Digital Innovation - Investigating the Sustainability of New Kinds of Web

Ramine Tinati
School of Electronics and Computer Science, University of Southampton, Southampton. England
rt506@ecs.soton.ac.uk

Leslie Carr
School of Electronics and Computer Science, University of Southampton, Southampton. England
lac@ecs.soton.ac.uk

Susan Halford
School of Social Sciences, University of Southampton, Southampton. England
sjh3@soton.ac.uk

Catherine Pope
Faculty of Health Sciences, University of Southampton, Southampton. England
cjp@soton.ac.uk

1. INTRODUCTION

From a technical perspective, the Web is a distributed information architecture that is based on the concepts of interaction (HTTP), format (HTML/RDF) and identification (URI) [5]. "Browsing", "navigating" and "information discovery" are the kinds of generic activities that web developers and information scientists concern themselves with, but the more common labels adopted by users to describe their online activities are Social Networking, Internet Video, Blogging, Online Banking, Open Source Development, Internet Porn, E-research and Internet Shopping. Specialist kinds of interaction (shopping baskets, playlists, blogrolls) are recognizable in all these activities, even though users may be simply "navigating web pages". Those web engineers and content providers building on the Web to provide Internet Shopping (e-commerce, b2b, secure financial transactions, product databases, stock control, warehouses and delivery) have different concerns to those dealing with Internet Video (rights acquisition, media streaming, content licensing, bandwidth negotiation, format transformation).

Following the socio-technical perspective described by Law [7], we can see the Web as a loose affiliation of semi-independent content networks (each a web in its own right) with their own practices, technologies, business models and ecology of producers and consumers. The Web is a network of *networks of stakeholders* mutually reinforced and stabilized by each others' success and by W3C standards and policies.

Innovation in the WWW occurs either through making improvements within an existing network (better ways of delivering Internet TV, for example) or by the creation of an entirely new web of activity to supplement the existing Web. Berners-Lee describes this process as 'magic'[1]; this paper identifies a recent Web innovation, analyses the quantitative evidence of its adoption and models the processes involved in its development in order to understand the magic of the web and to improve its chances of sustainability.

2. LINKED OPEN GOVERNMENT DATA

A recent development in the Web has been the Linked Open Government Data (LOGD) agenda – a set of worldwide initiatives driven by academics and governments in a bid to publish government data in a suitable Linked Open Data format [4]. A wealth of data – educational, financial, geographic, social statistics, public spending - is currently locked away behind non-machine readable formats, which potentially contribute valuable information about the world, and most importantly, enabling the government, country, and the world to function more efficiently [2]. The growth in interest has seen investment from countries within Europe, the Americas, Asia, Australasia. The countries are actively participating in developing large collections of datasets, providing a rich set of information for their citizens to exploit. The complexity of these initiatives are truly socio-technical, requires the adoption of new Semantic Web technologies, changes in government practices, and continuous commitment from not only the various invested stakeholders, but also the end-users.

As a result of the complexity of these initiatives, success is not guaranteed, highlighted by recent workshops concerned with overcoming barriers and resistance to LOGD [8]. Although the *principles* of LODG are established, the *practices* are not so well understood. The premise for the establishment of LOGD is that there will be continuous publication of data, offering added value to users, but, in fact, we know rather little about if or how the LOGD is growing. Is publishing LOGD becoming a day-to-day practice embedded within the structure of the government or will the initial hype fade to a tail-off in interest?

To address this, we investigate the activities of a number of data catalogs from a variety of countries participating in LODG initiatives. We then extend our investigation to consider the social phenomena that may influence the activities observed within the data catalogs - by doing so we develop a framework which ties together empirical observations and social phenomena observed in LOGD.

3. LOGD DATA CATALOGUE ANALYSIS

To explore the stability of this new movement towards open government data, we set out to track the data catalogs that provide a single point of access for their country's Open Government Data. These countries and the total number of records held in the data catalogs are shown in Table I. The frequency and rate of deposits are being be examined using ROAR, the Registry of Open Access Repositories[1].

TABLE I. OPEN GOVERNMENT DATA PORTALS

Country	Data Catalog Portal URL	#Records	#Bad Records
USA	http://data.gov/	4178	1623
Canada	http://data.gc.ca/	758	15
UK	http://data.gov.uk/	7151	1062
Spain	http://www.opendatacordoba.com/	2384	0
Australia	http://data.gov.au/	227	27
New Zealand	http://data.govt.nz/	571	0

[1] roar.eprints.org, used to monitor the rate of Open Access activity, has been modified by the authors to analyse Open Government Data repositories

As Table I shows, the number of records available within the data catalog varied considerably, as did the number of badly formatted records found. The two largest collections of records, data.gov and data.gov.uk both suffered from a large proportion of bad records - 39% of data.gov and 15% of data.gov.uk records had inconsistent or incomplete datestamp metadata. Many of these records where missing a datestamp entirely or were formatted with only parts of the date *e.g. just the year given, or the text "today"*. Another observation can be made regarding the number of records available vs. the age of the data catalog. Although Australia's and New Zealand's portals were established in 2009, the number of records available are low, especially compared with the UK and US[2]. Interestingly, Spain's portal, even though being one of the most recent has a large amount of records and no badly formatted datasets.

TABLE II. RATES OF DEPOSITS IN THE DATA CATALOGS

Data Catalog Portal URL	Rate of deposits		
	High	Medium	Low
http://data.gov/	1	11	130
http://data.gc.ca/	2	2	28
http://data.gov.uk/	1	55	170
http://opendatacordoba.com/	3	13	0
http://data.gov.au/	0	1	17
http://data.govt.nz/	0	8	53

Table II represents the size of deposits made during a single timestamp, which ROAR categorizes as a high (100+), medium (10-99), or low (1-9) rate of deposit activity. A healthy catalog would usually exhibit medium to high rates of deposits, indicating that there is a good wealth of data being produced. Furthermore, we can examine the frequency of deposits being made within the catalogs, illustrated in Figure $I._A$ and $I._B$ - the spikes represent the number of deposits made to the catalog at a specific date. Comparing the deposit frequency of data.gov.uk to data.gov.au, it can be seen that there are much more frequent deposits within data.gov.uk, all which are of a good size. In comparison to this, the deposit activities of data.gov.au catalog is far less constant, an indication of an unhealthy catalog.

The observation and comparison of the different data catalogs have highlighted some important aspects that contribute towards describing their behavior. A data catalogs activity can be described by a number of characteristics: the total number of records held, the frequency of deposits being made, the size of the deposits being made, and the consistency of deposits over a large time span. However, this only contributes to part of the overall picture of the activities within the Government Linked Open Data movement - quantitative measures provide a good description of the current state of activity, but we now want to take a step further and provide an explanation to what is being observed.

4. MODELING ACTIVITIES OF LOGD

There does not appear to be a uniform technical logic that is being consistently replicated throughout the different countries' initiatives – the range of outcomes observed are the result of the processes and activities of the interactions between technology and society. To model this, we adopt Actor Network Theory (ANT) as an analytical approach to understanding the interactions between human and technological actors[6]. ANT's *process of translation* provides a method of describing the outcomes of the interactions between actors involved in a network of activity[3]. By analyzing the formation of network within four key stages: Problematisation, Interessement, Enrolment, and Mobilization, we can examine the socio-technical relations between actors, highlighting the interplay that results in a functioning network.

Taking this a step further, we can combine this *process of translation* with the ROAR tools for monitoring data catalogues and produce a framework which incorporates quantitative and qualitative methods of analysis. Fig. II extends Berners-Lee's the magic Web development [1], overlaying its different stages with the process of translation. Each of these stages involves a variety of actors working together in a joint effort towards a common goal (in this case, sustainable LOGD). Sustainability is achieved through the continuous involvement and alignment of all the processes.

Different processes require the support of different analytical tools; the Problematisation and Interessement of the process requires identification of actors, setting roles and gaining interest at a micro level, whereas within the macro are tools such as ROAR, providing ways to monitor actor's Enrolment and Mobilization. The macro is heavily influenced by the micro, and vice versa. Furthermore, the process of design should not be regarded as separate social and technical components, but rather as a co-constitutional relationship, where the issues identified within the macro directly affect the design. Overlaying the process of translations reinforces Berners-Lee's concept that the development of the Web forms through a lifecycle, but by overlapping the process of translation makes it clear that the stages within the lifecycle are not mutually exclusive.

In order to maximize the likelihood of a stable, sustainable network of innovation, this paper presents an initial attempt at combining ANT/ROAR to provide stakeholders with evidence-based tools and a social model to reflect on their role and effectiveness.

5. REFERENCES

1. Berners-Lee, T. The Two Magics of Web Science. *Keynote Speech at 16th International World Wide Web Conference (WWW2007)*, (2007).
2. Berners-Lee, T. Putting Government Data online. *W3.org*, 2009. http://www.w3.org/DesignIssues/GovData.html.
3. Callon, M. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. (1986), 196-223.
4. Hall, W. The Ever Evolving Web : The Power of Networks. *Journal of Communication 5*, (2011), 651-664.
5. Jacobs, I. and Walsh, N. Architecture of the World Wide Web, Volume One. *W3C*, 2004, 1-51. http://www.w3.org/TR/webarch/.
6. Latour, B. *Reassembling the Social: An Introduction to Actor-Network-Theory by Bruno Latour*. Oxford University Press, 2005.
7. Law, J. Centre for Science Studies And if the Global Were Small and Non-Coherent? Method , Complexity and the Baroque. *Complexity*, (2003).
8. Share-PSI.eu. Discussion summary of SharePSI workshop: Removing the roadblocks to a pan European market for Public Sector Information re-use. *Transition*, May (2011), 1-12.

[2] Size of population should not affect the number of datasets

FIGURE I.A FREQUENCY OF DEPOSITS OF DATA.GOV.UK

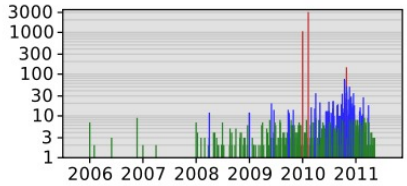

FIGURE I.B FREQUENCY OF DEPOSITS OF DATA.GOVT.AU

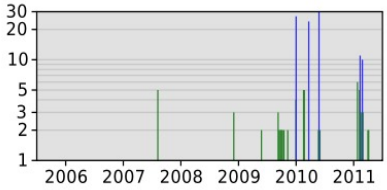

FIGURE II. THE PROCESS OF WEB TRANSLATION