Data Observatories
Decentralised data and interdisciplinary research

Thanassis Tiropanis[1]

University of Southampton,
Electronics and Computer Science
Southampton, UK
t.tiropanis@southampton.ac.uk

Abstract. Sharing data and observations has been at the centre of academic discourse for centuries. The Internet has enabled new promising methods of sharing, introducing a number of challenges. This paper discusses the concept of data observatories as decentralised platforms for sharing data and observations, and outlines a reference architecture for their deployment. It argues that decentralisation has been a necessity for interdisciplinary research and that it will benefit data-driven innovation beyond academia. It concludes with discussion on the necessary conditions of decentralisation and literacy for innovation and wider engagement in academic discourse.

1 Introduction

Publishing and sharing data for the purpose of statistical analysis is an activity that can be traced back to as early as the 16th century in England with the weekly publication of mortality statistics [2]. The purpose of those publications was to record deaths during outbreaks of plague in the City of London. It was on those data that John Graunt (1620-1674) was able to provide his “Natural and Political Observations on the Bills of Mortality” [2]; an analysis to understand changes in the population of the City of London and how those might relate to the plague.

The following centuries saw the proliferation of data gathering and statistical analysis in various sectors beyond public health. In recent decades, the Web has made it possible to share data on a large scale and to provide analysis (observations) on data from different resources. However, these activities have been problematic for a number of reasons. First, it is not always desirable for data publishers to make their datasets open to all, as favoured by most Web publishing platforms. There can be legal and ethical reasons behind this choice but it could also be that publishers do not know and cannot control the potential use that others can make of those data in a way that is acceptable to them; i.e. it can often be the case that a publisher places significant effort in the curation and publication of a dataset for free while another party can make profit by using that dataset in a value added online service. Second, it is not easy to discover what data are available or to establish their relevance to the analysis at hand, and their quality and provenance to that effect. This limits the quality of observations on Web data. Third, there is a barrier in terms of the digital skills required to discover, process and analyse data. While social media have lowered the barrier for publishing content on the Web there is no endeavour of that scale for sharing data.

As a result, the activity of providing and sharing ‘observations’ in recent years is primarily in the hands of individuals with higher digital skills who have access to datasets gathered by sizeable organisations running online platforms, and who are in a position to negotiate and implement the legal and ethical aspects necessary for data gathering and analysis. This not only limits the potential for data-driven innovation but also the potential of research. The decentralisation of the data ecosystem to make it more accessible to individuals with varying digital skills promises to benefit society and to further innovation and research.

The conundrum of secure sharing of data among people has many aspects. It is not just the technological aspects that can be challenging but also the legal, ethical and social ones. There have been proposals for Web Observatories as socio-technological artefacts for sharing data and observations. The concept of a Web Observatory started as a means to enable interdisciplinary collaboration in the context of Web Science [15] and it was subsequently developed as an approach to enable data sharing and analysis of a wider scope, beyond Web Science [14]. In recent years, ethical and legal aspects of data sharing in Web Observatories were identified and best practice as well as a new ethical models were proposed [18]. The technological challenges of real-time data sharing [13] and of sharing databases online [16] are two of the technological issues that were further explored. In the meantime, the increasing volume of IoT data, the characteristics of IoT datasets [10] and their potential for analytics present challenges for storage and computation on the cloud and at the edge [12], which adds to the technological requirements and challenges of data sharing; personal observatories on lightweight computers at the edge were developed to explore some of those requirements [11].

These developments necessitate one step further in the conceptualisation of observatories; namely, the concept of sharing data and observations not only on the Web and on the cloud but also on private cloud and edge deployments. They also necessitate frameworks for the codification and negotiation of access and use of data resources in order to produce and share observations. They also require mechanisms to establish accountability in data sharing ecosystems. Regulatory aspects for data & observation sharing platforms need to be discussed as well as the potential feedback loops of sharing observations within certain groups.

To that end, this paper presents the concept of decentralised *Data Observatories* as the next step of *Web Observatory* evolution. Section 2 discusses the elements of decentralised data observatories in terms of data resources, metadata and data protection, while section 3 discusses the utility of different aspects of decentralisation on sharing data and observations in the context of interdisciplinary research and beyond. Section 4 presents a *Reference Architecture for Data Observatories* in terms of how elements of data observatories inform concepts, concerns and services in a decentralised data sharing ecosystem. Finally, lessons from current deployments and guidelines for future Data Observatories are discussed in section 5.

2 Data observatory elements

From the initial stages of their development, Data Observatories are conceptualised as communities of people who engage with two types of resources: *datasources* and *observations*. The former can be files, databases or any query interface that can provide data to be processed by an individual or software. The latter are analytic applications, visualisations, statistical analysis or other processing of data obtained from a datasource in order to provide an insight or observation. Members of the community can publish and share those resources [14] within a community (i.e. within a Data Observatory). Those who publish resources are called user-publishers or for simplicity *publishers*. Those who access those resources are identified as *users*. Sharing is subject to the consent of user-publishers to the access and use of those resources by other members of the community. They are also subject to conformance to legal and to ethical frameworks set by the community. Sharing across communities, i.e. across Data Observatories, requires alignment of legal and ethical frameworks and potentially additional negotiation to establish the required level of trust. *Meta-information* or *metadata* to establish and confirm the provenance, intellectual property, licensing and overall quality of shared resources is also needed within and across data observatory communities. The term metadata is used to describe structured meta-information in a machine-processable format.

On a technological level, each Data Observatory fulfils its role by providing a catalogue of datasources and observations on a portal, with all the required meta-information. They also involve access control components, *Access Proxies*, to ensure that only authorised parties that have negotiated terms of use with the user-publisher get to access published resources. Datasources and analytic applications (observations) can be stored on repositories maintained by the individual user-publisher or on third-party cloud infrastructure. *Obfuscation* can be applied on accessed datasources or on meta-information on the catalogue; user-publishers can have the freedom to choose how much meta-information on their published resources is available to other users but also on the detail, number and frequency of records that can be returned when their datasources are accessed by specific users. Access Proxies can use obfuscation to implement access control policies on datasources but also on observations. *Access Agreements* can provide an unambiguous specification of access and use that a publisher grants to another user for a specific resource. Table 1 summarises those elements of Data Observatories and provides terminology that is used in the rest of this paper.

3 The case for decentralisation

The potential of data for innovation and research has long been discussed. Data science and Artificial Intelligence (AI) provide tools aiming to fulfil that promise while different disciplines have explored methods to engage with data in meaningful, effective and rigorous ways. The case for decentralisation has been adopted in interdisciplinary research communities and has informed the conceptualisation of data observatories. There is also discussion on the limitations of the centralisation on the Web in recent years, where a few successful websites have amassed big volumes of data [1].

3.1 In interdisciplinary research

The challenges of sharing data and analysis for computational social science have been discussed and a self-regulatory regime of procedures, technologies, and rules has been proposed [4]. In Web science, Web observatories have been proposed to address meaningful engagement with research data [15], while, in Internet science, the challenges of semantic catalogues for data and analytics in order to understand the state and impact of the Internet have been discussed [17]. From a Data science viewpoint, observatories can enable better discovery and engagement with data [8] but there are challenges in negotiating appropriate methodologies. There is discussion on issues of privacy, security, consent and trust and how those can be addressed in a wholistic way in the deployment of data sharing platforms to study socio-technical systems on the Web [9] or to deploy data trusts for AI innovation [6]. The reproducibility of AI [3] and bias [7] give rise to a number of epistemological, ethical and legal concerns.

These concerns were reflected in the design of data observatories in research environments. First, it was necessary to provide an architecture where research data would not be shared in a centralised repository but at the research institution that provided and curated those and their analysis; there were legal and ethical reasons for that choice. Second, analysis had to cite the data on which it was based for reasons of accountability and reproducibility among others. Third, meta-information on the methodology of gathering and analysing data as well as on the ethical safeguards had to be available. Fourth, it had to be possible to cite data or analysis but also provide different type of access to third parties and to the research community for legal and ethical reasons; that required access control and obfuscation techniques (including summarisation and anonymisation). Finally, it required different communities to engage in academic discourse over data and observations and necessitated the means to support that discourse.

Decentralisation in Data Observatories is conceptualised in the following design choices, which can be derived from earlier work in this area [14] and subsequent development:

- Identifiers; using identifiers (such as URIs) to identify individuals, communities, datasources and observations in order to determine permitted access and use. Decentralised Identifiers[1] can be considered.
- Meta-information; use of metadata to enable cataloguing, discovery and use of datasources and observations.
- Access control; publishers of datasources and observations control which users can have access to them; they also control how much of the meta-information on those resources can be available to third parties.
- Consent; publishers of datasources and observations can consent to specific uses of those resources and their meta-information by third parties.
- Portals; community- or organisation- maintained portals to host meta-information on data and observations available for sharing by their members subject to access control and consent. Any community or organisation (or even individual) is able to set up their own portal.

[1] https://w3c-ccg.github.io/did-spec/

- De-coupling of catalogue from storage; portals provide catalogues of resources and meta-information only. Listed resources can be stored in any third-party repository.
- Inter-portal communication; communication between portals maintained by trusted parties is possible. Such communication supports authentication of users, exchange of meta-information on shared resources, search, and negotiation of access and use across portals. Building a trust relationship includes compatibility of legal and ethical frameworks between communities. Standard protocols and schemata are used where possible (e.g. OpenID connect for authentication).
- Application Programming Interfaces (APIs); Where possible, portals provide APIs to support the development and publication of datasources and observations, communication across portals, and value added services for their users to help them monitor how the resources that they share are accessed and used.

3.2 In data markets and innovation

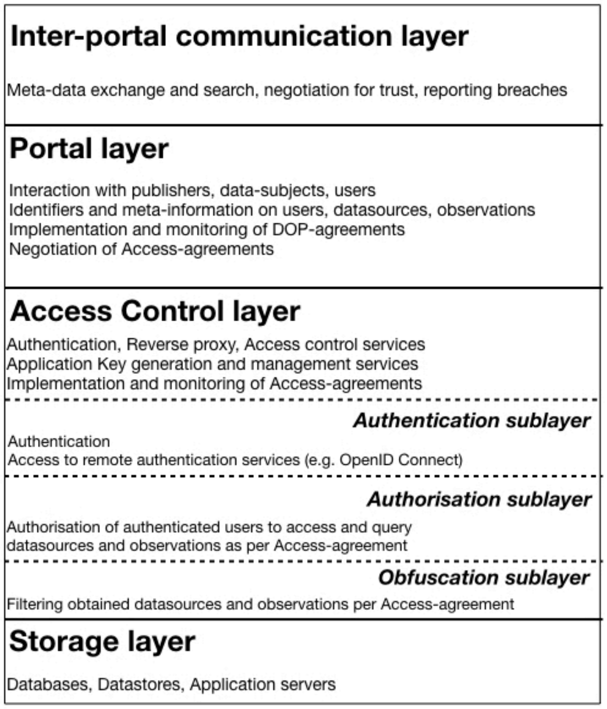
At the same time, technology suites to support re-decentralisation have been proposed. The SOLID suite [5] envisages a paradigm where Web service providers (such as online social media providers) do not store user data on centralised repositories. User data are stored and maintained in personal online datastores (PODs) instead. Service providers need to request permission from the users in order to access their POD instead of storing user data centrally. Web applications rely on access to the PODs of their users instead of a central datastore. Users can have control on whether to revoke such access and they have full control of their data at any time. Apart from SOLID, the emergence of blockchain can further foster development of decentralised data ecosystems by means of distributed ledgers the integrity of which can be maintained by a community and supported by consensus building mechanisms.

Nevertheless, data observatories present an additional challenge: it is not data that need to be shared but observations (e.g. analytic applications) on those data too. They envisage standard users and two types of publishers *datasource publishers* and *observation publishers*. This can be seen as a three-sided market that can foster cross-network effects, where, as more data become available, more observations become available and more people engage with both. However, at the same time, it presents complications when publishers need to control access and use of their resources. There is symmetry in that both data and observation publishers can control who access their resources and for what use. However, there is asymmetry in that observation publishers need to have one agreement with the data publishers for the data resources that they use, and a separate agreement with those accessing observations; those two agreements need to be aligned and accountability needs to be carefully established along the chain of data publisher, observation publisher and observation subscriber.

Even though these concerns seem to be specific to academic research environments at the moment they could be pertinent to future development of data sharing ecosystems. They necessitate a reference architecture that is specific to data observatories, where both datasources and observations are shared, as opposed to other models where only data sharing is envisaged. Concepts, components and best practice can be developed on such a reference architecture and inform its evolution in the future to support interdisciplinary research but also data-driven innovation.

4 A Reference Architecture for Data Observatories

An abstraction over existing data observatory architectures is presented by means of a reference architecture. It is informed by the principles argued in Section 3 and by previous work in this area [14]. It provides a number of layers to group components addressing similar areas of concern relying on the services of layer below. Figure 1 illustrates those four layers:



Inter-portal communication layer

Meta-data exchange and search, negotiation for trust, reporting breaches

Portal layer

Interaction with publishers, data-subjects, users
Identifiers and meta-information on users, datasources, observations
Implementation and monitoring of DOP-agreements
Negotiation of Access-agreements

Access Control layer

Authentication, Reverse proxy, Access control services
Application Key generation and management services
Implementation and monitoring of Access-agreements

Authentication sublayer

Authentication
Access to remote authentication services (e.g. OpenID Connect)

Authorisation sublayer

Authorisation of authenticated users to access and query datasources and observations as per Access-agreement

Obfuscation sublayer

Filtering obtained datasources and observations per Access-agreement

Storage layer

Databases, Datastores, Application servers

Fig. 1: Data Observatory Reference Architecture

- Storage layer. This is the layer that provides for secure storage of datasources and observations by publishers. The storage layer exists independent of data observatories and it can be instantiated as a database, file system, repository, cloud storage, application server.
- Access control layer. It provides for secure access to resources in the storage layer and it can host access proxies who can act on behalf of publishers. An access proxy can employ authentication services to identify users, authorisation services to establish access rights to specific resources, and obfuscation, as specified in Access-agreements.
- Portal layer. This is the data observatory portal layer that provides the means to list and discover datasources and observations and relevant metadata by users (subject

to DOP-agreement). The portal layer has its own storage of meta-information on listed resources. It can support negotiation of Access-agreements and it relies on services of the Access Control layer to ensure that those agreements are respected.

- Inter-portal communication layer. Communities of data observatories that have established and maintain relationships of trust can share metadata on resources that they list and enable each other's users to negotiate and obtain access to those resources. Information exchanged between portals is subject to DOP-agreements and it can support decentralised search across data observatories.

Implementations of data observatories such as the Southampton University Web Observatory[2][3] can be seen as instantiations of this reference architecture. A data observatory implementation includes the portal and the access control layers at its core and it can interface with a number of storage layer providers. Authentication across communities has been implemented using OpenID Connect[4].

5 Deploying decentralised observatories

Interdisciplinary research can be seen as a microcosm where decentralisation is essential, building on the structures and practices of academic institutions that predate the industrial revolution and even nation states. The remit of academia as an environment that fosters rigorous discourse has necessitated ways to share and cite resources and methods. The printing press and recently the Internet have provided new ways of sharing data and observations and they inform ethical frameworks and policy. Nevertheless, technical aspects of deploying data observatories are non-trivial. The reference architecture presented in this paper is an abstraction of current architectures that support decentralisation in terms of decoupling data (datasources) from analytics (observations), and by fostering mechanisms to negotiate and share those resources with trusted parties. There is no claim that it can guarantee the deployment of a decentralised observatory but it aims to contribute to the discussion on decentralised data ecosystems.

Beyond research, having discussed the benefits of decentralisation for data-driven innovation it is worth reflecting on readiness to adopt decentralised architectures for data sharing. It can be argued that innovation today relies on data collected by service providers via online platforms or devices used (and usually paid for) by consumers. AI provides tools to make observations on those data in ways that consumers (and often providers) are not fully aware of. However, those observations have the potential to generate wealth and they also have the potential to change people's lives by informing decision making. Further, it has been argued that the intensity and the mode of user interaction on online platforms not only enables service providers to predict but also to modify behaviour [19]. This can lead to vicious circles, where 'observations' by software that is not transparent and cannot be reproduced become self-fulfilling prophecies by means of feedback loops and behavioural modification. This could stagnate innovation and can lead to dystopian developments. In this light, the case for decoupling data

[2] https://webobservatory.soton.ac.uk

[3] https://github.com/webobservatory

[4] https://openid.net/connect/

from observations, for making the links between observations and data explicit, and for giving people control over access to their data could help avert such developments but it would not be enough.

Decentralisation requires people to be able to engage with data, methodologies and analysis so that they can make informed decisions on what to share with whom. Individuals need to be able to gain access to the possible knowledge that data can generate (rather than just the raw data) in order to appreciate its power. Their decision on sharing data needs to be reversible and re-negotiable. Where revenue is generated by their data they should be able to negotiate a share of that revenue. It is for this reason that the data observatory principle of sharing observations along with data can be empowering. Decentralisation and literacy in data sharing ecosystems could pave the way for innovation and wider engagement in academic discourse. Perhaps it is worth noting that John Graunt who provided those early 'observations' was not a demographer but a haberdasher by profession.

Acknowledgements

I would like to thank all colleagues on the journey of conceptualising and building data observatories in the EPSRC-funded projects SOCIAM (EP/J017728/1) and PETRAS (EP/N02334X/1), the EU-funded projects EINS (EU-FP7 288021) and VOICE (EU-CIP 621137), and the Web Science and the Internet Science communities. I am grateful to Alex Poulovassilis of Birkbeck for her feedback on this work. My sincere thanks to Wenceslao J Gonzalez from the University of A Coruña and the organisers of the conference on Internet and Science: An Analysis from the Structural and Dynamic Complexity for inviting me to speak and discuss on this topic.

Opinions expressed in this article do not necessarily reflect those of colleagues and funders of these projects.

References

1. T. Berners-Lee. Tim Berners-Lee on the Web at 25: the past, present and future. *Wired*, Mar. 2014.
2. T. Birch. A Collection of the Yearly Bills of Mortality, from 1657 to 1758 Inclusive, 1759.
3. M. Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018.
4. D. Lazer, A. Pentland, L. ADAMIC, S. Aral, A. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. Gutmann. Computational social science (2009). 323:721.
5. E. Mansour, A. V. Sambra, S. Hawke, M. Zereba, S. Capadisli, A. Ghanem, A. Aboulnaga, and T. Berners-Lee. A demonstration of the solid platform for social web applications. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 223–226, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
6. K. O'Hara. Data trusts: Ethics, architecture and governance for trustworthy data stewardship. White Paper 1. Web Science Institute, in press.
7. O. A. Osoba and W. Welser IV. *An Intelligence in Our Image*. The Risks of Bias and Errors in Artificial Intelligence. Rand Corporation, Apr. 2017.

8. C. Phethean, E. Simperl, T. Tiropanis, R. Tinati, and W. Hall. The Role of Data Science in Web Science. *IEEE Intelligent Systems*, 31(3):102–107.
9. N. Shadbolt, K. O'Hara, D. De Roure, and W. Hall. *The Theory and Practice of Social Machines*. Lecture Notes in Social Networks. Springer International Publishing, Mar. 2019.
10. E. Siow, T. Tiropanis, and W. Hall. Interoperable and Efficient: Linked Data for the Internet of Things. In *Internet Science*, pages 161–175. Springer, Cham, Cham, Sept. 2016.
11. E. Siow, T. Tiropanis, and W. Hall. PIOTRe: Personal Internet of Things Repository. Aug. 2016.
12. E. Siow, T. Tiropanis, and W. Hall. Analytics for the Internet of Things. *ACM Computing Surveys*, 51(4):1–36, Sept. 2018.
13. R. Tinati, X. Wang, T. Tiropanis, and W. Hall. Building a Real-Time Web Observatory. *IEEE Internet Computing*, 19(6):36–45.
14. T. Tiropanis, W. Hall, J. Hendler, and C. de Larrinaga. The Web Observatory: A Middle Layer for Broad Data. *Big Data*, 2(3):129–133, Sept. 2014.
15. T. Tiropanis, W. Hall, N. Shadbolt, D. de Roure, N. Contractor, and J. Hendler. The Web Science Observatory. *IEEE Intelligent Systems*, 28(2):100–104, 2013.
16. X. Wang, A. Madaan, E. Siow, and T. Tiropanis. *Sharing Databases on the Web with Porter Proxy*. International World Wide Web Conferences Steering Committee, Apr. 2017.
17. X. Wang, T. G. Papaioannou, T. Tiropanis, and F. Morando. Eins evidence base: A semantic catalogue forinternet experimentation and measurement. In T. Tiropanis, A. Vakali, L. Sartori, and P. Burnap, editors, *Internet Science*, pages 90–99, Cham, 2015. Springer International Publishing.
18. C. Wilson, T. Tiropanis, A. Rowland-Campbell, and L. Fry. *Ethical and legal support for innovation on web observatories*. ACM, May 2016.
19. S. Zuboff. *The Age of Surveillance Capitalism*. The Fight for the Future at the New Frontier of Power. Profile Books, Jan. 2019.

Table 1: Data Observatory elements.

Element	Description
Datasource	A dataset, data stream, database or file shared by a publisher on a data observatory. It involves a user interface (UI) or an application programming interface (API) for obtaining data.
Observation	The result of analysis on data obtained from one or more datasources. It can be a visualisation or analytic application which can be static or dynamically updated. Observations can also involve interaction with the user accessing them and they can be subject to obfuscation by the publisher.
Meta-information	Information or metadata on datasources or observations for the purpose of cataloguing, discovery, attribution, licensing, research use. Vocabularies such as Dublin Core, Schema.org, PROV and DCAT can be used to provide metadata.
Access-agreement	An agreement for access to a datasource or an observation between a user-publisher and another user. It sets out the access level (access rights, volume and detail of obtained records, frequency of access, use or licensing. It can also specify how much of the meta-information on a published resource can be available to other parties.
Data Observatory Portal (DOP)	An often Web-based front-end that enables a user to access the catalogue with the meta-information of available datasources and observations; it can also support negotiation with the publisher for access and use of those resources. API access to datasources can be possible for the purpose of developing observations.
User	A user of data observatory. Users who publish resources for sharing are identified as *user-publishers* or *publishers*. Users agree to terms for access to the data observatory portal (DOP-agreement) and engage in access agreements (Access-agreement) as users or publishers for access to specific resources.
DOP-agreement	An agreement between users of the portal that can involve access to the catalogue, publication of datasources, publication of observations, availability and use of the portal. This agreement is between each user and the DOP administrator, who acts on the behalf of the community or the organisation maintaining the portal.
Repository	Secure storage for datasources or observations, i.e. resources that can be listed in a data observatory. Repositories exist independent of data observatory portals. Access to stored resources is managed by the users-publishers. Those, in turn, can delegate access control to Access-proxies for users with whom an Access-agreement has been established in a data observatory.
Access-Proxy	A component that is responsible for access control to datasources and observations. It can authenticate users and implement access policy on behalf of publishers. It can also enable API access to software used by individuals with the necessary access rights. It monitors and implements the Access-Agreement employing obfuscation if necessary.
Obfuscators	Software components that can be employed by Access-proxies in order to perform obfuscation on data obtained by datasources or on observations as per an established Access-agreement. Such level of access can concern the number of returned results, number of queries over a certain period, the detail of returned records or frequency of access. Obfuscation can also be employed by the DOP-portal so that publishers can discriminate on how much of the meta-information on their listed resources can be available to different parties.