collective behaviour/cognitions informing models of social structure and human collective agency
Technical agency	Voluntary technical standards and agreements
Economic agency	Market mechanism for determining resource allocation through taxation, grants and pricing
Environmental agency	Standards for actions/rules/processes relating to ecosystem community impact
Political Agency	Actions of leadership groups and related notions of policy, accountability, transparency and representation
Legal Agency	Enforceable standards of behaviour with potential for punishment for non-compliance

Appendix E Interviews

Participants

AKA	Gender	Type	Role
Bella	Female	Academic	Architect
Ben	Male	Community	Innovator
Chad	Male	Business	Curator
Charlie	Male	Business	Curator
Chris	Male	Academic	Curator
Christian	Male	Business	Innovator
Clement	Male	Business	Innovator
Connor	Male	Business	Innovator
Craig	Male	Academic	Architect
David	Male	Academic	Architect
Davina	Female	Community	Innovator
Edward	Male	Community	Curator
Edwin	Male	Academic	Curator
Eleanor	Female	Academic	Architect
Gail	Female	Community	Curator
Herbert	Male	Business	Architect
Ian	Male	Community	Architect
Igor	Male	Business	Architect
Imelda	Female	Academic	Curator
Iris	Female	Academic	Curator
Irving	Male	Community	Curator
Isaac	Male	Business	Architect
Ivan	Male	Academic	Architect
Kate	Female	Community	Curator
Kerry	Male	Academic	Curator
Lana	Female	Academic	Innovator
Lester	Male	Community	Architect
Lorna	Female	Community	Innovator
Michel	Male	Academic	Curator
Mike	Male	Academic	Architect
Group	Mixed	Academic	Innovator
Group	Mixed	Business	Innovator
Group	Mixed	Business	Architect
Group	Mixed	Business	Curator
Group	Mixed	Community	Innovator
Group	Mixed	Community	Innovator
Group	Mixed	Community	Architect
Group	Mixed	Community	Curator
Group	Mixed	Community	Architect
Narinda	Male	Community	Architect
Nathan	Male	Academic	Curator
Neil	Male	Academic	Curator
Nick	Male	Business	Architect

Noah	Male	Community	Innovator
Noel	Male	Community	Innovator
Norbert	Male	Academic	Architect
Norman	Male	Community	Architect
Peter	Male	Academic	Innovator
Queen	Female	Academic	Curator
Quela	Female	Academic	Curator
Quentin	Male	Academic	Architect
Quinn	Male	Business	Architect
Sharon	Female	Business	Innovator
Stefan	Male	Business	Architect
Tamara	Female	Community	Curator
Tara	Female	Academic	Innovator
Ted	Male	Academic	Innovator
Terry	Male	Academic	Architect
Theo	Male	Academic	Innovator
Thomas	Male	Business	Innovator
Toby	Male	Community	Innovator
Tony	Male	Business	Architect
Turner	Male	Community	Curator
Ulf	Male	Academic	Architect
Uli	Male	Business	Architect
Uma	Female	Community	Curator
Ute	Female	Community	Architect
Victor	Male	Community	Architect
Xan	Male	Academic	Architect
Xander	Male	Academic	Innovator
Xantha	Female	Community	Innovator
Xavier	Male	Business	Architect
Xena	Female	Academic	Innovator
Xeno	Male	Academic	Curator
Xero	Male	Academic	Curator
Yosef	Male	Business	Innovator
Zoe	Female	Community	Curator

WST

In this section, we consider three interviews representing the WSTNet. The Web Science Trust (www.webscience.org) operates as a not-for-profit organisation promoting Web Science as a research discipline and advises on Web Science related issues. WST comprises:

- The WST Board of trustees from academia and industry who produce guidance on research, education, policy and advise at a government level on matters relating to Web Science
- The WSTNet spanning 20 university Web Science research groups globally who actively engage in Web Science research projects and train Web Scientists
- The WST Admin team where I have (ob)served the board and the WSTNet over 4 years.

Three senior representatives from WSTNet [Imelda], [Ted] and [Ivan] have been selected for IPA analysis to represent convergent/divergent conceptualisations of WOs framed within academia.

[Imelda]

[Imelda] is an experienced and senior academic and she talked with me about the WO from a visionary's perspective. She expressed a passion for the potential of the WO and positions WO as a tool for researchers to interact with data about the Web whilst characterising Web Science itself as study of humans interacting with the Web. She quips:

> "We are really building a Social Machine to observe Social Machines."

This binds WOs (or specifically Web *Science* Observatories) tightly to the definition of Web Science and to the study of Social Machines for [Imelda]. Our interview dealt with ideas grouped into the following themes:

- Observation
- Collaboration
- Social factors
- Data
- Trends + Benefits

A narrative map is shown below (Figure 6-1). The map highlights area of positive/negative sentiment, focus and objectives. Refer to the Appendix for a full description of the notation.

Appendix E

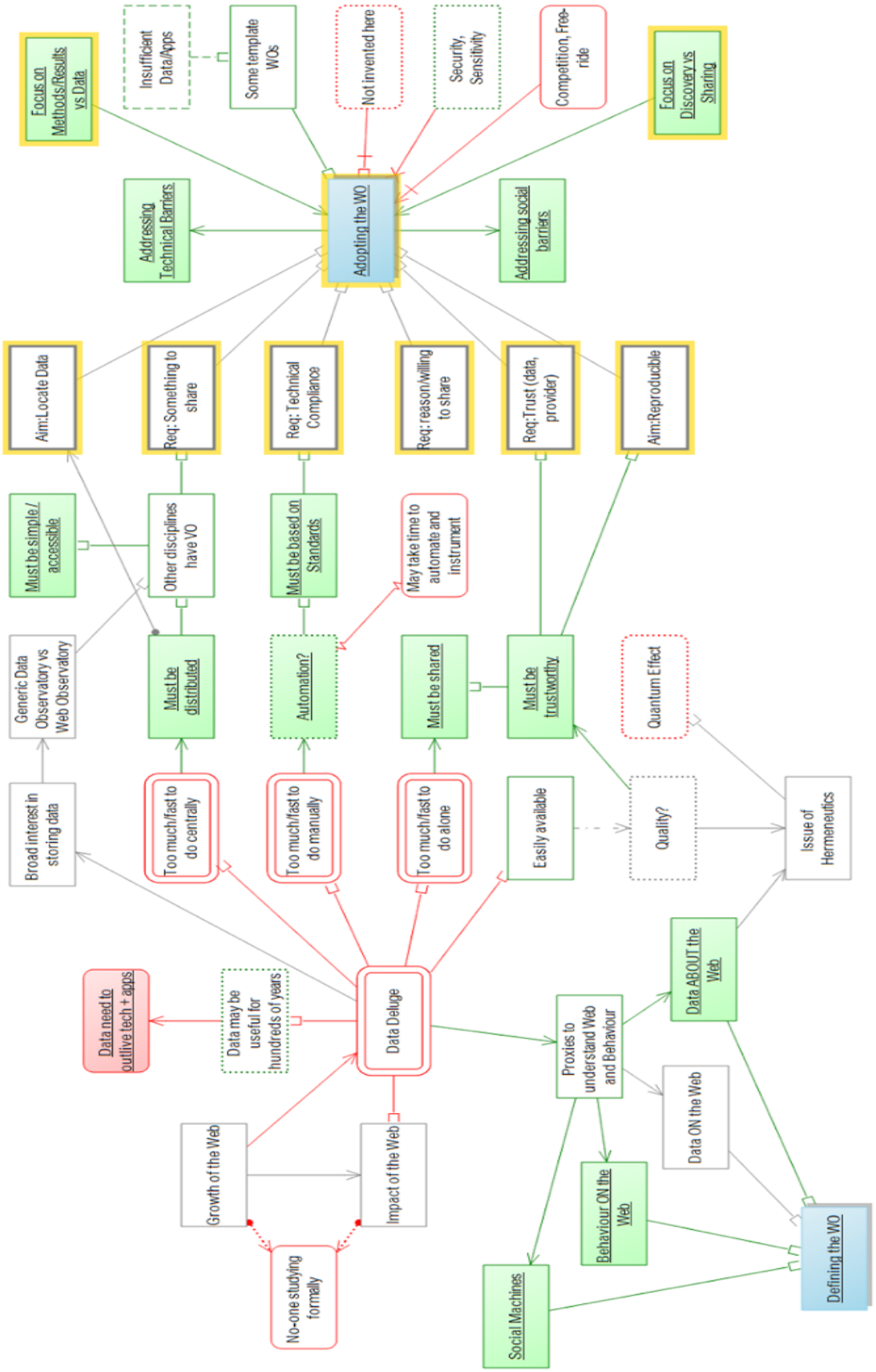


Figure E-1 [Imelda] Narrative

She explained that part of the idea/vision for WO took inspiration from a grand science fiction metaphor:

> "When we started Web Science we talked about the Asimov [Foundation] book, about this whole idea of psychohistory and [whilst] you couldn't predict what one person could do you could potentially predict or forecast what people do en masse .. by taking lots of data over time and analysing it on a longitudinal basis you could potentially start to forecast the extremes."

This encompasses several key Web Science themes including emergence at (Web) scale, prediction/modelling and longitudinal (even generational) perspectives suggesting a new type/scale of science.

> "Hari Seldon went off for a millennium before he [got] results: or his descendants did."

This is an ultra-high level view envisioning the life of curated data beyond the life of associated technologies/platforms in order to solve unknown future problems. [Imelda] thus places the WO in a broad historical context citing examples from climate science:

> "In order to understand [climate] you have to gather [weather] data from all over the world [ranging from] little people to big organisations..and how you do that? Sharing! It's become possible because they have now 200/300 years of data, initially hand-logged but increasingly computer-logged."

This left me with an intriguing impression of WO as being closer to Stewart Brand & Daniel Hill's notion of "a deep library for the future" at the Long Now project than to a pre-contextualised, filtered, ephemeral/snapshotted Google search.

She stresses that one cannot always know what data may be needed for even the short term making flexible thinking seem paramount:

> "it's all evolving so fast..if you set up these longitudinal things - by the time you've even thought about it, everything has changed."

Based on data at such speed/volume and the inability of humans to consider it in real time, we change our relationship to data and evidence with a consequential reliance on machine processing, trusting (even delegating) our understanding to the automated interpretation of data streams which no human may ever perceive/review directly.

In addition to having data to share and a willingness to share, [Imelda] stressed the framework/technology with which to share:

> " The other thing is standards. Standards are absolutely crucial because if you don't have data interoperability..without standards..you can't get any results."

Here [Imelda] is presenting agreed technical frameworks as “necessary-but-not-sufficient” and looks to the academic traditions of sharing in disciplines such as physics and astronomy who

> "have to share the information they get from the telescopes in order to build the big picture and do the data analysis."

She shared how the Virtual Observatories underpinning collaborative work in Astronomy inspired the WO and informed what could be done to bring diverse actors, projects and datasets together leading me to follow up with further interviews with this group:

> "suddenly it was this thought .. we need Observatories for the Web."

Although the core principles are shared [Imelda] knows the nature of data shared in Web Observatories is different to other data repositories:

> " the physics, chemistry, biology stuff tends to be much more homogeneous."

and she places limits on the comparison between Web Science and other "Big sciences" (e.g. Climate Science, Carbon Observatories) pointing both to observer effects and to stricter limits to claims of knowledge in Web Science. Unlike observing remote stars/galaxy where we do not believe our actions have any impact, the more embedded nature of working on/with the Web leads Web Science to being part of the system being observed:

> "You have a potential quantum effect because by just observing people they potentially change their behaviour."

The issues of data interpretation (evident in her own example of disputed conclusions in climate science) are not forgotten:

> "of course, it's incredibly controversial as how that data is interpreted. It's looking at the big picture of the climate and what impact we as a human race will have there."

She re-emphases the varied and sometimes competing results that can emerge as "truth" from the same source data given the multiple hermeneutics involved in a process where humans are 'sources of', 'processors of' and 'consumers of' data.

> "we're talking about human beings [being] involved, so we're not dealing with a natural phenomenon. You're dealing with human beings who are completely unpredictable in what they're going to do."

This touches on the idea of Psychohistory once more as a way of adapting to the inherent unpredictability of individuals. When I asked her to define WO for me, [Imelda] characterises it as having three parts which correspond to:

- The physical layer
- The content layer
- The policy layer

expressing how the system enables the things that users of the system will want (or be able) to do with which data. This bears a notable similarity to the emerging D+N+A model derived earlier. The idea of aims/policies is used to move the definition on from a purely academic Web (Science) Observatory position conceding:

> "That policy can be a policy that governments create or it can be the policies created by the big businesses."

Explicitly placing WO in a multi-sector context delivering:

> "[Insights] that could be to help businesses to do things better; it could help governments make better policies"

Whilst this position may seem antithetical to a WO for Web Science only, the situation is more complex given Web Science may, for example, be interested in the impact/process of government on/via the Web in a way which is distinct from government's own interest in governing on/via the Web (i.e., impact vs process). Thus, we may consider that requirements/motivations to OBSERVE DATA underpinning government on the Web may be different from observing data IN ORDER TO GOVERN on the Web. [Imelda] also talks about the "one vs. many" nature of a WOs or groups of WOs, and it seems there are no simple definitions here either relating to single or multiple systems/hosts:

> ". .we are partitioning the Southampton Web Observatory to allow people to put their own data onto our service, so you can have multiple Observatories, multiple instances of Observatories on the same server. Southampton may need to start to split its Observatory across different sites."

Focussing on the question of how [Imelda] viewed the WO, we talked about the difference between AN Observatory and THE Observatory and she conceded:

> "..this is quite hard - the semantics of it."

She gave an analogy for WOs and a "Web-Of-WOs" that has been used earlier in this document:

> "You have Web servers, you have Web sites, and you have the Web. It's the same analogy.. so Observatories are instances of data and THE Web Observatory is really the Web Observatory project which is cataloguing activities to tell people what's where and who they can collaborate with."

Imelda makes an implicit point about collaboration here – something which is broadly assumed and yet notably few tools/affordances are encapsulated in the WO concept.

> [THE WO] "is not actually doing the data analysis itself, it is just a cataloguing, library function really."

To explore [Imelda's] fundamental assumptions about WO, I asked if an Observatory could contain ANY data or had to focus specifically on data ABOUT the Web

> "For Web Science, it does .. It's data ABOUT the Web, about what people are DOING ON the Web . It's always about multiple things."

pointing to a mixture (at least) of inherently Web-encapsulated things as well as behaviour/choices made on the Web.

So while data ABOUT the Web is given primacy here data ON the Web is not specifically excluded:

> "of course, we can USE a lot of that data [ON the Web]. And so that's a very grey area as to where you start. I think the concept of DATA Observatories is a much wider one .. the concept is much larger than Web Science."

We see then WO grouped/mixed within a much larger notion of data hosting/sharing system that may incidentally be ON the Web but not relating back to inherently "Webby" concepts or activities with Web-as-the-primary-message and not only Web-as-the-medium.

Having situated WOs in a very broad potential ecosystem of sources and uses which range widely in the types and sensitivity of the (shared) data involved I asked about groups collaborating, sharing and exchanging data:

> "this has to be, by the nature of it, a distributed activity. You have to collect data in a distributed manner, and you can't just drag it all onto one server and do big data analytics on it because of the premises of security trust issues."

and so minimum standards around sharing the existence, availability and re-useability of data are required:

> '"At the very least we want people to share their metadata about what data they've got. And they can just put a cross against sharing to start with, but at least you know that they've got some, and they link it to the results they have published .. for the moment this is telling where the data is that they might interested in for their experiments, or to learn how to do experiments or to do shared experiments with"
>
> "this is a culture change that gets people to share at least at the level of repeating experiments and getting people to share at the level of the results, what analysis you've got from the data and how you did the experiment so that people can repeat it."

She hints at some down sides to the 'academic tradition' of sharing and chides those people who might game the rules to avoid sharing:

> "to hide behind [privacy] - I can't let you have that because it's got to stay anonymous, it's .. top secret."

or citing financial and competitive reasons:

> "it cost us so much money to collect, we're not going to give any unless we [have] the first mover advantage."

Whilst a little impatient with what comes across as a tactical data considerations versus a more strategic view of WO data [Imelda] does acknowledged the inherent sensitivity of some data but focuses more on social structures and exchanges.

The social aspects and indeed the social 'embeddedness' of WO come through here not only for boosting the initial adoption of WO as an idea but also for the continued growth and sustainability, and it was suggested that a broader self-sustaining Web of collaborators sharing was required rather than one or two central systems as the only hubs.

> "We have a number of "Web Observatories" all over the world but very little data in them .. generally, the further you go out from us the less data there is. And the less data there is in the Observatories, the less other people are going to start doing it. Because there's not a reason to do it."

And thus understanding what draws users in or keeps them away is a vital insight:

> "Avoiding friction may be as important as finding inducements and drivers .. if you overcomplicate something nobody is going to use it "

> "[We need to] work out what are the things that are going to tip people into using it, and then the lessons from history, if you study how the Web evolved and how the social networks evolved, the simpler things are, the better."

There is no "land grab" for Web Science or WO apparent here beyond a conviction in the importance of studying the Web as [Imelda] is comfortable to cede core parts of the service such as dataset discovery (essentially part of THE Web Observatory) to other providers.

> "And I always say to people – 'if Google ends up doing [that] for us, that's fine' ..we need a simple search engine to start with - Where are the datasets on Ebola?"

Specifically contrasting the technical with the social aspect [Imelda] agreed that the technical challenge was about:

> "open standards and data interoperability. But what's really interesting, I think, in terms of this is what language do we have to speak."

This was a rich interview which helps us to understand the nature of the WO and how its supporters are working towards broader adoption. [Imelda] recognises the opportunity for academic collaboration and for government/business insight and innovation but I feel she positions WO as a shared legacy for future researchers who will want to know how the Web developed, how our behaviour changed it and uniquely how it changed our behaviour.

Imelda Themes

Super-ordinate themes	Sub-Themes	
Observing	.. micro/macro trends	Using WO to detect current and longitudinal changes
	..causing a change	aka the quantum effect
	.. via Social Machines	behaviour in sociotechnical systems
	..vs. searching/analytics	As a distinct process from Google et al
	.. vs. storing	Going beyond owning or collecting centrally
	.. behaviour	Web proxies for choices and actions
	..as Psychohistory	the prediction of broad trends over time
	..to support guidance	Using WO to inform policy and best practice
Collaboration	Interdisciplinary. effects	tensions + synergies btw disciplines
	Network effects	Amplification through wider adoption
	Use of standards	Formats to support interoperation
	Trust/FUD	The risks/benefits of participation and sharing
Social	.. challenges vs. technical	Social issues dominating technical ones
	.. embedding	Seeing WO as part of social exchanges
	.. machines	Hybrid systems arising between people/machines
	.. element of adoption	Valuing relationships within the adoption process
	.. sharing	Viewing sharing as a social process

Appendix E

Data	..about things	Data on the Web
	..about behaviour	Data about users on the Web
	.. about the Web	Data about the Web
	.. at scale	Challenges at scale
	..for robust science	Underpinning conclusions with robust data
	Hugging data	Reluctance to share data
	Longitudinal data	Observing over time to notice macro trends
Trends	Global challenges	Broader focus on engagement
	Big Data	"Datafication" and the data deluge
	Centralisation of Web	Counter trend to decentralised ethos of Web
	Data as a solution	Leveraging data to underpin policy/strategy

[Ted]

[Ted] is an experienced senior academic with a particular focus on what WO can deliver as evidence for research endeavours. He views the WO as a scientific instrument and focuses on what can be done with such a tool (rather than its content or design) citing outcomes in Academia, Business and Government. This indicates both the diversity of his own involvement and the framing of an approach/tool for which the context/framing can be switched on demand. Observatory work is not seen as 'beyond' any group with access to appropriate data and skills in a broader ecosystem experiencing the democratisation of data and where [Ted] says:

> " The overwhelming direction of travel is for the wider exploitation of data assets between all major stakeholders."

Our interview dealt with ideas grouped into the following guiding themes:

- Competition
- Collaboration
- Datasets

- Analytics
- Eco-systems
- Definitions/Models

A full breakdown of themes is shown in the Appendix and Figure 6-2 shows an overview of the narrative in context.

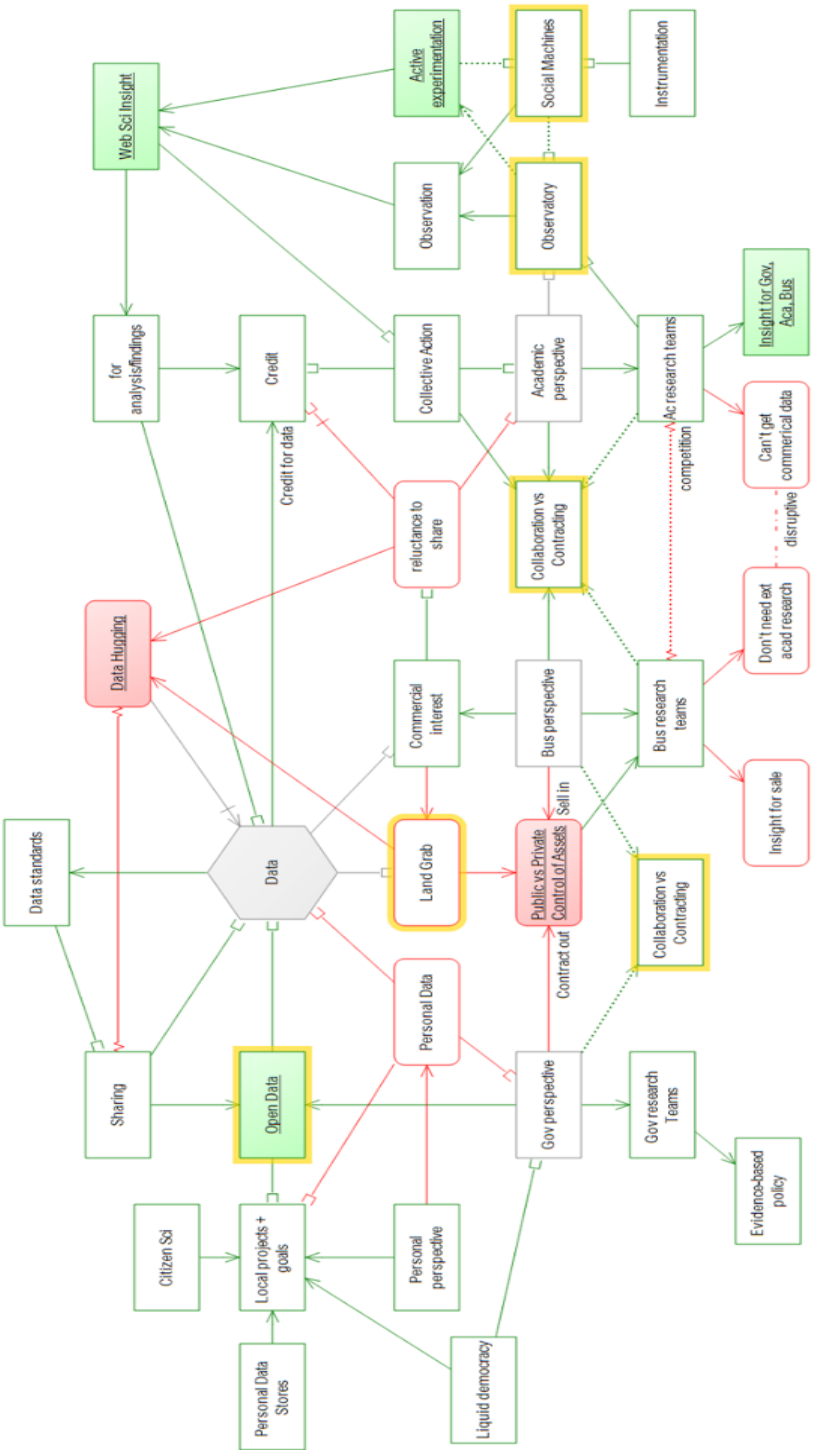


Figure E-2 [Ted]'s Narrative

I asked [Ted] where the idea for a WO for Web Science had come from.

> "We were aware of the context in which there was a 'land grab' for lots and lots of data on the Web, whether personal or corporate or whatever, and that .. corporates, were simply harvesting the Web… there were big [Observatory] efforts underway in the Far East, in particular, Singapore. That work we'd been made aware of."

He recalled that he and his colleagues were looking at:

> "New kind of offers we might come up with to facilitate research..[and]..we had a strong interest in data-at-scale on the Web .. for anything that wants to think of itself as a reasonable science or indeed even a systematic engineering discipline, you better have your primary data to hand or be in a position to notice it or collect it..[so]..the idea that we needed a significant scientific resource to do analysis on seemed kind of self-evident."

This characterises WO partly as a natural response to the growing data deluge and partly as an innovation (or even competitive) opportunity in response to external trends within the ecosystem, in business and between academic peer groups.

Basing this on the idea of Virtual Observatories used by other disciplines:

> "Wendy [Hall] was saying we need an Observatory, the equivalent to what astronomers have, and [while] that seemed like a trite analogy at first, the more you think about it, the more it's interesting."

[Ted] builds on the observatory analogy, pointing out that astronomical Observatories work even better in clusters:

> "Instruments collectively can sample the electromagnetic spectrum in a way that any [single] one can't. So collectively, you could see things that you couldn't see out of one or the other."

Inspiring the idea that WOs might work in concert and, I believe, identifying what seems to be a key difference between observatories and other standalone Web analytics:

> "Certain areas would focus on [different sources] perhaps even different aspects of the same Social Machine."

Once the idea of an Observatory for the Web was coined a focusing image was quickly chosen:

> "One of our first kind of actions immediately was to find a picture of the Hubble telescope and then stick our logo on it [laughter] - to instrumentalise the idea in a very concrete way whereby the thinking around what Observatories and observation were was partly founded/formed."

The use of metaphor here seems to have been particularly relevant here creating a much-needed understanding of what Web Science was trying to do and presenting it in a way that was accessible/understandable by non-specialists while retaining the academic credibility of existing scientific approaches. [Ted] made a very interesting observation about the notion of the observatory metaphor:

> "I think an Observatory [analogy] can lock you into thinking too passively .. Lots of science works by injecting signals into a system or various disruptions to processes, or indeed you process it into systems and then observing what happens .. [like] medical scans insert contrast medium to map flow/structure."

And so [Ted] looks for the WO to move beyond passive scanning:

> "Become an active scientific instrument as well as a passive one..[since]..living off the exhaust fumes of data is not the same as generating your own fuel."

Given that WO is based on comparable ideas in other disciplines I asked [Ted] how he responded to the challenge that WOs were not novel at all and merely "old wine in new bottles".

> "Most of the Web Science that has been described as old wine in new bottles is because we are trying to coordinate and convene disciplines, many of whom have long traditions of doing their work to a different set of questions .. I think an awful lot of what we do in science and engineering methodologies can be traced back to earlier antecedents or indeed to equivalent efforts elsewhere."

and pointing to key difference between Web Science's focused data collection of Web data and

> "Particular datasets from this piece of the Web's history or this company. But it's rather serendipitous what you have, so the principled collection at scale, according to standards which promote the re-use and analysis of that is something new."

[Ted] effectively gives a pragmatic definition of the purpose of a WO from which we may infer the required structures and functions:

> "If we are interested in the natural history or evolution of Social Machines, we have to get hold of the signature of that activity. So whether that's transactions on the Web,

> interactions with the Web, data emitted as a by-product of activity or services .. it's a notion you want to harvest as much of the data as you can. In the case of astronomers, it's photons. In our case, it's something different, partly as a collective resource to do the research on in the first place."

Looking at the notion of metrics in academia for collective research I asked about a suggestion made to me in another interview with an academic, during this study, that they would 'get no credit' if they used somebody else's dataset. [Ted]'s response was both immediate and critical:

> "I think that's just wrong-headed."

He suggests instead an evolution of a model to be more inclusive/collaborative:

> "Two lots of people get credit, the originator of the data and the author of the new insight over it ..[creating a] kind of citation, certification, virtuous cycle."

There are however, [Ted] says, plenty of situations where people will perceive benefit from holding on to their data but he is looking for a more collaborative view:

> "What we haven't quite got is enough use cases where people say, 'You know, everybody gets richer' [when the data is made open]."

Building on the idea of where much of the data is currently harvested, I asked about potential Observatory-like systems operated by commercial organisations. I asked if he recognised, say, Google as a Web Observatory:

> "Yes. Of course, I mean, I think it's lots of Web Observatories"

I note that this is *not* a view shared by all the participants and speaks to the variation, even within fairly homogenous groups, of conceptualisations in this area.

Focussing on large commercial data owners, I asked about the impact if firms should increasingly come together to aggregate even more comprehensive data - would this be 'good thing' saving us the effort of collating data or would this be a breach between academia and business?

> "Our bigger challenge is that the people who are doing this at scale are certain larger tech companies, and they have been doing this longer, harder and more intensively than we have, and that is an interesting challenge for us .. we're now starting to see a situation where large Internet companies are able to undertake research that's actually out of reach of academics."

This paints a picture of academic research into Web data being commercialised in the way that [Ted] characterises advertising/marketing adopting:

> "Psychology or mass observation studies where the interest was originally academic .. [and spawned] an entire secondary industry that developed around personnel selection or .. market segmentation."

[Ted] cites former academics with an appreciation for data who are building systems at scale outside of the academic context.

> "So many of the disruptive, large Internet and Web companies were born out of academic institutions .. they get how important the data is [and so] for us there is a real question about the viability of our [academic] research going forward if we're not able to master the data."

Though [Ted] makes it clear that for Web Science this should not be about a massive acquisition of data without purpose:

> "We shouldn't be collecting everything, and I think the Web Observatory concept needs to start to embrace corporate resources, as well as academic ones."

With so much commercial data in the hands of company research teams, I asked if this might leave the WO as an approach restricted to Open Data without commercial partnerships.

> "Well, I think it could do, but I think we have to work very hard to actually bring those organisations to the table."

With this [Ted] is partly signalling a shift/split in the nature of research between open/free resources and well-funded commercial groups with commercial offerings. The resulting competition/tension as the traditional models/roles of who gathers and analyses data are seen as disruptive. There is an implied need for well-funded international research to negotiate with (rather than compete with) the ultra-large data holders.

> "The original DNA of the field was about large-scale collaborative data sharing and astronomy is, in that sense, a very good example because as far as I can see, these have always been pan-national efforts because of the size and the scale."

though he is not specific as to whether this should be organised directly between academics or funding agencies.

> "The challenge is that a lot of what is interesting to us is actually company proprietary data [so] one of the real questions going forward is how we, as a field, make and arrange research concordance to get at some of this stuff."

This raises 'red flags' about how personal data becomes proprietary through the provision of free (sic) apps/services and platforms. Moving to governments, I asked if they were equally focussed on the use of digital data. [Ted] pointed to the historical roots of government using data citing multiple examples:

> "They did 'Observatory' [Observed] for years. I mean, what was the census [the] Domesday book?..[Government] has large-scale data going back years on traffic flow and accidents. It runs intelligence operations ..[and] ..the office of national statistics."

But in terms of moving from traditional forms/sources of data to digital data, [Ted] points to government wanting different outcomes:

> "Beginning to put in place their own data analytics capability to do two things [1] to look at how government [..] can exploit the insights from data in the delivery of services, in understanding what its citizenry want or are doing and [2] to see how government itself is working."

The delivery of such systems in practice and at national scale, however, is more challenging:

> "The capacity to actually stand up these systems and then know what to do with them and take insight from them, that's a long way off. That's an aspirational ability, though they talk about it increasingly .. government, in general, the government machine, has come to a view that - and they talk about it now in the UK explicitly - 'data-driven government' ..[However].. It depends which bit of government. They're very, very differently motivated."

This presents a void into which both academic and business will move to provide analytical insights and models. It is clear that business is willing to do so but at the cost, [Ted] warns, of losing control of data assets as recently evidenced by the selling off of UK national postcodes.

> "Many times the public sector has been contracting services out and lost control of its own data assets, which has recently come to be a problem in a number of areas."

In terms of openly sharing data between groups, I asked about the ways in which this happens.

> "Academia should, in many cases, be more proactively pushing data and insights their way [Academia to Government]. I think government to private - it's often been problematic, partly because often the private sector has been looking to sell services in."

This causes tension for the data ecosystem around sustainability/viability if data is made openly available and then sold back at a profit, re-released as open data, etc. Requiring at least results (if not underlying data) to be accessible and requiring the data/results to be freely available even if not available 'for free' (at no charge).

This also suggests open government data both as a source of data/analytics (for varied reasons/objectives) but also as a type of Observatory though with an important distinction:

> "Open government data, as a class and portal, can be regarded as a class of Observatory and in some cases very explicitly being so. But I mean this is the difference. If you're actually building a systematic way of observing Social Machines on the Web, that's a different set of requirements than being able to release the related value in data on the Web .. is this data ABOUT the Web or ON the Web: there's a big difference."

Focusing in on this key distinction: whether everyone made this distinction when using Data-ON-the-Web [Ted] conceded:

> "I think that's our Achilles heel .. [Data-ON-the-Web] is important data. But often that data will relate to transactions or interactions that are not even in the digital space. They're physical journeys in cars or the amounts paid for [items]".

I asked about the ecosystem BETWEEN data owners and to what extent those users might be motivated to share data assuming technical standards existed to do so. He paints a picture of (commercial) pressures in some areas of academia changing default open/sharing behaviours:

> "There's a lot of instinctive data retention in organisations. And this just goes back to the issue between closed, shared, and open data where there simply isn't an awareness that data released in perfectly secure ways can add to the innovative solutions."

[Ted] seizes on an important distinction in the overlap between Web Science and commercial Web analytics:

> "Academics struggle to share sometimes as much as anyone .. [and] .. in this area [Web Science] are in a peculiar position that isn't entirely shared in an area like astronomy, for example, where our original concept of the Observatory came from .. Astronomers

> don't, in general, have serious commercial contenders who are out there trying to equip themselves to do this [kind of] observation."

Constantly bridging contexts between Academia, Business and Government, [Ted] seems to be considering/resolving these challenges through all three lenses, citing moves to counteract academic/business divides by making government and broader academic datasets available:

> "The interesting example there is the administrative data centres, where there's been a real move forward to make available for academic research large amounts of sensitive socio-demographic data, data held by Treasury, and other things."

There is a real feeling (and frustration) here that without such initiatives more and more data will simply end up in commercial control placing further pressures on academic groups to find accessible, relevant/robust data sources.

> "Any kind of Observatory is only interesting at the point the data is open to you, or the insights are open to you. The underlying dataset may not need to be open but the visualisation of the 'so what' [must be]"

[Ted] broadened our discussion further by asking if my research separated out the role of the individual in this ecosystem citing:

> "Citizen science platforms, liquid democracy platforms .. personal data stores (in another context) are infrastructures where you can argue that the ultimate decentralisation, democratisation of the information asset is at the individual level."

introducing further subjective framings around the context, application and therefore the meaning of data and data interactions beyond socially-learned group interactions at an institutional level to subjective/ideographic models.

I asked about [Ted]'s practical experience of studying Social Machines using an Observatory:

> "To understand the phenomenology and the process of Social Machines, you've got to have primary data and you've got to start thinking about capturing it and organising it and noticing it and curating .. SOCIAM has found the Observatory so useful, you know, how [else] are we going to understand these systems, unless we can pay attention to their vital signs and the interactions that are being thrown off by them.."

I concluded by asking about the ease of observing systems externally from their natural data exhaust vs. specifically instrumenting them to create data in useful formats and levels of detail:

> "It's an immaturity in the field. I think a lot of Social Machines out there would not have thought of configuring their services so that there was a well-formed API that could actually act as a set of junction points .. like what Facebook did when they engineered explicitly feedback to see how a population would respond and got castigated for it."

Ted implies here that observatory requirements may impact the design of future systems bringing us back to quantum effects of participating in the systems being observed and that by association, the choice of the measures will be driven by the motivations of the groups owning the data.

Ted Themes

Super-ordinate themes	Sub-Themes	
Competition	Competitive imperatives	The drive to compete between academic groups and/or between commercial groups
	Credit data vs. findings	Citation split between contributors
	Data "land grab."	Market competition for Web data
	Hugging data	Reluctance to share data
	given loss of privileged academic position	Growth in importance and capability of business research teams
	Disintermediation of academic research teams	Business doing advanced Web research

Appendix E

Collaboration	Collective capability	Need to share efforts to ensure breadth of analysis
	vs. Contracting out	Working together vs. working for
	In a shared body of knowledge	Building towards data and knowledge commons
	Politics of sharing	Drivers for retaining or sharing data
	new academic/business models	Engagement models between business and academic models
Data/Analytics	Instrumentation	Building detection/monitoring into Social Machines
	Portals/Analytics	Delivering insights inside/outside a platform
	Data ON vs. data ABOUT the Web	Distinguishing between inherently Webby data and Web-delivered data
	Standards	Conformance supporting interoperation and sharing
	Data-driven orgs	Feeding data as evidence into policy
	As robust proxies	Choosing meaningful data to represent aspects/behaviours of interest
	Openness/Accessibility	Key requirements for sharing
Ecosystem	Cyber vs. Physical	Border between Webby and Web-mediated
	Commercial WO	Google and others as WOs
	Gov data holdings vs. systems	What is held vs. what/how it is surfaced
	Repositories vs. Observatories	Storage vs. gathering/testing of data
	Open vs. Free	Accessible vs. zero-cost

	Cross-sector values/sources	Interoperation across different social groups/frames
	Role of the individual	Drivers and actions of ungrouped individuals
Analogues	WO as a Social Machine dashboard	Looking vital statistic of Social Machines
	WO as Open Data platform	Looking at open data sources
	WO as an active vs. passive scientific tool	Injecting changes into studied system before observation action/reaction
	WOs as personal contextual systems	Staging local/contextual system to further specific local aims

[Ivan]

[Ivan] is an experienced senior academic. He spoke with me about WO with a strong focus on what WO should ideally be in order for it to be distinctive and valuable and with a particular eye on the notion of WO as the Web (Science) Observatory rather than the more generic idea of observing/sharing any data on the Web. He argues convincingly that WO is an Observatory ABOUT-the-Web and not simply ON-the-Web and argues for pragmatic human-centric processes that will allow the WO to grow in breadth, adoption and automaticity over time.

Our interview dealt with ideas grouped into the following guiding themes:

1. Data
2. Definitions
3. Pragmatics
4. Web Observatory
5. Eco-systems
6. Perspectives.

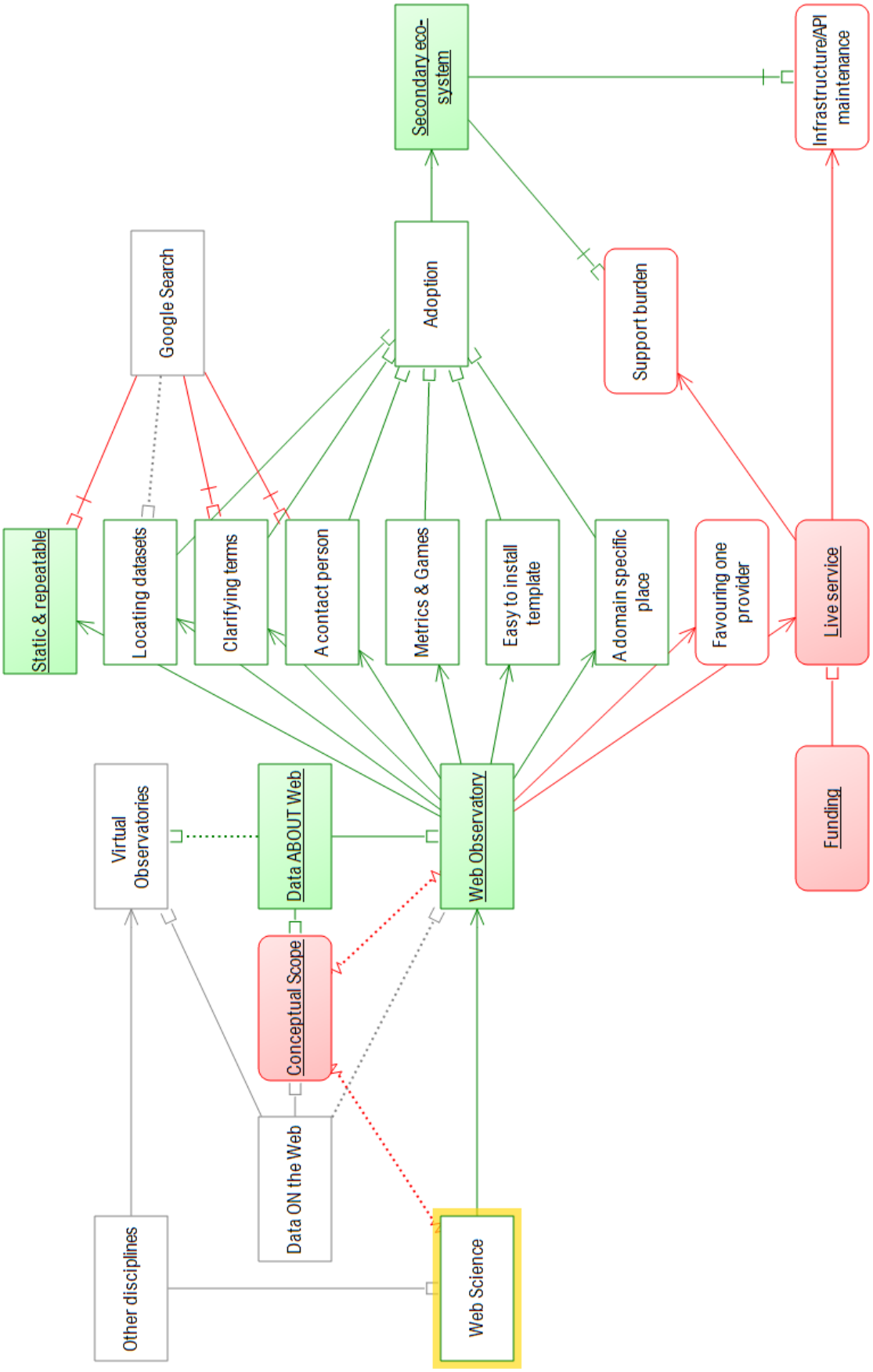


Figure E-3 [Ivan]'s Narrative

When I asked about what WO meant to him, we spoke a little about the history, but this appeared incidental and mainly [Ivan] conceived of WO as a direct and natural output of doing research i.e. WO as "a-thing-you-do-Web-science-with".

> "I think the feel all along has been that the goal is to enhance research by people trying to study the Web originally."

but goes on, with some discomfort, to acknowledge that *how* that is done has become more flexible.

> "I think we were moving more towards data about the Web. Now, I'm not quite sure. It seems like some are going one way, some are going another .. the vision is drifting a little bit, just because many, many more people are now interested in sharing data rather than specifically Web-related data."

This shapes the requirements of WO to align completely with the goals of Web science research:

> "I would be happiest if the Web Observatory [were] an affordance for doing Web science .. I would like to see the Web Science Observatory being a Virtual Observatory about the Web."

and while [Ivan] fully acknowledges the broader interest in data collection and other applications of similar technologies he creates a personal boundary between Web science and other uses:

> "I guess I differentiate making data accessible on the Web and things that really directly relate to Web data..I am a Web guy - that [other stuff is] just not the kind of research I do."

while arguing that other disciplines are already studying other non-Web topics leaving the Web to be studied in Web Observatories. There is a deeper concern implied here that broader inclusion of data without an explicit connection to Web Science questions "dilutes" the WO brand and perhaps by implication the distinctiveness of Web Science itself.

[Ivan]'s clear demarcation allows for a more direct evaluation of what a WO should (vs. could) be used for and [Ivan] uses the term "Web Observatory" throughout the interview to mean the system(s) that contain information of interest to Web Science research. Underlining this with:

> "Just because you're storing arbitrary data, it's unclear to me why you're in the Web Observatory."

He compares this to businesses engaging with shared research facilities in other fields

> "I've been a little troubled by the emphasis that this Web Observatory is going to be really useful to companies. Is WO really a company thing? Carbon [Observatory] isn't and nor is IVOA .. I don't know what companies are paying for Palomar .. I look at where it's succeeded, it's still academia."

[Ivan] feels there are more appropriate places to share generic research datasets:

> "But if we're just saying this is all about data sharing, [and] it doesn't matter what kind of data, [then] I start to feel like, why don't we just cede control to the RDA?"

Even the sorts of templates that the WO is based on such as Deep Carbon Observatory (DCO) are entirely distinct here:

> "Even though they're using some of the same standards and things like that, I would have a lot of trouble thinking DCO is a Web Observatory. It's a Virtual Observatory, but it's not a Web Observatory."

He makes a mildly sardonic (but compelling) point about collecting data together (in the sense of a classification) suggesting that if one approached the DCO with a dataset their first question would probably be:

> "Is it about carbon?"

[Ivan] is quietly suggesting that to include data that is NOT about the Web in a Web Observatory might seem inappropriate and even a little bizarre. From this he implies that the same pragmatic question be asked about the data to the WO:

> "Is it about the Web?"

> "How do you study a Web science issue? You need access to information about Web and Web use and that sort of data."

Though interestingly there is already a linguistic 'de-restriction' tacked on here ("that sort of data") beyond the specific declarative use of data 'about Web and Web use' so [Ivan] still leaves the definition open to broader "sorts" of data going on to qualify that Web Science itself may attract different goals/definitions:

> "I'm not going to define for you what Web Science is, and [so if] you put something on there [WO], and people use it - great.

Any division may be arbitrary and, to some extent, a simplification and [Ivan] acknowledges the blurred line between types of data that are (1) inherently about the Web, (2) the less distinct type which are Web-mediated and perhaps cyber-physical hybrids and (3) those which are inherently physical (non-Web).

> "The problem is it seems a lot of people are..saying the whole Deep Carbon Observatory is a Web Observatory."

Some examples are very clear for him:

> "Information on how people of different ages are engaging with technology, particularly Web technology - I'll take 'Web technology' broadly to include apps and things like that - strikes me as something that's very Web Observatory-ish [whereas] MOOCs are not WOs though you can have an MOOC Observatory to feed a study on how people interact in learning environments on the Web."

I touched on the distinction between AN Observatory vs. THE Observatory and asked if he thought the distinction and the "airbrushing" between terms was intentional (artful) or unintentional/unconscious:

> "I'd say it's unintentional because I see different people using it differently .. I think the current terminology is using 'Web Observatory' more like 'Website', in which case, we haven't got really the name for the 'whole' of them, except to call it THE Web Observatory: 'A' versus 'THE' is not a really good distinguisher."

[Ivan] has a clear picture of the difference between single and interoperating WOs though I feel he is not alone in struggling to find a memorable/meaningful name that would displace the ease with which user airbrush around the term Observatory:

> "I think the analogy to an Observatory is where some of this gets in trouble .. I think the reason for that has to do with Web Science, not to do with Observatories".. I'd be happier if we had a term like an Observatory platform, or an Observatory site, or an Observatory catalogue or something, and then the interacting set of those as sort of THE Observatory."

Notably there is a discrepancy between [Ivan] and [Imelda]'s understanding of the catalogue as being part of the WO versus being part of the Web of Observatories. I asked [Ivan] if he believed Google was a Web Observatory and his answer was "No": different from my own pre-conception and from [Ted]'s view. [Ivan] goes on to argue why:

> "The Internet archive .. is the same thing [as Google], except it says 'here's what it looked like at this day'. I can use that to study behaviour over time, to study changes, to answer questions about the Web..going to Google the search engine and getting my current answer strikes me as different than creating a dataset about people's use of Google or Google's answer to Web problems."

[Ivan] relaxes his earlier humorous data question "Is it about the Web?" and ultimately cedes that the distinction/hurdle is blurred but is not all-encompassing perhaps softening to a different question:

> "Is it helpful in understanding the Web?":

> "The notion of the Web Observatory to me was about helping people study this Web thing. That does include information about the Web and information about use of the Web, and that may include information "on" the Web, but it's not all information about everything."

Thus we are left to consider how many elements of the Web process (the content, the connections, the behaviours, the metadata) are required to understand the broader Web experience and it is interesting to note that [Ivan] specifically critiques the notion of conflating the Web with search

> "I actually feel that the notion that the Web is 'only search' has been a terrible thing that's happened to the Web."

It should be considered that while it has previously been argued, Brown (2013), that search differs significantly from observation this does not preclude the possibility that search technology (rather than the user-facing service) may be based on a paradigm comparable to a WO or that Google more broadly across its many services may be characterised as a WO (or OSO).

[Ivan] takes an interesting position on the near future of WOs in terms of automated/unattended access:

> "I still feel that Observatories, for a long time to come, are going to have humans involved..I don't see us in the near future doing lots and lots of live access to a lot of this data"

And he goes on to highlight the arbitrary complexity of processing data at Web scale chiding, to some extent, those who may expect this to be an easy thing:

> "To assume I'm going to somehow magically go to your Twitter data and pick it up and integrate it with mine, or your MOOC data, or your X data, or your Y data, strikes me as naïve given the current state of data technologies, not so much Web technologies."
>
> "It's about the metadata catalogue. There's not really an effort to put all the data together into one big database as it were. It's to help people find what they need. Then apps and things start growing around particular pieces of data.
>
> "I'm not really so big on curation in that sense because that implies 'liveness' and has costs and things associated."

[Ivan] moves here to hypothesise resource-intensive development/support of query APIs and infrastructure that is required AROUND the dataset, ultimately putting pressure on long-term sustainability due to resource costs to curate the infrastructure beyond the cost of curating the data itself. So when I suggested that some talk about maintaining Virtual WO environments to allow for the deprecation of databases, APIs, OSs and other technical platform dependencies [Ivan] smiles patiently and says:

> "There's kind of this myth that no one wants to share their data .. we have willingness to share stuff, we don't really have any funding to maintain the SPARQL endpoint .. you see very little data sharing in the database world .. because these things are hard as opposed to no one wants to do it."
>
> "These things die when the money goes away."

I asked then if WO was to be more of a social network than a technical platform:

> "It's certainly a Social Machine .. it's interoperability, sharing, finding, all the Web stuff. (1) how do they know that dataset exists; (2) how do they find it; (3) how do they know what the rights and things are ."

[Ivan] astutely and pragmatically illustrates the futility of systems that have no narrative or social connection to mediate the social exchange underpinning the search for data and no shared method/format for doing so:

> "I can arbitrarily publish it anyway and hope somebody finds through Google, but I'd rather be somewhere where people who may be looking for particular kinds of data about particular things know they're more likely to find there .. the problem is if we have a network of Observatories using a common standard so they can be inter-operable, but I can't FIND them."

> "I go to Google to search for something. If I don't find what I'm looking for, I call my friends. I'd like the Web Observatory to be sort of where I go to as the second step before I call my friends…"

[Ivan] likens WO less to a static file space and more to a GitHub model which keeps the data fresh by embedding it in a community:

> "I'm putting my data in a way that keeps it alive. Think code, right? If I want some code and I'm going to GitHub. I expect a live project

We spoke about the adoption of WO, the use of templates to kick-start and other complementary/parasitic pieces and [Ivan] raised parallels with the adoption of the Web

> "When I first went to the Web, you went to THE page at CERN .. there needed to be a centre of gravity to grow enough stuff that it can become non-exclusive .. in the early days of the Web, that was the CERN site. People went there, and that's where they learned about Web technologies .. it needs a centre of gravity. It needs a point of attraction to get started .. an easy set of software that you kind of install this, you put your data here. What your problem with that is turns out 'you put your data here' isn't quite so easy".

> "It [WO] needs a centre of gravity. It needs a point of attraction to get started."

Other systems and providers may, [Ivan] suggested, grow up around WO and it is not necessary for these to be provided by the "owners" of Observatories

> "Once people start using it heavily, then it doesn't really matter whether that same entity owns it .. Google never owned the Web [and] Search wasn't a primary mechanism until the Web was a certain size."

Ivan Themes

Super-ordinate themes	Sub-Themes	
Data	Curation vs. Metadata	Content or information about content
	Broad and localised	Multiple sources around a topic from a trusted source
	Dataset vs. data service	Providing raw data or a supported access method
	DaW vs. DoW	Data-ABOUT-the-Web vs. Data-ON-the-Web
Definitions	Web Research	Web science vs. research using Web tools
	Search vs. Observation	Comparing WO with Google et al
	Web Science	What is in scope and out .. Tied to WO
	VO vs. WO	Virtual Observatories in other disciplines
	WO vs. W^3O	One vs. many WOs
	WO	Web Hosted Observatory vs. Web Science Observatory
	Google	Search ≠ Observation
Pragmatics	Adoption	Enabling, promoting use of WO
	Centre of Gravity	Exemplar to kickstart WO
	Funding	Limited ability to support WO without ecosystem
	Essential utility	Discover data, describe it + terms (manual!)
	Politics/Power	Status and vested interests as a factor in adoption
	Location vs. Curation	Ability to find what's needed over automated interfaces

	Templates	Create ease of deployment + replication + education
	Metrics	Gameify/rankings to promote behaviour
Web Observatory	WO ← WebSci	Defining Web Science (hard) gives you WO definition
	Context	Importance of context/theme for WO vs. Repository
	Social Machine	The social group, convening nature of sharing
	Longitudinal, Static	cf. Changing Google filtered snapshots

[DataCo]

[Charlie]

[Charlie] is a senior technical manager. He talked broadly about his views/experience of the project issues in terms of the embeddedness of [DataCo] content and systems in a wider Web ecosystem. He was less focussed on the specifics of individual [DataCo] commercial offerings and more concerned with an existential view of [DataCo] as part of that ecosystem framed by a key assumption:

"The basic problem, in the long term, would seem to be that we can't compete with the Web."

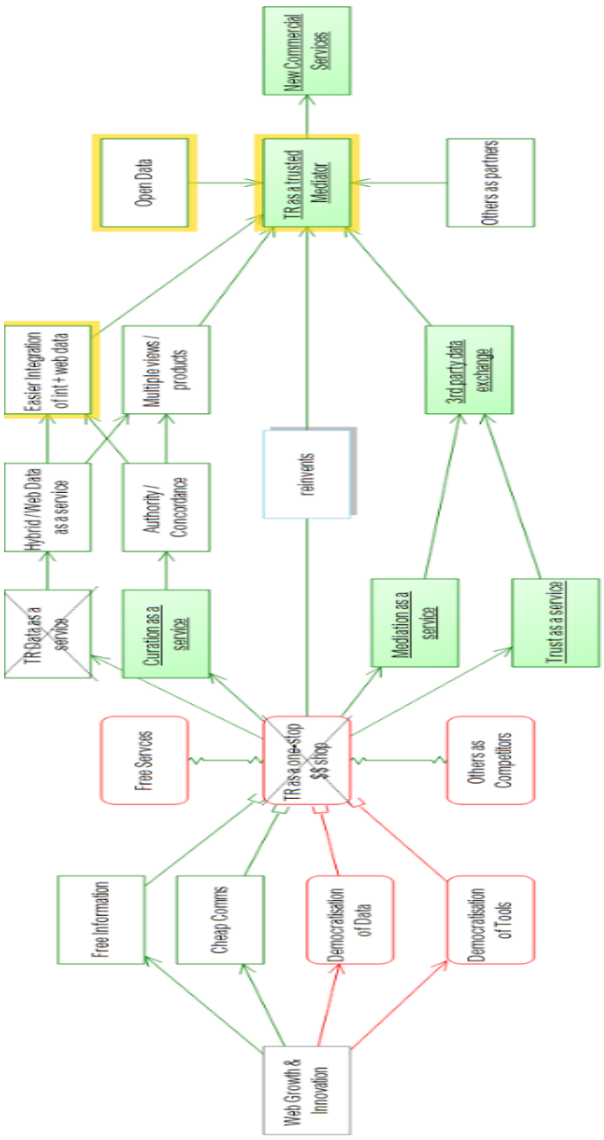


Figure E-4 [Charlie]'s Narrative overview

His framing of the [DataCo] experience is based on the idea of continued disintermediation and disruption of the status quo by the Web. He shows a high level of confidence that the sheer scale of innovation inherent in the Web overall will sooner or later address all the technical challenges that currently grant large corporates a privileged status within the information marketplace (e.g., privilege of scale/resources, of location, of license, of skills, of information). This view seemed in no way pessimistic but rather adaptive/agile and argued for the need to adapt the fundamental approach to market engagement:

> "It seems clear that the way we've done things in the past isn't going to be as sustainable or add as much value in the future."

The interview dealt with themes which I have grouped into the following guiding themes:

- Human engagement
- Data
- Web Impact and trends
- Value Models.

Having discussed the basic model behind the proposed WO [Charlie] spoke about how [DataCo]'s fundamental operations of collecting, and wrangling internal datasets had parallels to this idea and whether [DataCo]'s systems might themselves have to be extended to consider non-[DataCo] data from the larger Web ecosystem:

> "As we move forward, have to look a lot more at how we look at other sources on the Web and perhaps the way we do things is likely to change over time .. because of the existing Web and the existence of data that's exchangeable more directly between people who are connected to the Web and according to technologies and standards that are designed to make that easier."

This indicates how pervasive and disruptive [Charlie] feels the Web has become and, to some extent, how this might represent both an opportunity and a threat for the established businesses in the information space. [Charlie] foresees [DataCo]'s position and established identity as a:

> "..One-stop portal for everything that a professional might need.."

being potentially unsustainable in its current form in the longer term thus creating pressure for [DataCo] to re-invent its offerings and redefine its engagements both individually with clients and also via new partnerships - though the form of this it was unclear:

> "It's a bit of a question of [whether] it's on a formal basis with partners (..) or on a more open or informal basis simply between any interested parties on the Web."

This undefined future comes over as being grounded in [Charlie]'s highly pragmatic view of the 'relentless march' of technology and innovation on the Web and he suggests that even fundamental assumptions about the current business model may be challenged:

> "The better the Web gets, the hard[er] it's going to be for us.. we can still add value, but not necessarily by directly carrying the information."

This is a consequence of [DataCo]'s long years of experience in managing information curation and integration and exchange within their own federated structure:

> "That involves a lot of secondary things, especially things like integration, navigation, creating relationships. Also, doing things like managing rights and obligations and to some extent providing a certain amount of provenance information."

[Charlie] characterises sharing data sources at scale as being extremely challenging without standards for agreeing meaning:

> ".. To somehow coordinate what I do with what all other providers and all potential consumers are doing. If left to my own devices, that is an almost insurmountable problem ..[creating].. huge amounts of work because different publishers chose different standards or reference points.. From a technology point of view, if all of those other sources are reachable..but they're all using their own versions, their own means of identifying what they're referring to, their own way of describing quality, their own way of doing everything, then clearly the interpretation problem for the financial institution is incredibly high."

Thus allowing [DataCo] to support information exchange through *trusted curation services* where:

> ".. Institutions will want to take data which has been validated, which has been managed for quality, which has been integrated, [and] which provides an authoritative view of what's going on in the market. And that's what we do..if we can provide services, metadata, tools to those people, it helps them do more in common. Clearly, it benefits them in a sense that now they're having to do less work, but also it means that when the result of what they've published reaches our customers, it's more readily "integrate-able" in the stuff we produce, the stuff we do provide. "

This identifies the potential for central roles within a WO context to bring together disparate open data sources under canonical naming schemes whilst potentially linking these curated definitions back to other commercial services from providers such as [DataCo]. [Charlie] is painting a picture of offering the tools and services that [DataCo] use internally to manage their own complexity out to the market as a new level of service and value where data curation is becoming everyone's problem.

> "A lot of organisations have an internal challenge in terms of how they integrate and describe stuff. And some of that will move to more open methods, or publicly available methods can also become more prevalent."

Their technique has been to tame complexity through publishing/naming standards rather than creating a single centralised database:

> "We weren't centralising as you say. Instead, we were creating subject-oriented, content operations where the information was published using standards so that access [was enabled] to any of the information .. we weren't standardising database technology, we didn't have a single warehouse to collate all of this stuff into. Instead, each part of the organisation was working on its own specialised areas of information, and then publishing, and making available that information using the standards."

> "By creating an authoritative source for those things, it can act as a common point of reference."

Again this shows a strong parallel with the conceptualisation of WO as bringing together *access* to multiple disparate sources without storing/staging these sources in some notional vast centralised repository.

I asked whether there might always be "sacred cows" in the information space which would always require highly formal and authenticated processing - something that would always be reserved for [DataCo] to do. Even here [Charlie] tends to pragmatism:

> "I would tend to favour even those things eventually becoming a more of a ‘meshed exchange’ rather than mediated by a hub. Simply because the technologies [supporting] the Web will get better to deal with those problems anyway."

This idea of the "hub-and-spoke" giving way to the Web/mesh prompted a question about how market dominance and commercial influence might allow the biggest players to dictate the standards/tools that underpin the ecosystem standards he was discussing and whether key technologies would make the difference:

> "The basic problem is how to deal with meaning and who pays the cost. I think that's where the generic challenge is."

> "..There is an opportunity for us here in the sense of mediating an agreement between two parties, which can enable some form of commercial relationship in terms of that exchange, which does then carry obligation with it. But where we're not actually, necessarily carrying the concept..the shift is purely about where the organisational boundaries are. It's not so much what the technique is.""

Here [Charlie] is exploring the idea of assuming liability between parties in an ecosystem of previously unmediated, unvalidated data exchanges. We discussed trust-as-a-service having characterised the impact of the Web as having moved the focus from a lack of data (pre-Web) to a lack of information (pre-apps) to a lack of focus (pre-search) and currently to a lack of what is trustworthy.

> "Basically, the thing that stands behind the [DataNames] is not the format of the number or even the format of the URI..it's saying that the process and the quality of the method are the same for many different types of identifiable groups. And then by creating an authoritative source for those things, it can act as a common point of reference."

And it should be noted that these sources currently offered by [DataCo] for free may also be seen as an 'on-ramp' for other commercial systems.

> " ..There are power structures around different organisations and they may want to try to leverage power around their particular way of describing data … in the long term that .. starts to work against them as information in general starts to become more easily exchanged."

Using the example of Apple iTunes as a content (vs. a pure storage) cloud model, the parallels between music and datasets were explored. Apple themselves do not seek to produce all the content but rather to carry the content, mediate the commercial exchange and enforce (where possible) the terms of the content license. [Charlie] looks to a range of models where [DataCo] offer curation tools/services or, like Apple, carry and distribute content on behalf of others whilst enforcing license terms as required.

> "And if we put more effort into how we agree meaning all of us can reduce the amount of effort we have to do to understand what anyone's saying."

The value (truth) of a hybrid dataset may, however, be fuzzy or contextually based on the users requirements - users seeking early rumours/gossip about emerging opportunities will be looking for different sources and levels of curation than users looking for formally reported news or financial transactions. More curated data (filtered, processed, etc.) is not necessarily better - the choice however of a single, persistent identifier to wrangle multiple sources/references to potentially the same entity, say, a company (reported in rumour, news or official reports) needs to be disambiguated or "concorded" as [DataCo] put it. How each piece of content is then evaluated will depend on the usage and the provenance:

> "The on-going problem for them is they have to be able to identify all of the companies and the people, all of the significant entities that occur in a news story in order that they can provide analytics on them and describe the results to the beneficiary..It's 'a single version of the truth' in the sense that there is something for people to agree on.

[Charlie] talked with enthusiasm about human engagement - the role of people and expertise - in this process ranging from the creation to curation in blended solutions such as TAMR and particularly in ultimately contextualising the data for specific purposes and giving it meaning. Albeit that he did foresee that automation would continue to reduce the numbers of people involved in the mechanical curation process.

> "the Observatory, would [also have] the whole dynamic around how meaning is agreed. Because it's a very human thing, it has nothing to do with machines."

Ultimately [Charlie] sees the move towards a WO-style paradigm within [DataCo] as a natural extension to enterprise (federated) data management in which trusted parties, authoritative sources and partnerships based on open standards will reduce the costs of integration so that meanings can be agreed between diverse parties using diverse datasets for diverse purposes in an eclectic mix that goes beyond the curated data and services offered by providers today.

[Charlie]'s interview depicts [DataCo] data as becoming an important part of a wider opportunity and he focuses on the need for preparation, communication and facilitation of this opportunity.

Charlie Themes

Super-ordinate themes	Sub-Themes	
Human Engagement	Humans as producers of data	People's behaviour/choices/profiles as native input for data source
	Humans as processors of data	People transforming/correcting native data
	Blended models of processing	Machine learning + expert moderation as an approach to correcting/curating data
	Incentives for interoperation	Community contributions vs. paid crowd tasks
Curation of Data	A single version of the truth	Looking for "contextually useful" and semantically consistent cuts across diverse datasets
	Commoditisation of curation	Debate as to the level to which curation can be automated/operationalised and therefore loses perceived value
	Persistent Identifiers	The creation of non-system (super-ordinated) identifiers that are designed for use by machines to point to concorded entities
	Cost of integration	Poor naming and continuous bespoke integration work increases the cost of integration data sources
	Curation tasks	The arrangement, naming and alignment of data items using classification and curation
	Raw Data vs. Curated Data	Highlighting the difference in value between unprocessed and curated data

	Value in process vs. content	The perception that the organisation/enhancement of data has a distinct value from the value of the content itself
Web Impact + trends	Changing perceptions of value	In the process of disruption the perception of what is worth having, the cost/benefit and specifically which facets of a service are often transformed.
	Constant Web improvements	The observed progress over time of the Web providing innovations and solutions to previously unsolved problems
	Context collapse	The phenomenon by which previous barriers such as distance or cost of storage/production are made trivial though the digital nature of the Web
	Data, Info, Attention, Trust	The evolving process whereby the search for raw data changed (in the presence of large amounts of data) to a search for contextualised and interpreted information. This in turn gave way (in the presence of large amounts of information) to a search for what items might be most relevant (using techniques such as search). Ultimately in the presence of large amounts of sometimes conflicting information - the focus on which sources are "true" and which can be trusted.
	Disintermediation / Disruption	The process of technology-fuelled changes in market structures, capability and dominance in which earlier (e.g. physical) forms and methods are replaced by newer (e.g. digital) methods

	Hub → Mesh	The tendency of earlier centralised, monopoly style networks resembling a hub-and-spoke to be replaced with multiple hubs and ultimately a meshed network of (in this case) more equal/less dominant market participants.
	Sacred cows	Processes that are resistant (immune) to technical/cultural disruption due to the embeddedness of the assumptions around the current form of processing
	Web Data vs. Curated Data	The distinction between datasets found "in the wild" via the Web without warranty or liability and those formally offered by an organisation and notionally supported by that organisations reputation.
Licensing of data	Synthetic - Hybrid licenses	The notional license that results when two or more data assets with different licenses are provided as a unit to a third party.
	Non-open licenses	Restrictions placed on the re-use
	Licenses vs. Systems	The requirement for the meaning/impact of licenses to transcend what system they are produced on (i.e., That they carry over to be meaningful on other systems
	Standards Licenses ODRL	Open data rights language standard as a possible method of capturing rights and responsibilities across system boundaries
Nature of data	Structured vs. Unstructured	The distinction between datasets in databases and document sets in repositories
	Data ecosystem - Open/Closed	The class of data assets across which services and products are assembled

	Fuzzy vs. Formal data	The distinction between explicitly stated values and inferred possibilities/probabilities
	Hybrid/Triangulated data	Datasets constructed from two or more sources
	Many renditions/perspectives	The ability to construct multiple views of data from the same sources
Web "value models."	Content Marketplace/Surfacing	The perceived value in a service which locates and or surfaces/presents access to data sources
	Freemium model	The offer in which a portion of the data/service is offered at no cost while a more complete/functional service requires a subscription or fee.
	Indirect value models	Methods by which the provision of X creates value in area Y
	iTunes model of content + control	Apple's model of providing a content licensing + access system for which it provides little/none of the content itself
	Liability/Trust as a service	The notional value of dealing with a known counterparty with a positive reputation
	Manage vs. host vs. own	Models for [DataCo] in which they broker, host or buy/sell content
	Network Effects/Partnership	The increase in value which accrues to a service when the number of service users increases
	Sustainability	The method by which the current process (business) is funded such that the process is seen to sustain its own costs either directly or indirectly

Partnerships	Open innovation	Allowing technology to be leveraged (spun-out) to the market for better engagement and development
	TAMR	ML/Expert learning curation systems
	Services around Third Party content	[DataCo] building services around data bought in from other companies
	Alignment of Objectives	Making decisions which generate (mostly) favourable outputs based on what is measured for each group involved in the decision.
Quantum effects	Conspicuous Commentary	The tendency of behaviour to change and results to change when experimental subjects are (aware of being) observed due to the presence/influence of the researcher.
Sharing of Data	Cost of service - integration	The resources/costs involved in wrangling multiple data sources to form a service or product
	Publishing standard vs. storage standard	The notion of agreeing on the name/meaning of a data item vs. actually holding the item in central places and/or according to standard formats
	Privacy vs. Transparency in WO	The trade-off between the rights of individuals to not be surveilled and studied vs. the more useful nature of complete/detailed data
	Free Access vs. Free of Charge	The notion that users are free to choose to see a piece of data (potentially accepting a fee/charge) vs. the notion that access to the data item inherently carries no fee.

	Variety + Dissonance	The notion that the scale of the Web will inherently comprise different (and even conflicting) ways of doing things
	Repurposed/Segmented data	The notion that set of source data/documents may be re-package, filtered or otherwise prepared to suit a particular usage or customer
Trust	Trusted Data	Data items that are believed to be valid/true
	Trusted Agents	Counterparties that are believed to be who they say they are and who will do what they say they will do
	Provenance	Knowing where an item came from and some level of the transactions/transformations performed upon over time
	Quality/Reliability	The idea of fitness for purpose and freedom from processing error
Use of Standards	Format standards	Agreed ways of holding/transmitting standards
	Process standards	Agreed set of steps in processing data
	Internet/Web standards	Technical (W3C) standards
	Centrally agreed vs. centrally stored	Notion of the meaning versus the implementation of data being key to interoperation
	Market Power vs. Adoption	Notion that commercially successful (powerful) organisation may be able to push the adoption of standards owned by or produced by themselves

[Thomas]

[Thomas] is a senior manager. He appeared to engage with this conversation as a business-focused innovator, at times passionate about the opportunities that new approaches and techniques might bring. We spoke at length about several companies innovating in the WO space and the increasing use of hybrid (Social Machine?) machine learning + crowdsourced expertise solutions.

He is directly involved not only in finding uses for [DataCo] (meta) data but in promoting these solutions both internally and externally in order to convince business units to adopt them and embed them as part of the natural operating process in a sustainable/value-adding way.

His key themes related largely to the transition from traditional business models to newer types of framing and to some extent the tension between the contrasting views/assumptions underpinning such models and Figure 7-4 summarises his narrative.

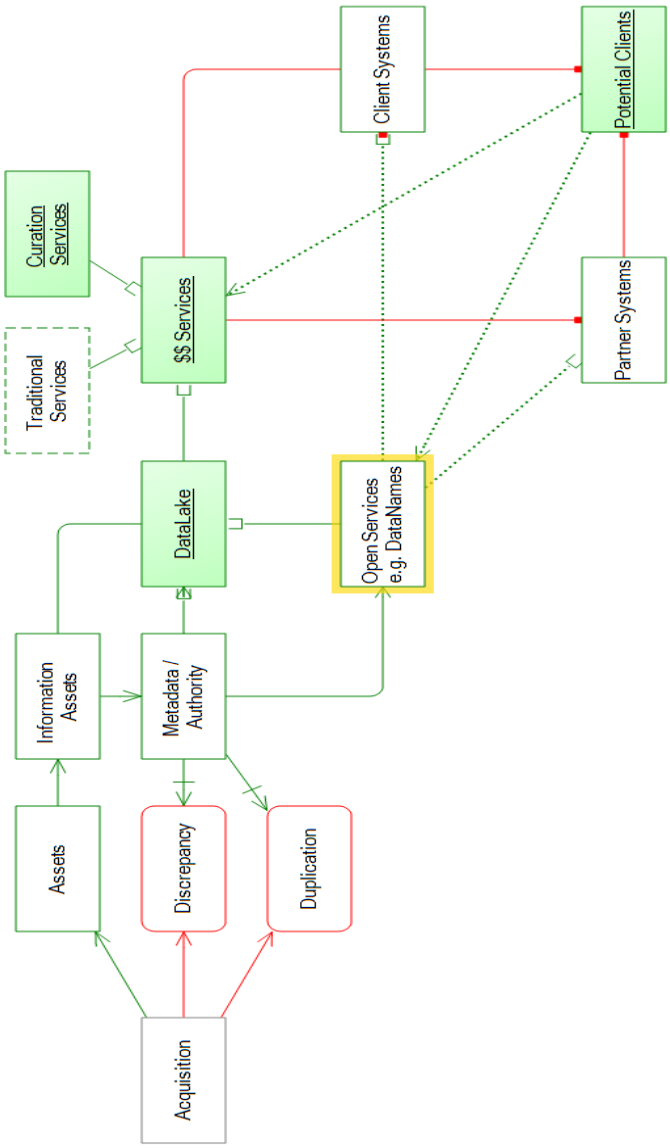


Figure E-5 [Thomas]' Narrative overview

[Thomas] started by situating the challenge in the [DataCo] merger, which unfortunately (but unsurprisingly) left the organisation with conflicting and duplicated data sources and an attempt to address this using data storage and querying techniques for:

> " Common metadata authorities..and the programme by which those are governed, called [DataLake], all of that came together following the [..] merger"

But [DataCo] realises that this is not a problem unique to their own operation

"We've got now these core authorities that we're using internally for identity, that's a problem shared by all of our clients, and generally on the Web."

"We .. ingest data, push it into a Hadoop infrastructure, get it into a triple store data warehouse and then push it out to [DataNames] through an API and some search interfaces. So we get to show people the end to end capabilities there."

And so a more strategic view of [DataCo]'s data in a wider context was formed.

> "We began enhancing the ability to connect our data to client data and client data to [other] client data, etc. Basically, create a connective substrate for all of this data."

Given huge volumes of data are now available and this data deluge attracts clients who previously worked only with carefully curated commercial data there is a new challenge to replace (or enhance) the commercial data with open data that has to be curated/validated or combine the traditional sources (in a meaningful way) with easily available open data to create hybrid views. [Thomas] expressed a core unresolved question around [DataCo]'s positioning going forward:

> "In operations we saw this trend happening in advance of the business folks seeing it, because I think they [Sales] are talking to end users who are used to more of a desktop sales process where the end user, an analyst or a lawyer, isn't digging too deep into where the data has come from."
>
> "What role does [DataCo] play in all of that?.. are we best off trying to own all slices of the value chain to deliver that through to the desktop? Do we specialise in connecting the data and helping the world do that at scale for professional users?"

For either strategy to be effective, both internal business units and customers must see the value of the approach and [DataCo] has dedicated a team to work on education and promotion for this.

> "[Working on the] strategy for [DataNames] as well as the conversations we're having internally to evangelise that initiative, and then the similar conversations we're having with clients."
>
> "In particular, folks in the ecosystem that we're trying to develop and the partners."

[Thomas] showed a keen awareness of network effects and community engagement with open standards and platform technologies which he described as:

> "A transformational strategy to think about putting these under an open data license to facilitate growth of them, to facilitate feedback on them so they just scale better, above and beyond what we're doing here".

Internal education is a key activity due to concerns around the broad cannibalisation of revenues by substituting free content for premium content/service - something about which the business has reservations and has called for some compromises around the "openness" of the project outputs:

> "So there's still apprehension about that. You've probably seen some licenses we used. We ended using a mix of creative commons that allowed commercial reuse and some that didn't allow for commercial reuse, but at least it got us out to the point where we are."

And now the service is live, the feedback and engagement phase has begun which prompts the desire for additional features and a broader scope of published ID's. [DataCo]'s strategy team want to surface the key debate over "the value of open" and how value may be measured in terms other than direct revenue.

> "If they [Customers] want that data, is it worth *giving* it to them in the interest of furthering the partnership and getting feedback on that data, getting adoption of it, etc? Or is this something we really want to commercialise? .. It was that kind of tension we wanted. "

[Thomas] sees that trusted metadata and curation services may be an effective path to new sources of revenue for [DataCo] given so much raw data is being combined by organisation with little experience in the disciplines and pitfalls of data management at such a large scale:

> "We're already seeing a lot more blending of data. If you don't have the provenance all the way back to source, which in [DataCo] is difficult to do, and I can't imagine what it's like in a place like Citigroup or JPMorgan … if you don't have that provenance."

> "Knowing the history, the paper trail for the data is challenging but absolutely necessary if you are know how it can be used "

I asked [Thomas] about combining data that belong to one of many [DataCo] companies, (an operational problem) versus delivering data come from six different sources using six different licenses. He immediately conceded the licensing issue:

> "Yeah, I think that's a huge one in all of this."

The ultimate intention here seems to be the creation of a more ‘collaboratory’ ecosystem of providers, each bringing some area of expertise and tooling to an overall ecosystem of end-users with specific needs for the complex, hybrid multi-source services and models they want to build.

[Thomas] is attempting to anticipate where the problems will be for clients as they attempt more complex roles within the information supply chain perhaps taking on the mantel of an information provider to their own clients/partners as [DataCo] does. If [Thomas]' team has predicted correctly, then [DataCo] will be there to add value and vertically integrate (from source to service) rather than only horizontally from industry sectors A-Z.

[Thomas] seems very focussed on the customer (both internal and external) and his role is that of the innovator, searching for fruitful uses and applications of the technologies that [DataCo] and their partners can bring to bear on the problem at hand while retaining and growing [DataCo]'s service profile with customers. This is a creative and entrepreneurial viewpoint about leveraging techniques and assets in a new way.

Thomas Themes

Super-ordinate themes	Sub-Themes	
Acquisition	Natural variation / Complexity	Geographical diversity and acquisition create an unavoidable variation and overlap in corporate systems and data assets
	Wrangling across BU's	Federated businesses are typically left with high levels of autonomy which does not easily fit with radical corporate-wide re-engineering
Business vs. Tech Ops	Alignment of Objectives	The measurement of business units in terms of financial revenues may misalign objectives between Tech Ops and Sales
	Commercialising Assets	The monetisation of information and techniques
	Education, Recruitment	Convincing internal / external groups to back the idea
	New models of business	Looking at the contribution to the business beyond financial revenue

	Network effects	(in this context) Scaling the effects of a system/approach beyond a single company
Data	Authority + Disambiguation	Creating a concordance for specific entities
	Data in Silos / Interoperation	Mobilising data beyond organisational boundaries
	Prov, Trust and Curation	Having confidence in the source and treatment of data assets
	Synthetic Data licenses	Combining data sources under different rights and obligations
Partnerships	Innovation Hybrid solutions	Leveraging Social Machine solutions comprising human expertise and machines
Tech Trends	Democratisation (Data, Tools, ML)	The broad availability of data, tools and techniques at a low cost of entry

[Quinn]

[Quinn] is a senior technology manager. He talked about the project partially in terms of opportunities for the business but more broadly in terms of the discipline of 'doing data right'. His viewpoint was pragmatic in terms of the tensions between technology, business and the demands of the market and was very 'data-centric', with data quality and curation being positioned as central to offering a credible and trustworthy service at scale. His key themes reflected the realities of corporate perspectives, conceptual beliefs based on experience of "how things are" and themes around curating, managing and structuring data from a technical perspective (Figure E-6).

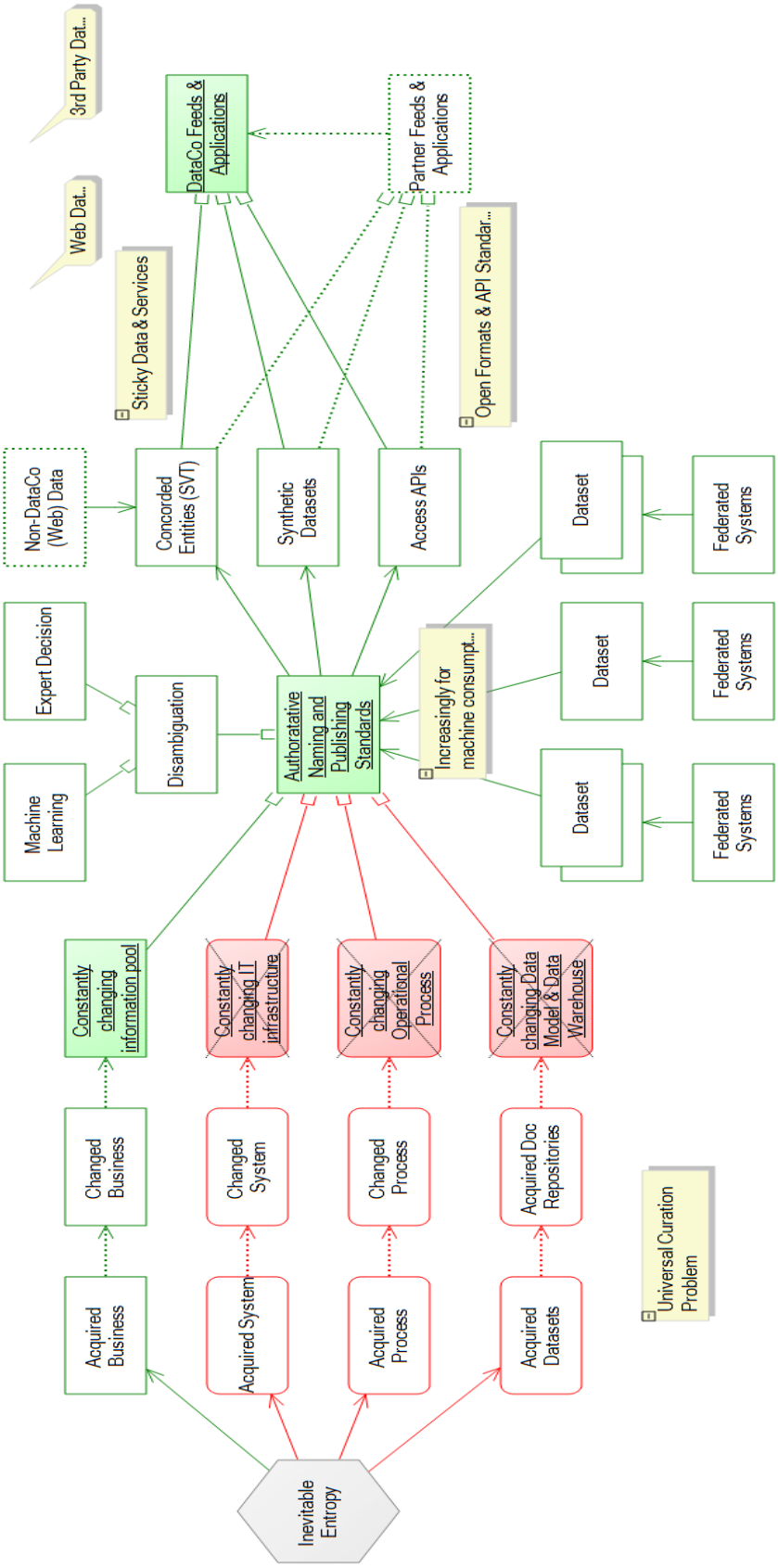


Figure E-6 [Quinn]'s Narrative overview

[Quinn] situated the project in terms of the challenges for the merged [DataCo] business when each part coming into the new structure had already previously acquired several businesses before merging operations:

> "[DataCo] is built through acquisition, so we have many examples where we bought a company and found that there were duplicate data sources within that company and we have to rationalise or concord those in order to deliver the value proposition."

> "With ever-increasing amounts of data available in the world and our clients demanding that we provide ways of integrating that data, we needed to start to be much more consistent about how we named our data when we gave it to our clients."

[Quinn]'s characterisation of an on-going need to manage repeated (endless) re-organisations of changing data assets within the group reminds us of the reality that organising (big) data from multiple sources is unlikely ever be 'finished' and that an agile approach of delivering datasets/services that are good-enough-for-application-X may be a more realistic ambition for WOs than expecting to categorise all data sources "once and for all". [Quinn] explained the journey that [DataCo] had been on:

> "While we'd like to think that we did that [merger] work [strategically], and we come up with overarching plans, [in fact] usually it was hacked to the desktop level."

> "So about five years ago we started a project to be much more disciplined about how we match the data and how we concord the data. And that was called .. [DataLake] and it was an internal project purely to solve the problem of multiple sources."

Something which reflects the challenges of marshalling multiple sources for WOs. The cost and efficiency of operation were at the heart of these changes:

> "We needed to change the cost of ownership of our data and recognise that our data was merely a piece of the puzzle, and so we realised that [DataNames] could be extended to solve that problem".

This view of the importance of data curation was primarily a technology initiative given the historical tendency to fix these things more tactically:

> "It was very much driven by technology. It was very much driven by our understanding of a need for data and information models, and an understanding of the way the world was going in terms of machine consumption."

Commercially [Quinn] sees the challenges of managing data-at-scale and variations between local representations as being inherently complex and the need to find "a single version of the truth" as a common problem shared by many large organisations.

> "I think that we already see plenty of customers who have this problem because of silo businesses or because of geographical businesses."
>
> "Our approach is one that's actually been built over 20 years of problems .. 20 years of hard mistakes learned and lessons learned. So I think yes it will resonate with different companies for different people, [the] problem of having silos across businesses is very common, whether you're built through acquisition or whether you just run your business from a geographical perspective or whether you are just poorly managed."

[Quinn] talks about the disruption and disintermediation effects that the Internet and the Web have brought to business sectors where the costs of providing comparable services (the financial barriers to entry) were previously very high:

> " 15 years ago it was hard for people to set up their own satellite networks, their own broad distribution capabilities - now it isn't."
>
> "it's about an evolution of the business model once you recognise that distribution is solved by the Internet."

And [Quinn] talks about the nature of that business model for [DataCo]

> "The distribution will be the easy bit. Referring [my emphasis Ed.] to things will be hard because there'll be so much noise, so much badly organised and acquired data. So the model, the value is in the data models and the naming mechanisms..Information infrastructure-models to wrangle the nature and meaning of noisy data are required."
>
> "it could be brokering .. maybe the future is [that] we are a broker of information access, but we don't necessarily distribute the information anymore."

which comprises a fundamental shift in commercial identity and business focus and one which appears highly relevant to commercial WOs and WO services.

Similarly, preferred techniques and approaches have moved on as the sheer scale of data on the modern Web makes previous approaches impractical:

> "20 years ago, you'd do a big ETL job, build a data warehouse, and you're done, but most people know those projects crash and burn long before they deliver value, and so people are much more sophisticated about it."

such that companies are increasingly employing solutions which are built for machines to consume the data and analyse it directly rather than prepare/render data for humans to analyse. [DataCo] themselves are employing a hybrid solution which combines machine learning with expert review using:

> "A company called TAMR who spends a lot of time trying to solve the problems of building data, building cross-siloed views of data .. the old style way to do that would be to build a single database. You'd have to do an ETL across those other databases and build. But these days, TAMR takes the approach that they'll do a machine learning algorithm to try and drive to concorded data."

[Quinn]'s *technical* view on the essence of this problem appears fundamentally grounded in the structure of the data rather than the nature of the content, the systems or the organisational boundaries:

> "As people start to recognise that data is the new software, the way people are going to derive products, derive value in the future, they're being much more strategic about it and then they wrestle with these problems which is how do we concord this stuff..?"

> "If you're serious about solving those problems you've got to do a much more precise job."

and [Quinn] stresses that [DataCo] has both the tools and the skills to apply to this problem and that

> "There is a huge amount of intellectual property around the way we acquire, organise, and present information .. and the integration tools and the models by which we acquire, organise and integrate that information are the mechanism by which people discover that as well."

This is something which will affect the competitive nature and charging structures for commercial WOs and the extent to which free/commercial services and sources will interact and blend with free/open data being carried by commercial organisations in order to add layers of trust, accountability and precision.

[Quinn] feels one of the keys to this data-centric model is the machine-centric nature of how the data is increasingly being consumed

> "If you move to a world where it's less about terminals and more about machines, if [..] maybe the data doesn't flow through our database anymore, but maybe we still provide the intellectual property that we've always done as a model and as an intermediary."

Using open data and open standards creates a problem for many businesses who base their investment and project decisions on purely financial measures such as IRR and ROI.

> "We also see the broader play around building an ecosystem of partners and customers that use our data in a way which is fundamentally 'sticky' and valuable so we don't have a direct ROI associated with the project, but we do have strategic leverage of the assets."

There is an apparent challenge or tension with this approach across some sectors of the business who do not have a nuanced appreciation of the opportunity costs or longer terms carry-over costs of tactical vs. strategic solutions:

> "I think that that's down to the fact that most business people don't really understand the technical issues associated with data management .. If you tell a business person, 'Don't worry, we'll hack it in the desktop', [they'll often say]: 'Well, that seems fine. Why would we not do that ..?' [Laughter] "

> "I think it's very difficult to measure ROI in that sense, but I think that, the ease at which you could say, Well let's just hack it one more time in the desktop."

It may be hard for [DataCo] to see a very different alternative future at a time when they are successfully pursuing the model they currently employ:

> "What does a modern information company, a modern information *aggregator* look like in, say, five year's time? .. [We] recognise that over time more and more data will come from alternative sources. We won't be a one-stop-shop the way we have been in the past, and so [DataName] is a way of solving that as well."

> "In five year's time, more and more information will not be in our data centre. More and more information will be available publicly in a way that clients will wrestle with how to manage that information."

What is the value of creating an open offering next to other commercial [DataCo] services? This is a difficult question for the [DataNames] team to answer. They seem to be looking to work more closely with customers and regulators to build more open solutions:

> "Open data, specifically was really just a recognition that if you tried to do something in a proprietary way that customers would look at that cynically as a way of locking you in

> "We'd had a long history of being very unwilling to open up what we called [PrivateNames], which was the original way to do naming which was really done for humans but gradually was used for machines as well and we've been very unwilling to open that up and so customers were suspicious of another proprietary standard."

[DataCo] also believe that establishing [DataNames] and other authoritative data sources will reduce the cost of development and integration for their clients by giving them access to high-quality authoritative sources of metadata:

> "It supports other engagements, whether it's making our data easier to use."

The specific impact of this approach is hard to quantify:

> "It's hard to measure the success in those terms, in terms of what is the ROI associated with doing this."

It is clear that [Quinn] feels that the more extensive/laborious curation approach is superior to using search which is not thought to be as precise or adequate a tool for this class of problem

> "Most of our customers demand precision and recall and provenance in a way that we can't do with simple search..Search on its own is not quite good enough because it's fuzzy and because it requires too much human cognition to make sense of the data."

> "Search solves a problem if you've got a human at the end of the day to do the disambigu[ation], to be the ultimate [arbiter]. But that's not the way the world is moving. The way the world is moving is towards machines doing the work.."

> "So yeah, you're absolutely right. That is why search will fail. So, of course, it doesn't do what we want, or our clients want."

In delivering this strategy to the business [Quinn] summarises the elements of the thought process and in doing so justifies the original spend as being a cost of business:

> "We originally started the project as a mechanism by which we solved an internal problem which no business person was aware of or really cared about."

He qualifies the project as an internal success

> "We've made enormous strides .. and it's being used as a method in which we join assets across the whole of [DataCo] now.

and moves to the justification of extending the benefits of the programme to clients which as a free-of-charge offering does make the tie back to other [DataCo] data and services more easily achievable.

> "We [now] see it as fundamentally making our data easier to use and therefore sticky..it isn't about saying that's going to be commoditised per se. It's about saying that the way that is monetised will change."

> "What's the benefit that we can bring to the whole rather than thinking purely in terms of [DataCo]?"

> "You know, it's not [purely] altruistic. We believe it changes the cost of ownership of our data which will increase the usage of our data. Most of the clients who use our data today are big clients who can afford to use lots of technology and lots of people to solve problems."

[Quinn]'s interview seemed highly focussed on building strategic solutions from firm foundations and principles, and I characterised this as an architectural view: [Quinn] does not seem to be particularly exercised by the specifics of the data, the sources, the applications or the different communities that may use them but does seem very engaged in the idea of the right structures (models), disciplines and processes that underpin both [DataCo]'s and the customers' use of the data.

This architectural view is designed to meet exacting requirements and to smoothly support current systems and to "weather-proof" them for future operations in an elegant and robust fashion.

Quinn Themes

Super-ordinate themes	Sub-Themes	
Corporate	Business View vs. Technical View vs. Customer View	Resolving priorities/measures of success for business, technology + market perspectives
	Disruption/ Disintermediation	The changing structure + changes in what is valued
	Openness	The growing size of, and economic models behind the open data ecosystem
	Power	The nature of (independence from) commercial control (Internally / Externally)
Philosophical	Complexity / Silos	Inevitably varied and silo'd nature of organisations and systems
	Non-financial valuation	How to measure the direct value (ROI) of actions when the return is neither direct nor financial
Technical	Data/EDM	Concordance as the central focus for Enterprise Data Management) vs. system-level, product level integration or centralised repositories
	Organism → Mechanism	The growing role of machines in not only producing but also consuming data leading to the insufficiency of the search paradigm
	Curation	Contrasting with search and employing hybrid models

Community Interviews

In the next section, we review three interviews with [Davina], [Gail] and [Ian] with each selected to represent a particular framing or perspective on the WO itself rather than on the core mission of the project.

[Gail]

[Gail] is an experienced business/industrial executive with a current focus on what WO can deliver to enhance the operations and services of the South Australian government. She views the WO as a learning and accessibility tool to liberate and automate the underlying data resources which she describes as "locked up" in current systems. She appears to focus on gathering data more broadly from the Web, combining this with internal data and synthesising new data/understanding. WO is an enabler both at a technical and social level bringing together different groups and individuals and marrying them to resources/services in an intelligent and insightful way.

Our interview dealt with ideas grouped into the following guiding themes (in E-7):

1. Conceptualising WO
2. WO as a learning tool
3. Data
4. Resource allocation
5. Social dimension of Web & WO

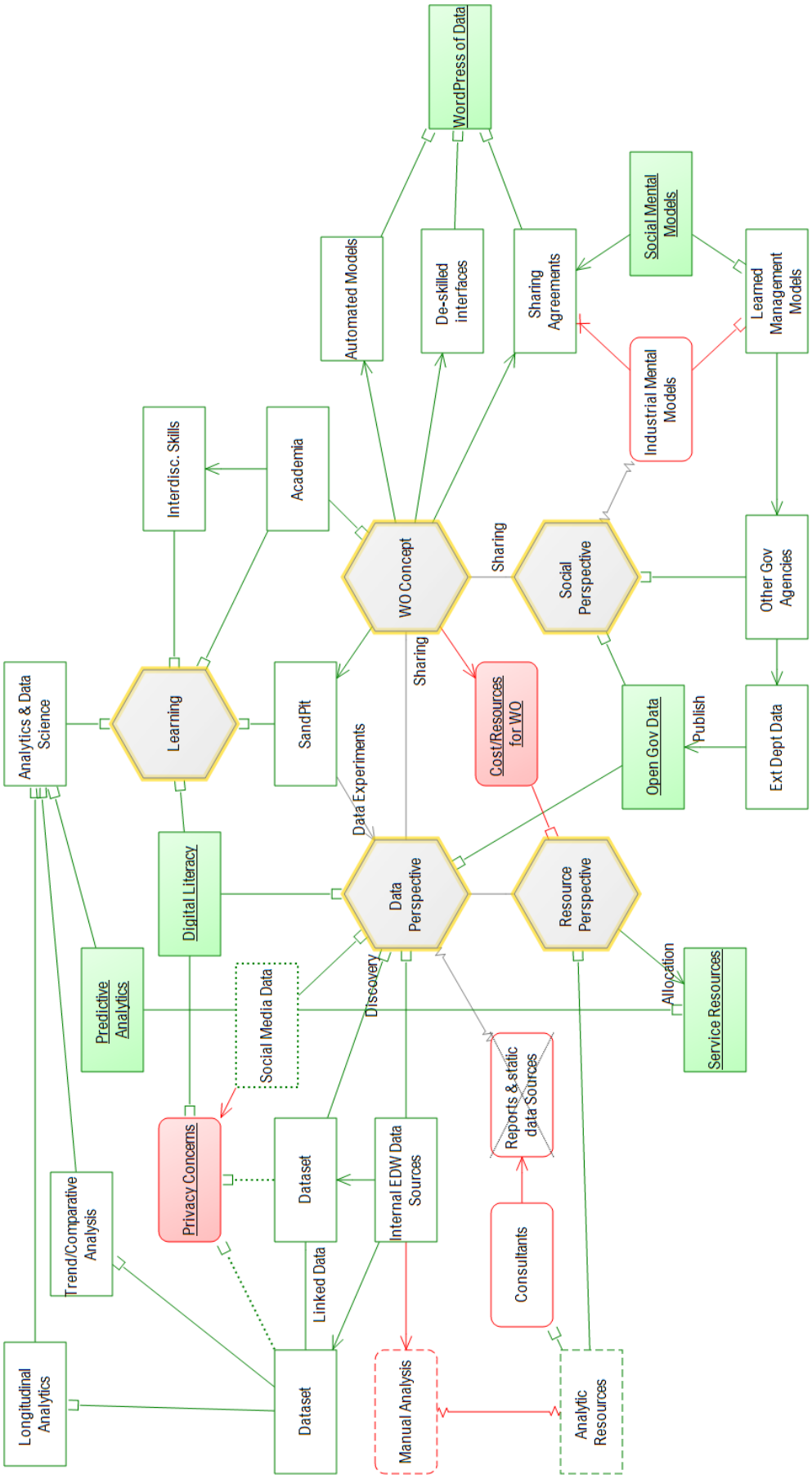


Figure E-7 [Gail] IPA Narrative

[Gail] frames her engagement with WO with the context of her role within government saying:

> "The journey for me around the Web Observatory [is] what we think it might be able to do for us as a government agency."

I didn't form the impression that she saw WO simply as automating or refining existing services (though re-use and speed were potential benefits) but rather that this opened a new way of working and offered data at a different scale. She related how she had learned about WO during a visit by Wendy Hall to Australia and how, despite her experience in both industry and government in the use and deployment of purely technical solutions, how impressed she had been with how WO had been characterised as:

> "..very much about the Social Machine, the anthropological and the sociological aspects of the Web and not just the technical."

I asked if she were particularly conscious of a link between the social and the technical in her own work she confirmed this based on her experience industry:

> "Through the microcosm of being within a business and realising the behavioural shifts that the technologies were enabling ..I did some knowledge management work .. and it was less about the technology and more about the behaviours, [but] the behaviours were enabled <u>because</u> of the technology shift .. [I see] the interaction between the technology and the people as being quite fundamental."

Thus the embeddedness of "social" in technical contexts runs through [Gail]'s observations around the challenges her department and government in general faced - often reducing to two things: ‘data’ and ‘people’.

She talked about the data challenge within her own group and seemed to express the idea that they could (should?) be doing more in a sense that ‘the bar is being raised’ in other areas:

> "We get a couple of million calls a year in relation to issues […] the analytics work that goes into .. those areas: it's deep, but it's very manual .. We have (sort of) trend reports, but it's at a very [basic level] - it's not with any scientific rigour. "

She stresses the sheer scale of research effort that goes to support evidence-based policy within Government and there is clear frustration here as the resulting outputs often don’t deliver underlying data and are therefore hard to analyse and cannot easily be re-used or extended:

> "Government is using .. all of that evidence-based decision making, and we spend a fortune on research reports and surveys of particular sectors, and I shudder to think

> how much government spends on it .. I started pushing with people that we actually get the raw data .. and start capturing it into something other than a Word document with the consulting team analysis .. normally when we do those reports it's a one-off, and then if you go out again two years later, you start again from scratch."

[Gail] is focussed on WO as an opportunity for a much more flexible use/re-use of their own data than current systems allow in order to build longitudinal views and enhance then from multiple sources:

> "Our data is still locked away. We're doing some data warehouse work, but up until now it's been very much locked away in our system ..[whereas].. if you've got this [WO] infrastructure and the datasets, we can just keep adding to it and it can become more sophisticated as time passes ..[and].. connect it with other departmental data."

[Gail] characterises WOs potential impact as much wider than a single department, as:

> "A whole of government opportunity for us .. [which is] .. the opportunity to extract it and then make it available and then connect it with others like Australian Bureau of Statistics data, Department of Employment data, Tax office data."

Critically sustainability (via impact and value for money) are seen as key criteria to enable an evolving WO.

> "The beauty of it is that if we can set it up and it's effective and sustainable .. you can keep tapping back into [and] adding to it."

Indeed funding is flagged as a potential barrier to this objective with education (both in terms of skills and general digital literacy also flagged as strong contributory factors:

> "[Data Science is] not a skill set that we're going to deeply invest in any time in the near future. That is partly because of budget, but partly because of the lack of understanding .. I see it even in younger generations of senior executives over here, they're still not as [digitally] literate as they should be - they're just not familiar enough with it and the skill set of applying the models."

While [Gail] sees there is a deficit in their own understanding around the uses and applications of digital data she is not arguing for everyone to become a data scientist:

> "There's no way a government [agency] is going to match the kind of skills and capability .. because the cutting edge stuff that the academic institutions are interested

> in [is] set up to a structure to sustain and keep going .. We don't have to become data scientists; we just need to know the problem we want to solve."

This speaks strongly to the way [Gail] sees her own agency and other government departments being (tactically) funded in comparison with the notion that academia has longer term (strategic) funding. It is worth noting that wanting access to such skills/tools without acquiring internal expertise creates the need for partnerships either between Government/Business for commercial services or Government/Academic for an academic engagement with government policy. The politics and economics of sharing are potentially complex.

Underpinning this idea is to be able to share datasets between systems on a broader basis (W^3O) without having to develop a single centralised WO system that she fears would simply become another walled garden. They should instead be looking outside their own agency both for data and skills:

> "Tapping into other datasets and indicators out there and global trends, matching it against UK data or US data, that's when it's to my mind fulfil the huge promise .. [and using] .. Academia for those particular skills around data and visualisation and the mechanics of making sense of data." This is particularly relevant when not only the technical infrastructure that an organisation has is incompatible with the kinds of integrations required but also where the mental models do not foster what is needed to succeed within a modern Web ecosystem.

She points to new young employees already skilled in Web-oriented collaboration and crowd communication lamenting that:

> "Those GenX's coming in (GenY's actually), who were used to collaborating and used to being very social - they'd walk in the door and we'd give them an e-mail box and a shared drive and [say] 'There you go, knock yourself out, I hope it works for you' "

Outdated mental/technical models such as Australia's patchy success with BYOD (Bring your own device) has often prevented users from accessing or being well integrated into existing corporate systems so that users experience *less* transparency/collaborative communication at work than they do at home in private. Willingness and ability to share data therefore go hand-in-hand for [Gail].

> "Bring your own device over here has probably possibly failed to live up to expectations to be honest, because of a lot of legacy infrastructure and as you say, risk aversion especially in government - 'you can't put that on my network!'."

For [Gail] choosing to invest in organisational infrastructure to enable corporate/private integration combined with promoting sympathetic attitudes and training to social working and security are the key social/cultural factors in an otherwise purely technical problem. Once these cultural barriers are overcome she sees potential in working with social data:

> " … there are social media sources that can be proxies for what people are talking about - for what people think ..[though].. I don't know that we understand enough of what we can do with it, so ..where can we go .. so that we can experiment without having to make the massive investment ourselves, and we can learn along the way."

[Gail] agreed there was also an aspect of using the WO as a *bridge* to academia as much as to access specific data resources.

> "We've got to get a whole lot better at understanding data and using data, but to my mind, the Web Observatory took it to a whole new level, and there was an opportunity to learn so much from an involvement with it."

There are fortunately already examples of departments looking to leverage data-centric approaches. [Gail] cites a police group who are:

> "Very interested in data analytics around predicting work injuries and management of work injuries, because that's a real issue for them ..[there is].. Strategic potential around where government is spending its money ..[with].. the huge advantage that I see is the potential when they all grow and start interconnecting."

Fundamental to this process is ensuring ease of access and ease of use for those without deep technical skills - [Gail] creates an interesting meme when she calls for WO to be as easy and dominant as tools in the Web publishing space

> "It [WO] has got the structure and the potential to be the 'WordPress of data'."

[Gail] sees social processes such as education as key to promote the willingness to preserve and enhance data. Her WO offers simplified access, seamless integration across group boundaries and non-technical analytics on diverse, distributed data sources to guide better policies and better government.

Gail Themes

Super-ordinate themes	Sub-Themes	
Conceptualising WO	.. as a generic solution	WO as an approach within Government
	.. as a specific solution	WO for targeted problems/services within Government
	.. as an academic system	WO as an output of and contextualised tool for Academia
	.. as an investment	WO as commitment of time/resources/funds
	.. as an integration hub	WO bringing together walled data gardens
WO as a learning tool	.. for digital literacy	WO helping to teach digital literacy per se and also to apply digital literacy to problems
	.. between disciplines	WO creating bridges to learn interdisciplinary techniques
	..for social media analytics	WO as a platform to study social media data and learn techniques for analysis
	.. for skills transfer	WO as a bridge to academic expertise in data science
	..with sandpit metaphor	WO can offer known (non-sensitive) datasets as learning material to "play" with
	..deskilling complex processes	WO as the "WordPress" of data
Data	.. in a comparison or benchmark	WO providing insights from the difference/similarities across datasets/sources

	.. in a trend or series	WO using longitudinal (incremental) data for insights on changes over time and future outlook
	.. trapped in docs / reports	WO helping to mobilise data in system/machine-oriented models vs. doc (human oriented) formats which are difficult to analyse automatically, reuse incrementally or share with other systems
	.. linked to other data	WO leveraging semantic Web and linked data formats
	.. made easily available / visible	WO as part of the democratisation of data and tools
	.. in a Web of data	WO helping to leverage both datasets on the Web rather than datasets created from the Web
	.. in Automated analyses and models	
	..and its metadata	
Resource allocation	.. from Web data	Using proxies from social media and other Web sources to focus activity
	.. from a proactive vs. reactive stance	Using traces to pre-plan/allocate activities rather than react to requests
	.. from a predictive model	Using data to simulate/model issues before they occur
Social dimension of Web & WO	.. in collaborating	WO can bring groups together around a single issue/challenge
	.. around the privilege of funding to engage	WO can require a substantial investment of resources to implement

	.. in disrupting	WO can change the typical dynamics of who knows what and has the insight to take action
	.. in framing	WO fits with a contextual framing and is adapted to it
	.. privacy	WO may expose identities personal data more easily than other methods
	..of learned social models	WO may transcend older industrial age models of value
	.. using social data sources	WO may leverage social media sources for insight
	..dominating technical dimensions	WO may be affected by social factors which prove to be more impactful than technical issues

Appendix E

[Ian]

[Ian] is an experienced strategy consultant focusing on supporting the strategic goals of the South Australian government through open data and smart city initiatives. WO has a palpable cross-over for [Ian] as part of an architecture comprising sensors, open data, open licenses and analytics underpinning "an advantage for South Australia in data". He views the WO as a potential lever: part of an integrated set of processes and techniques representing a solution to the question he is asking:

> "How do you take the concept of open government data and expand it to societal proportions?"

This characterises him primarily as looking for sound methods in that he is less concerned about what the data is (as long as it openly licensed) and more about the pipelines (and the technologies they use) leading to the agreed goals (rather than what the goals specifically are). He appeared to conceptualise WO in a highly developed and strategic way, looking to integrate WO into several aspects of his proposed solutions. There is a contextual (occupational) framing of [Ian]'s WO - something used to create specific solutions for the policy innovations of his group.

Our interview dealt with ideas grouped into the following guiding themes (Figure E-8) comprising:

5. Conceptualising WO Models
6. Data Focus
7. Teaching/influencing
8. Impact Focus

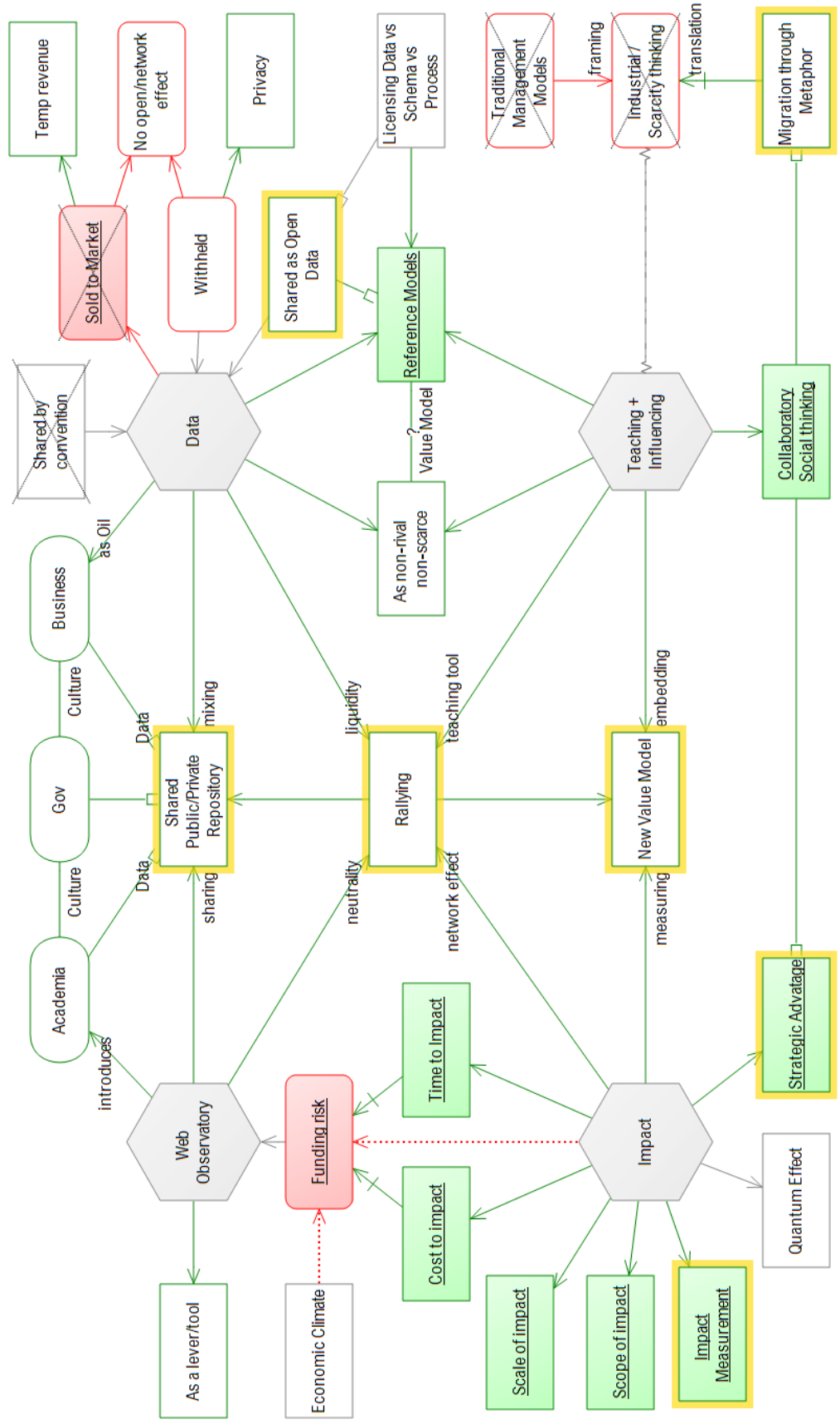


Figure E-8 [Ian] IPA narrative

I started by asking [Ian] what his interest in WO was and how he felt it would contribute to his work:

> "The Web Observatory, the archetype and even the title - the Observatory - grabbed me instantly and I thought there's a play in there that we need to be involved."

and based on the archetype he framed/contextualised it as a natural extension of earlier work around the release and licensing of open data which represented:

> "A fundamental shift in our thinking and our position with respect to data to whom that might be of value and in the concept of ownership and licensing."

His broad vision was expressed as:

> "Taking the open government concept if you like and expanding that to societal scale where public data and private data are shared in a meaningful and respectful way and that data is being increasingly generated by a smart interconnected devices (IoT)."

This is a non-thematic, operational definition which characterises [Ian] as a solution-builder for his group leverage the technical approach for sharing and combining data. He had previous experience of confusion (and political 'fiefdoms') around the ownership vs. the control of government agency data whereby each party:

> "Thought that the data they generated belonged to their agency .. in actual fact all the data generated by the South Australian Government is owned by the Crown and not individual agencies."

This was not conducive to agreeing the process and detail of licensing datasets across so many stakeholders and hence the ownership of the data was asserted centrally such that:

> "By centralising ownership we were then able to make an essential decision about the licensing and then once we put in place a programme and a capability to help people to understand how that applied to them - and in actual fact protected their copyright."

Ultimately this had a positive effect such that agencies:

> "Adopted it rapidly because it unblocked problems for them and it had really interesting side impacts .. like it became suddenly much easier for agencies to share data between them by releasing it openly."

This highlights a core issue for WO in which many diverse license types may be represented by numerous data owners and for which WO may wish to promote standardised licenses to reduce the friction of complex hybrid agreements.

Their intention in reducing friction around publishing had been to treat data as something to be refined into more valuable products:

> "What makes society and markets work is the free flow of commodity and information and data as a commodity to all ingredients it's right near the base of the value chain(s) which when you act upon that stuff you can do really interesting things ..[it's]..a raw ingredient like iron-ore out of the ground you can turn it into buildings and turn it into bridges and anything you like and data is a very similar kind of story."

Given fundamental economic differences between traditional rivalrous goods and digital data I asked if this manufacturing metaphor was widely used/encouraged:

> "In one sense data is to the information age what oil was to the industrial age or coal or steel. It's the new oil without which the information age doesn't exist..[but in another]..It's an industrial age paradigm of supply and demand ..[and].. it's an artificial construct that I'm using."

I explored what the impact/value of metaphor here was in moving towards the agencies goals:

> "Ultimately I work in a bureaucracy which is an industrial age mechanism .. in a strategy job which is at the horizon to be able to take contexts and explain them in a lexicon that our leadership understand and can, therefore, act upon, I need to use stories and metaphors and analogies that they're comfortable with."

And so we discussed that an element of the process around introduction such systems was a process of translation and framing for colleagues who:

> "See things based on mental models that were established in the past and are more relevant for a different time."

I asked if the Metaphors/Models were essential (or just helpful) to bridge thinking between old + new models and [Ian] talked about the need for speed in agreement:

> "Trying to find a common language so that I guess I don't need to wait for the rising tide of this trend to sort of lift all boats ..[and].. I need to engage <u>now</u> if we're going to find some comparative advantages for South Australia in data."

> "I'm also using the opportunity to inject new terms and phrases gently into the conversations and dialogues so that I am teaching - but slowly - because it's really hard to teach people who learnt their craft [mental models and patterns] in a different time and era using different management techniques"

I asked how moving from a more competitive/closed to a more collaborative/open mode would help achieve the intended goals:

> "My aspirations for the Web Observatory are to provide a point of reference, a capability that will enable us to do two things. One is to observe. So I would like to be able to point the Web Observatory at online activity and have it shed some light on what it means .. The second thing is that what underpins the Web Observatory is the kind of public-private data sharing capability. Now that is different than what we have .. there's no way for private sector data and community data to be mixed together on that platform and for it to be made available. It's purely government."

Noting that these aspirations were generic/operational I asked what such a capability would off beyond his interest in the management/control of the data:

> "[We are looking at] policy outcomes, better decisions about business, where people want to live. Anything you could imagine."

[Ian] is far from confident that a WO of open data will be without challenges based on a number of key factors:

- Budget security
- Depth (liquidity) of the Open Data market
- The perceived need to share through OD agreements
- The level of internal re-use of Government OD

> "Our open data movement in South Australia is [based on] goodwill and a sense of feeling like it's 'the right thing to do'. But as budgets shrink and demands increase - and we've got some tough economic times in South Australia at the moment - those things that fall into that nice-to-have bucket which .. tend to be the first thing that are stricken from the work plan."

Additionally, since a certain amount of sharing pre-dates the idea of open government data not all agencies see this approach as novel:

> "There are existing memoranda of understanding and agreements in place for data sharing between agencies that were in place long before the open data movement kind of took hold here. So there are still a bunch of agencies that are just 'doing what they've always done'."

[Ian] sees the true potential of open data is far from being realised due to slow data release though the reader is referred to Ch4 where the analysis of UK Open Gov data seems to demonstrate a 'long tail' structure:

> "We've only got a small amount of data that's publicly available .. less than 1,000 datasets ..[and].. I would say that we're not actually releasing our high-value datasets yet .. things like location data, data about conveyancing and houses .. they're still currently sold piecemeal to the market, and that's not yet been cracked open through the open data movement."

In contrast to the UK government reports of high levels of ingesting its own data [Ian] reports:

> "The South Australian Government in the open government context is a major supplier of the data probably more than it is a consumer of its own data."

So at an occupational level [Ian] focuses on the arrangement of technologies, standards and licenses in order to open up a new dimension/axis of analysis for his agency whilst flagging that there is still much to learn:

> "We are in a place and time, where we can measure things in society like behaviours, [at] a resolution and time scale that we've just never been able to in history .. we don't have models that help us to understand how that works .. we're in kind of uncharted territory .. We understand the economic multiplier effect through research, but it's really hard to understand what the impacts are of taking a piece of data and (if it's priced) copying that ad infinitum. What's the impact of that? We just don't have strong models for that."

More broadly [Ian] not only senses that digital data is different in terms of scale but that the economics/politics of digital society may be profoundly different and he is flagging a number of core Web Science research themes in his assessment of what needs to be understood in order to govern a digital society effectively.

> "What a Web Observatory could do to understand the impact of what open data at societal scale might look like and how tipping more data in, releasing it, shifts, ebbs, flows, tides and waves of data might materially impact things in the real world .. I'm

quite passionate about .. the concept of tipping all this data into society …[and]… how do you know what's going on with that data?"

Ian Themes

Super-ordinate themes	Sub-Themes	
Conceptualising WO	.. as a capability	WO as a tool for Government to apply
	.. as a starting/rally point	WO as a way to engage around an issue
	.. as an academic system	WO as an output of, and link to academia
	.. as a means to an end	WO a specific solution to a specific problem
	.. as a "nice-to-have."	WO being non-critical and under funding threat during austere times
	.. as a marshalling tool	WO as a place to gather datasets
	.. as a mixing bowl	WO as palace to mix datasets
	.. as a joint/valve between public/private data	WO a way for public/private players to control and regulate access to their shared material
	.. as a tribal/cultural meeting	WO as a neutral space to mic techniques and culture

Data	.. as the new oil	WO as central to the idea of data underpinning the modern information economy
	.. as non-rivalrous	Understanding the economics of zero-cost perfect copies
	..vs. schema/process	Underpinning the commercial idea of where value/IP comes from
	..as "lubrication" for transactions	Secondary oil metaphor for transparency fuelling economic activity (despite the idea of value through information ASYMMETRY)
	.. licensed openly	Data released under suitable license terms allowing for re-use and clarity of liability
	.. as both output and input	Referencing the idea of releasing raw data as open source and then re-consuming elsewhere as raw data within government and/or as derived (upcycled data) from a 3rd party
	.. ownership of	Referencing issues of public domain data and ownership vs. control
	.. as valuable/saleable	Referencing the opportunity of government to derive revenue from the sale of certain datasets
	.. from smart cities (IoT)	Referencing the idea of IoT Observatories

Appendix E

Teaching + Influencing	.. analogy/metaphor	As a method to bridge (frame) older-style models of thinking
	.. away from industrial age thinking	The intention to educate internally and raise digital literacy and new models/opportunities
	.. by translating to shared ideas	As a requirement to create "a journey" from existing value models and approaches to digital concepts
	.. about scarcity vs. digital models	In order to frame data and non-rivalrous and infinitely copyable and shareable
Impact	.. on business decisions	Referring to the cross-sector advantages of WO
	.. of quantum effect	Referring to the closed-loop impact of feeding back results (inputs) into societal systems which make the "observer" part of the system
	.. for strategic advantage	The use of WO for insight to take advantageous action
	.. on reducing friction	Transparency as a path to reduce cost, delay and unnecessary processing/queries
	.. on economic multipliers	Referring to network effects from broad adoption and participation
	.. on policy outcomes	Referring to the use to cost-effectively plan and measure the impact of policy through digital data sources
	..on real world challenges	Referring to societal challenges vs. academic research

	.. on big picture thinking	Encouraging longer-term investment in data systems/assets and keeping them open as a broad stimulation of the economy rather than short-term sale of data assets which excludes those without funds.

[Davina]

[Davina] is an experienced executive in the business, government and non-profit sectors. She views WO not only as a tool for observing Social Machines but as an example of a Social Machine itself - orchestrating people around concrete processes/outcomes and cites the example of "Government-as-a-Social-Machine" (though this is perhaps a broader definition than Berners-Lee's). [Davina] is predominantly focussed on outcomes - placing less emphasis on the specific data/inputs and more on the impact for the communities she is serving. She is very concerned with accessibility/ease-of-use as ways to address friction, barriers to adoption and retention of users. There is a sense of "assembling" rather than evolving solutions here which echo's a more high level architectural view but [Davina]'s views here are pragmatic and meritocratic, quickly delegating ("I'm not interested in the technical side") the responsibility for the technical components to those best suited.

[Davina]'s WO assembles people/systems through the Social Machine metaphor to create the most efficient path to the desired outcome (innovation). Our interview dealt with ideas grouped into the following guiding themes (Figure E-9):

1. Adoption
2. Education
3. Framing
4. Funding
5. Impact
6. Social
7. Web Observatory

Appendix E

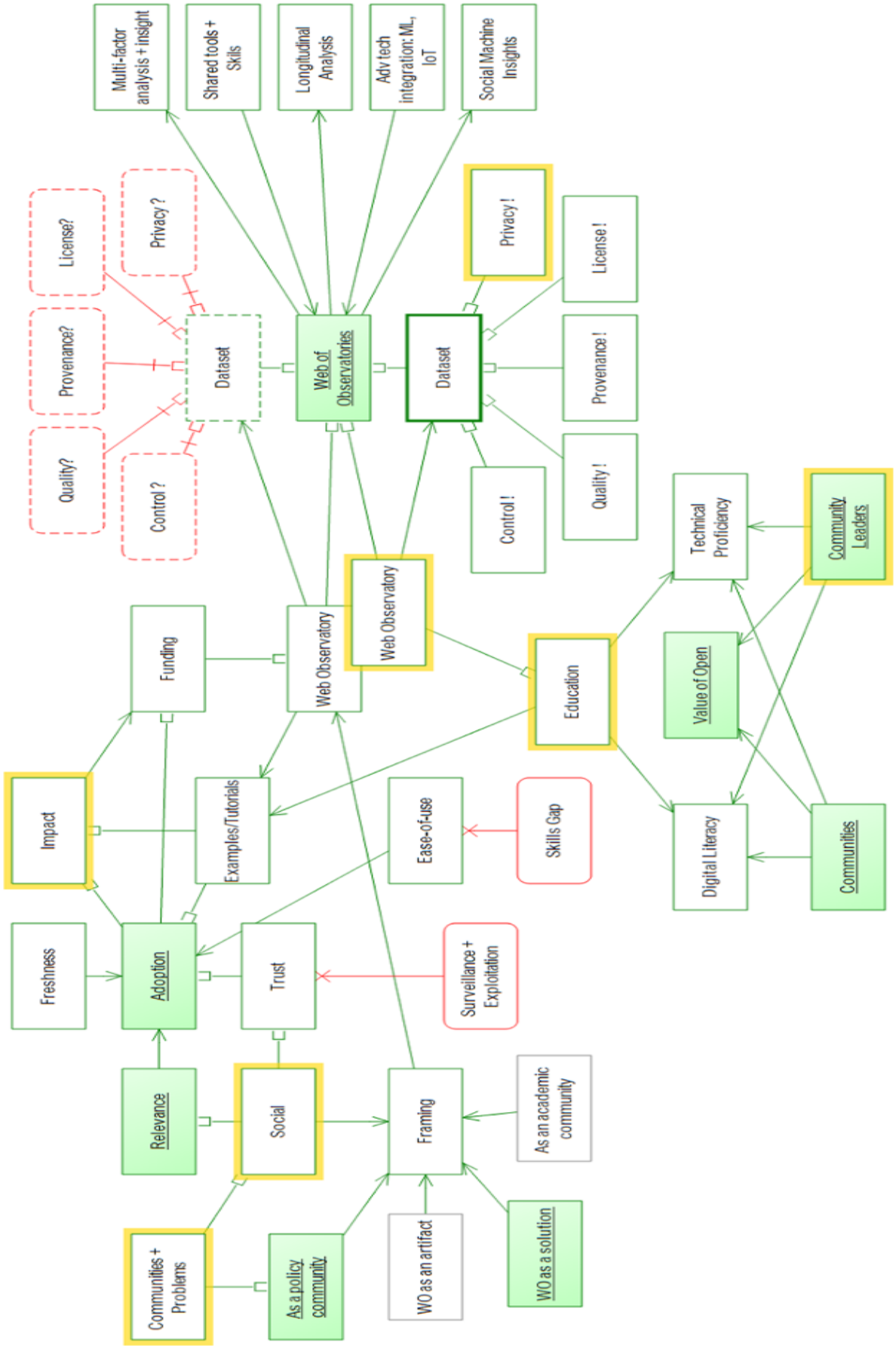


Figure E-9 [Davina]'s narrative

I started by asking about [Davina]'s initial impression of WO and what had interested her:

> “[At first] I just thought it was a data repository .. OK, well how's this different from data.gov.au or data.gov.org? I probably thought, this is something in the academic space..for academics who put data up and 'play' with it."

she initially characterised WO as something tentative, safe and theoretical rather than practical she suspected:

> "A playground for testing academic theories .. At the minute you can play, because nobody's going to die ..”

Even now - some time later - [Davina] is still concerned that not enough valuable data is being shared for the WO to be a serious tool for policy and that we are missing the valuable/sensitive datasets and tools that will add real value to policy.

> "How can you get a safe and secure online environment, and then bolt other things on? So whether it's machine learning or natural language processing, you've got to start with the data..data underpins everything. I think [at the moment] you'll get people who'll put datasets up to play. What you really want is to be able to link the Observatory with the sensitive datasets. I think once you can link the Observatory to real time real datasets that are actually valuable, then you can start solving some problems. At the minute, I think it's a playground to learn."

> "In other words, they [Government?] are happy to share data that's not that valuable. They won't share data that is valuable.. that other stuff can be somebody else's data that might be sensitive, but it could also be publicly available data. But it's actually that mix; it's the whole linked data piece.”

I asked if everyone viewed WO in the same way - as a playground..

> "The presentations [we] did to academic audiences, particularly in computer science - they did the whole: 'well so what?' thing .. It's something that uses APIs and has data and databases - it's just an apparatus. What you're testing [as a computer scientist} are algorithms, and the Web Observatory is not currently an Observatory of algorithms. If you make the same presentation to social scientists or economists or lawyers or sociologists, you get a very different reaction."

[Davina] cites trusted data and data access management as being an important factors sitting between fully private and fully open data:

> "When [a colleague] presented to government, they made the intuitive link as soon as [he] described the issue of the data provenance, you could see them all go, 'Oh thank God, this is something I can use.'. So it was that element of control: I can put stuff up there. I've got transparency. I can see who's used it. I can get permission. I can close it. I can make it open. I can make it public. I can make it private. I can suddenly turn it off if it's a stream".

So [Davina] reports different groups apparently framing the WO differently and portrays a wide range of reactions based on the context and framing of the WO concept.

> "[There were] a whole lot of people there - it was the fact that they imposed their own concepts on it ..[and]..they put on an overlay .. a bridge between a need and a solution."

With such different conceptualisations I asked if WOs would cluster into academic, educational systems, businesses and government/policy systems:

> "I think that in terms of the way that we've promoted it if you like, it's very much sat in the academic space. But as with all sort of good things in the academic space, you then start saying, -Well hold on, academics are trying to solve academic problems and education and policy are single perspectives, etc., but particularly it's [about] the PEOPLE."

Engaging with a range of group cultures/objectives I ask if she were suggesting that "tribes" were quite important for observatories and whether one needed to understand what tribe you're in and which other tribes you can/can't trust, and hence what you're risking?:

> "Do you envisage the ability to collaborate and share data between those organisations, and/or between businesses in academics or business in government? Do you see that as being possible, likely or do you see it as conceptually quite hard? [Researcher]

> "I think it's entirely likely, as long as you've got the legal framework and you've got the trackability and .. the accountability built in."

> "There needs to be [a] very, very clear idea of who does what. I think we need to be very, very clear about where the boundaries lie .. now, what we need now is the legalities."

We discussed the notion of boundaries, licenses and closed groups for WO and noted:

> "If all the data that you put on a WO is closed to everyone, then it has no value because nobody can see it .. so in that extreme case an Observatory exists but has no value at all,

> because nobody can get to anything .. but when you give access to the Observatory, it's open (accessible) data for someone:"

We discussed the idea of having WOs and *not* sharing benefits more widely but rather for more local constituents.

> "One of the interesting things about the digital catapult is that their reason to exist is not to be equally valuable to anybody out on the Internet. Their purpose is to create something that's more valuable specifically for UK businesses."

how international businesses engaging with the DC would restrict their insights and the information they extract from data sharing to a purely UK context remains unclear.

[Davina] conceptualises the continuum of systems partly in terms of the value/sensitivity of curated datasets and potentially sensitive data in WOs at one end comparing them with low-risk open government datasets:

> "At [one] end, we've got data.gov [and] things like CKAN and Socrata, those sorts of systems where you're basically shovelling completely open datasets there. And these guys are saying: No risk, it's free, it's open, we don't care who sees it .. it is completely open data and no risk, no strings [but] also no business model behind curating it ..[and therefore].. most people's experience of open data is that they go to their local open data or their local government, and they see a whole lot of files. They're not curated; they're not updated, and why would anybody bother because nobody's using them .. [And you need this to] start to drive a need from the data owner to say, "Well, I better curate this'."
>
> [For WO] "we have got to create a sustainable business model."

This highlights not only a technical issue but also a sustainability concern around WO systems which add some of their value by existing for long periods of time and building longitudinal datasets of high-quality data.

[Davina] suggests that technology is lagging behind ambition in this space citing an Australian example:

> "Adelaide had the 'multifunction policy' in the 1980s before anybody else had thought about it. And they had this concept [around multiple datasets] which they couldn't build because the technology wasn't there .. what they've got is a browser, and then they've

> got a whole lot of different databases behind it running it. So you still haven't got the data being able to talk to each other."

And so the project team was attempting to revisit this idea rallying different data sources (both public and private around the goal of measuring the city against the requirements of an ageing population to help with:

> "Adapting cities to be age-sensitive" given that "Ageing has bipartisan government and community interest and support globally"

according to the final project report. Leveraging the WO concept to build a system for government ageing quickly became the preferred choice as something that could not help one local community but could be replicated across other WOs:

> "So can we start to really .. link - not just within Australia but globally .. to collaborate and use it to solve a problem. That problem though, needs to be something that they're all linking to .. let's ask a policy question and see what datasets we might need."

In the ageing population space, the provision (supply) of facilities was available from public datasets while the demand side - who was using the services and their demographic profile - was *not* publicly available and so the project needed a partner with access to person data [Davina] explained:

> "The department of aged care. ..[have].. all the seniors' data. We have a senior's card .. everybody over the age of 65 gets a card. They've got all the data."

Privacy becomes a significant here issue requiring the support of local government agencies and a broad "opt-in" recruitment to allow seniors data to be captured and analysed for the WO in support of aged care service models.

Given the differences in focus I asked then about the role of WO systems for groups operating systems at different levels of detail, sensitivity and privacy - potentially each looking for different outcomes:

> "Does a WO have the option to link/bridge large open systems with the rigour and security of more focussed tightly managed and curated systems?"

She was convinced this would be possible under certain conditions:

> "if you had a secure space on the Web Observatory that was around a problem, and you were actually able to link to other datasets that were quite valuable around a specific problem, and the provenance and that was well policed, you can have that linkage.

> "in order to get the [open] benefit of this combined with the rigour and the curation of that, you .. bring those two together, which we think may be .. where the Observatory sits is the continuum in the middle, which is actually, I've got some closed data, I've got some stuff that's sensitive. It's actually only really valuable to me and anyone else when I start mixing it with other stuff

I asked if Government should be building or at least funding these bridging systems:

> "The challenge in the short term, government doesn't have that money."

Indeed though in the first instance some money was made available through ANZSOG and SA Government grant:

> "So that funding meant that we could actually develop and put something up."

This still left the project to seek both public and private funding to match the shortfall even to the extent of covering basic hosting costs for a bespoke WO at the local university whose individual departments were contributing valuable manpower to the project but who at the accounting/billing level [Davina] felt:

> "..didn't yet see the value".

There were several groups interested in the SA Web Observatory but repeatedly the issue of funding comes up [Davina] quoted an attendee at one of their project presentation

> "This is really interesting, but for me to put the resources of my people on this, we have to be solving a problem that we've got a stake in solving."

[Davina] characterises the next stage to be finding the problem where it's "part of their day job"

What is next for WO I asked [Davina]:

> "The Web Observatory now needs to really come out of the technical space .. it is still technically difficult for a non-technical person to upload data or do anything with it. [What it needs is] 'Press the button, upload the file.' Make it really, really, really simple."

I asked if WO reflected the user experience and the social nature of the WO she characterised it as:

> "No, because it's too passive .. once it gets to the visualisation phase, it starts to become a little bit more engaging .. [but we still need to know] .. How does this visualisation link the data that is there, to my problem .. That's the next step, I think, which is to really link it to a problem."

She stressed that WO needed to deeply understand the social element and build that in as technology gets cheaper and easier such that:

> "Once the pipe gets bigger, and the technology gets smarter, you better have done your thinking."

[Davina] seems optimistic but sees WO as being a work in progress:

> "I don't think it can necessarily solve problems yet. I don't think it can yet because I don't think that you're going to get the sensitive datasets up there yet."

These are social/cognitive issues rather than technical issues though the former may be underpinned by solutions to the latter. For a fully operational Social Machine [Davina] envisages a community of WOs focusing not only the data stored/referenced but also on the relevant issues faced by the communities that are served by it:

> "For something like the Observatory to be naturally effective, it almost needs to be a repository of problems as well as a repository of solutions .. and that is the Social Machine. Unless you get the humans who continue to use it, update it, make sure it's right, and link it to problems.

She doesn't necessarily see only a single thread running through all WOs

> "You might have an academic research playground, and then you might have a more secure commercial type sandpit where they can go."

> "What is really required is to create not just one Web Observatory, but a 'Web of Observatories', which will enable the sharing of data, analytics and visualisations across datasets, across jurisdictions and between organisations. This is what will yield true insight and enable much greater transparency as to how the Web is developing, both within Australia and around the world.

Davina Themes

Super-ordinate themes	Sub-Themes	
Adoption	.. Via fresh content/apps	The need to keep content relevant to attract and retain WO users
	.. Via ease of access/use	The need for low Ux and skill barriers to WO usage
	.. Balancing "wins" over concern over data release	The need to address concerns over the loss of control and possible implications of releasing data publicly while highlighting benefits
	.. Planning	The need to consider adoption strategies BEFORE the general availability of technologies
	.. Based on key features	Recognising that controlling data effectively and deriving confidence around quality and sources is key for adoption
Education	.. For applied learning	The need to make digital skills concrete through practical application
	.. Around Digital Literacy	The need to explain the implications and not just the skills around digital
	.. On the value models for "Open."	Resetting industrial-era value models around "buy/sell" for "share/exchange"
	.. At different levels/speeds	Expectation of iterating through n easy → challenging levels of skill to achieve mastery vs. a single "brain-dump."

Appendix E

	.. For deployment and maintenance	The need for technical/operational WO skills to complement/support user digital skills
Framing	..as a policy vs. academic community	The desire to frame WO in terms of a specific problem space rather than a solution (for any problem)
	.. In application terms vs. Computing/infrastructure terms	The difference between the framing of WO as a solution vs. an apparatus
	.. Relative to measurements of success	The distinction between viewing WO as an output/artefact per se vs. a bridge to some other output
	..as an academic exploration vs. a real-world challenge	The distinction between theoretical/exploratory work and applied work with consequences for people's lives
	..as distinct classes of WO problem → classes of WO system	The consideration that certain types of problem may require distinct types of WO with different features/processes in Academia Business and Government
	.. Within and across disciplines	Giving rise to interdisciplinary perspectives (frictions?) and/or new insights

Funding	.. As a limit to progress	Recognising that investment in WO is required to create and maintain systems/services
	.. In proportion to the complexity of data integration	Recognising that scope of each WO will e limited by the coast vs. perceived value of each additional source/feature/integration making the value model quite critical
	.. For dedicated teams vs. community contributions	The recognition that (potential) community members who don't have WO as their "day job" will be limited in their engagement
	.. From different tribes	The discussion around funding/support from academic research, business applications and government policy research
Impact	.. On government policy	Relating WO work to planning, evolving and validating government policy.
	.. Of personal vs. Open data	Considering how personal can be safely integrated into hybrid personal/public models
	.. Of integrating vs. displaying mixed sources	Distinguishing between data-level mixing of sources vs. presentation of distinct datasets on the same screen
	..of longitudinal data	Recognising the value of incremental data (trends) over time on long-term policy planning and measurement
	.. Of enhanced machine learning	Recognising that ever larger data volumes will require policies to be evolved based on data that no human has looked at

Appendix E

Social	.. impact of disclosing sensitive data	Recognising risk aversion as a uniquely human/social factor
	.. nature of exchange across tribes	Recognising the boundaries across which data and ideas and concepts flow
	.. embedding of problems vs. technical nature of tools	Recognising the social context for problems and for the fitness of solutions
	.. focus on crafting solutions for specific communities	The process of matching data/tools to people in the creation of a WO solution
	.. concept of value	Reflecting on whether people give away "valuable" things or only trade them
	.. framing of problems within a context	Recognising the translation of tools and terms into contextually recognisable elements of an understandable solution
	.. part of the Social Machine	Recognising the human and social elements in the sociotechnical system
WO	.. As a repository of problems as well as data	Envisaging WO as a community around which problems are highlighted and solutions shared.
	..as a sustainable model	Recognising that systems and datasets need to be curated and supported
	..as a teaching tool	Using WO to "play" with tools, models and non-critical datasets to build skill in analysis
	..as a research tool, business tool vs. policy tool	Characterising academic vs. financial vs. social outcomes for WOs with associated features/costs
	..as a bridge amongst+between communities	Envisaging WO as neutral ground between organisations to share data

	..as a focus for funding/engagement	Envisaging WO as a neutral ground around which to build large-scale projects
	..as a technological artefact	Identifying WO as a tool amongst other tools where the social differences may be more distinctive than the technical differences
	.. As a Social Machine for observing Social Machines	Recognising the WO itself involves social elements which are distinct from the social elements being observed
	..As a bridge between corporate EDW and Open Data repositories	Envisaging that the deep/narrow private nature of enterprise data warehouses and the broad/shallow nature of general open data repositories suggests a middle ground for WO
	..as a platform for IoT data	Envisaging IoT data as a suitable use case for WO to flow into public policy
	..on cheap, ubiquitous hardware	Envisaging WO feeds from cheap custom hardware such as Arduino/Pi or ubiquitous mobile devices.
	..within a legal/accountability framework	Suggesting that adoption requires an understanding of rights/responsibilities with the provided data and that this is a high priority