# 6 Web Observatories: Gathering Data for Internet Governance

**Wendy Hall, Aastha Madaan, and Kieron O'Hara**

The World Wide Web is the most significant application of the Internet, a simple, easy-to-use information space indexed by uniform resource identifiers on which are built most of the services accessed by Internet users today, including search engines, video streaming, and social media. Although the Internet predated the web by decades, the web brought a technical revolution, with the most significant social impact changing many aspects of how communication takes place and social behavior is shaped in politics, economics, leisure, entertainment, scientific research, commerce, and social interaction. As noted by Mueller and Badiei in chapter 2, the creation of the web (together with the invention of the browser) was partly responsible for the emergence of the Internet as a mass public medium, which has made Internet governance such a key issue. It has evolved from an efficient, although not unique, document repository to an active socio-cognitive space in which people express ideas and emotions across geopolitical boundaries. Its impact and reach have emphasized the importance of interpersonal integrity and issues of data ownership, privacy, trust, and surveillance in mainstream research, while the data, tools, and platforms that constitute and enable the web are distributed geographically, across legal jurisdictions. They are also used differently, with different levels of effectiveness and embeddedness, by diverse cultures, genders, age cohorts, and economic classes.

The social challenges of context of use and analysis make the task of studying the web complex. Neither are the technical challenges trivial, as the data that are generated may or may not be preserved. Rapid changes make it harder for researchers to find evidence to test hypotheses related to research questions pertinent to governance and policy making. For example, consider the prominent roles played by private social networking entities in political campaigns run on the web (Stieglitz and Dang-Xuan 2013)

or during disasters (Ngamassi, Ramakrishnan, and Rahman 2016) when these platforms play a critical role in providing physical aid and basic living facilities to the stakeholders in consideration. In one specific example, Facebook has been implicated in the apparent violence directed at the stateless Rohingya people in Myanmar (2016–2018), as a means of orchestrating anti-Rohingya sentiment (Banyan 2017), as a means for jihadis to spread extremist messages (Singh and Haziq 2016; Stevens and O'Hara 2015), and as a means for the Rohingya to inform the world of their problems (Wong, Safi, and Rahman 2017). The effects of social networking may be found in more stable societies too—for instance, the rise of smartphone use among US teenagers correlates with (though may or may not have caused) a sudden rise in suicide rates over the same period by 31 percent for teenage boys and more than 100 percent for girls (Twenge et al. 2018).

In this chapter we argue that, while it is important to address various Internet governance issues at the protocol level (or design level), it is also critical to understand the affordances of Internet use for the interactions among stakeholders. For that, effective, ethical, and secure methods of gathering and sharing data will be required. In this chapter, we consider the challenges to creating and disseminating such methods and describe an architecture for that purpose, which we call the Web Observatory (Hall et al. 2013; Tiropanis et al. 2013). While the Web Observatory is situated in the web context, using web protocols to organize data, its value for Internet governance stems from the importance of the web as the gateway to the use of the Internet. The architecture is designed to meet a set of technical, social, and legal challenges that will stand in the way of any kind of evidence-based Internet governance. Additionally, although we don't highlight this in this chapter, the idea of a Web Observatory is intended also to provide a pragmatic means to facilitate the sharing of data; to this end, it has been argued that the Web Observatory is a potential architecture for a *data trust* (O'Hara 2019), identified as a key enabler of the growth of the artificial intelligence industry (Hall and Pesenti 2017).

## Governance, Content, and Data

The web is a socio-technical construct (Hall and Pesenti 2017), and as such its effects ripple through its embedding societies as emergent macro-level phenomena such as the formation of "crowds" as a response to real-world

events, the spread of emotional contagion, the site of opinion markets that affect the results of a real-world event such as an election, and the host of data marketplaces where data can be traded for monetary benefits while compromising personal privacy. It follows that, on top of the web's technical development and its open standards, those with an interest in governance need also to consider content usage and access patterns to understand issues highlighted by DeNardis (2010)—digital equality (Hargittai 2010), social media (boyd 2008; DeNardis and Hackl 2015), identity (Turkle 2005), knowledge production (Benkler 2006), Web 2.0 critiques (Lanier 2010), and copyright and information intermediaries (Vaidhyanathan 2007). Data can provide deeper insights about large-scale populations in real time, opening out research questions about the social behavior germane to a user-centric view of Internet governance (Dutton and Peltu 2005), on issues such as social and political movements; political participation and trust; crisis prevention, preparedness, response management, and recovery; individual, group or community, and national identities; and personal, local, national, and global security (De Roure 2014). To take one example, Wikipedia's underlying technical platform remains more or less the same as it was in the beginning, but the way people interact through it has varied significantly over the years (Kittur et al. 2007), and so the regulation of Wikipedia cannot simply be a matter of developing protocols.

A rich literature on Internet governance is available that describes various perspectives, including governance of platforms, layers of Internet infrastructure and applications where governance needs to be separated, distributed governance based on geopolitical boundaries, and data governance. Some of these also highlight how effective governance policies require working at a variety of scales, from the micro level of detail of individual protocols like HTTP (hypertext transfer protocol)[1] or HTML (hypertext markup language)[2] to macro-level emergent behavior such as blogging, spamming, or e-commerce and how the social phenomena emerge onto, diverge within, and submerge into it.

For example, Google knows much more about us than we know about Google (Hall and Pesenti 2017). Even if we are not involved in any power play or asymmetrical encounter with Google, such asymmetry has ramifications. This has sparked debate about user privacy, unwanted profiling of customers (people) by technical giants, and tracking of their online activities, preferences, and location (Hildebrandt 2016; Zuboff 2019). These

mandate governance policies to be inclusive of these new forms of interactions and relationships on the web.

Because of these social and political imperatives, a shift in the traditional Internet governance view has been proposed in the Tunis Agenda, successor the UN Working Group on Internet Governance, and elsewhere (Wagner 2016). Whereas the traditional view of governance focused more on the technical functions and standards required to keep the Internet open, unfragmented, stable, resilient, and secure, the UN group advocates "the development and application by governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet" (Wagner 2016). This will become more critical because of the rapid change and maturing of information technologies, including machine learning, natural language processing, face recognition, robotics, blockchains, and cryptocurrencies. Ultimately these technologies will converge with the Internet of Things (IoT), and as they do the interrelationship between humans and machines will pose unprecedented challenges for human societies and how they are governed (Fry et al. 2015). Meanwhile, many social activities increasingly take place online using the functionality of networks via commonly used connected devices, creating what have been called social machines, which themselves generate important quantities of data (Shadbolt et al. 2019).

Hence, an infrastructure is critical to overcoming the main barrier of web data collection and analytics essential for evidence-based study on the web. In addition, the speed at which data interactions occur on the web means that the data become obsolete and outdated at a rapid pace even if regular snapshots are taken (Hall et al. 2013).

## Governance Challenges for a Distributed Data Infrastructure

### The Need for an Infrastructure

Data have become the new fuel empowering decision-making in almost every sector. Governments make a wealth of public data openly available (on sites such as www.data.gov, United States; www.data.gov.uk, United Kingdom; data.gov.in, India; and www.data.gov.au, Australia) under different licensing based on their use. But observing a restricted number of siloed datasets that are deemed nonsensitive provides a narrow view to any problem

or issue. The real value of data comes from the broader perspective of multiple datasets brought together around a specific question or issue and from a range of sources (Fry et al. 2015), which suggests the idea of a platform to bring together data for the purpose of analysis and interrogation that furnishes methods and tools that enable researchers to locate, analyze, compare, and interpret information consistently and reliably (Hall et al. 2013; Hall et al. 2014; Tinati et al. 2015; Tiropanis et al. 2013).

It goes without saying that a centralized data store cannot possibly scale; the model proposed here is a platform where anyone with a dataset who wishes (and has the right) to share it could display the metadata to enable the data to be discovered by potential users. We call such rights holders data owners as a shorthand—they include data controllers whose data are personal under European data protection legislation, although the data shared via the Web Observatory need not be personal. Such data owners could, in principle, be anyone able to exercise rights to share, including governments, corporations, nongovernmental organizations, educational institutes, scientists and academics, or even private individuals (for instance, someone who wished to consider sharing his or her medical data with selected health care providers or researchers). However, joining the Web Observatory and advertising metadata does not mean that the data owner is obliged to share; the owner remains in control (and in particular, a controller of personal data would remain the data controller) and makes the decision to share. The Observatory is a distributed infrastructure to enable data discovery and sharing, not an automatic distributor of data. Hence, sharing of the data takes place *only* when the dataset owners receive a request and *only* in accordance with the owners' constraints, legal and ethical requirements, and business models. The governance of individual datasets remains with their owners; the responsibility of the infrastructure would be to provide protocols to support data discovery and sharing. To answer open questions and ensure open access, institutions owning data, or with common interests in sharing data, need to come together within the Web Observatory to build an active community engaging in experimentation and innovative problem-solving, involving the generation and sharing of both qualitative and quantitative data for evidence-based decision-making (Verhulst et al. 2014).

When data are especially sensitive (and this is not only personal data—consider, for instance, geologic data relevant to a fracking inquiry, which raises no privacy concerns), it may be that the data owner does not consider sharing

it at all. The owner might accept and process queries from third parties and return the results, perhaps using differential privacy techniques to ensure that the queries in aggregate aren't disclosive. Or the owner may even refuse access to the data but may post visualizations of results from processing.

Indeed, any data owner could allow access to visualizations of data whether sensitive or nonsensitive or access to tools or techniques that have proved valuable for data analysis or visualization. The Observatory can be the repository for anything valuable to the data analytics community, not simply metadata about valuable datasets. So, for example, the metadata describing transcripts of sessions of the Internet Governance Forum described by Cogburn in chapter 9 might be posted to the Web Observatory; those who could add value to those data could contact, via the Observatory, the rights holder and ask for access. Similarly, the mailing lists whose analysis is described by Ten Oever, Milano, and Beraldo in chapter 10 could be discovered. The mailing lists may be sensitive, so sharing might take place only under very specific conditions (e.g., the terms and conditions of access might preclude publishing direct quotes). Meanwhile, limited access could be granted to the commercially sensitive data held by major search engines and social media companies, discussed by Jørgensen in chapter 8—if not to the data itself, at least to summaries, statistics, and visualizations, or perhaps such companies might accept a limited number of queries from academics and policy makers without compromising any comparative advantage the data are perceived to confer. This could be especially valuable for Internet governance, given the importance that such sites have in the ecosystem of the Internet. The value of this kind of eclecticism would help address difficulties in studying the private sector, multistakeholder governance, and overstudying open systems (see DeNardis's chapter 1).

Though a distributed data infrastructure could and should be open to private and public actors, an early win could be the augmentation of current government digital platforms and open datasets. Increasingly, digital governments require access to data services that are internally and externally produced, which often creates a complex ecosystem of social systems (Tinati et al. 2015). Thousands of datasets are now openly available through open data portals describing local businesses, city-sensor networks, and live transport and traffic data. But not all government data can be opened for access so easily, and not all the data governments use are generated or curated by them, what with outsourcing and other policy delivery

partnerships with government. In this context, there is an opportunity for a wrapper to expose and share government data, which importantly allows third parties to access and produce analytical and visual representations of the data while still retaining access control (Tinati et al. 2015). Furthermore, a data infrastructure would offer an opportunity for governments to link to nongovernmental datasets, enriching existing data and providing new insights not originally envisioned (Hall et al. 2013; Tiropanis et al. 2013).

From the point of view of Internet governance, governments and nongovernmental actors could observe socio-technical phenomena from such an infrastructure via data from domains using web protocols (e.g., social media, crowdsourcing platforms, health care, city sensors). Data could be made available for those needing to make evidence-based policy decisions relating to Internet governance without exposing it as open data or indeed without the data owner having to surrender control over access at all. The Observatory provides the means for discovery and communication between potential consumers of the data and their owners and may itself be open or closed to new members. To the extent that the data are sensitive, the Web Observatory will inherit all the legal, ethical, and political issues relating to sharing sensitive data. The Observatory infrastructure does not solve these problems, but because control over the data remains with the data owners, neither does it create new information flows over which nobody has any control. In particular, it remains the decision of the data owners whether to share sensitive data (e.g., about a smart city) with someone who has requested access, to restrict its use for Internet governance purposes only, or alternatively, to share with a wider range of data consumers. To the extent that potentially sharing data about socio-technical phenomena has implications for wider society, the proposed Observatory infrastructure minimizes those implications. As a consequence, data governance, access to datasets, usage tracking to understand derivations of datasets, and user trust are major concerns for such an ecosystem. In the rest of this section, we describe these concerns in detail.

### Challenges

An important aim is to provide strong support for *user-centric transparency* by enhancing end users' awareness about their choices for data sharing, dependent on interactions with other stakeholders and the sensitivity of

datasets, requiring case-by-case analysis of the use of data. Crafting the terms and conditions of data sharing requires addressing the questions of liability and responsibility of data sharing on the part of operators, data-sharing companies, and other stakeholders. The growing imbalance of power created between citizens and companies through the privileged access that corporations have to information on our collective social lives is set to become an increasingly pressing social and political issue (Davies 2013; O'Hara 2015; and see also Jørgensen, chapter 8). Some key challenges need to be addressed by anyone proposing to share data:

*Privacy* is critical for understanding data processing on shared datasets of personal data (or data derived from personal data), with cultural context an important consideration. Global rules may not be applicable, or sufficient, at the local scale (O'Hara 2019).

*Open data* and the field of data integration generally has raised questions of interoperability, transparency, accountability, and reusability.

*Trust in the integrity of data* is becoming even more important with automated data collection through smart devices and analytics being performed by a range of actors (even before we consider issues concerning the increasing tendency of malign actors to attempt to mislead—an economy that supports fake news farms could equally support fake data farms). Understanding the provenance of data and measuring its quality are key challenges.

*Sovereignty* has made the questions of where data are stored and accessed especially relevant in the context of cloud storage, especially in light of increasing moves toward data nationalism (Chander and Lê 2014–2015).

*Uncertainty* about issues such as liability if data are misused and the security of data following a share has been identified as a key blocker to data sharing (Hall and Pesenti 2017).

An infrastructure for amassing data for evidence-based reasoning about Internet governance will need to bring together diverse communities for data sharing and reuse across geopolitical and application contexts and across public, private, and nonprofit sectors. These are questions of corporate social responsibility, as well as the ethical and legal consequences potentially faced by corporations involved in these controversies (Dutton and Peltu 2005). The challenges laid out here are even more critical with the rapid growth of technologies such as the IoT and artificial intelligence, which promote machine-to-machine interactions and capture a huge amount of potentially personal data about individuals.

Data that are personal lead to much concern over the expansive digital surveillance practices of some countries, the rise of nation-specific data-localization policies that cite privacy concerns as a justification for requiring providers to store data within national borders (Chander and Lê 2014–2015), and emerging and controversial policies such as the right to be forgotten ruling in the European Union (DeNardis and Hackl 2015; O'Hara and Shadbolt 2015; O'Hara, Shadbolt, and Hall 2016). The emphasis of the EU General Data Protection Regulation on a notion of data protection delineated by a rich jurisprudence contrasts with countries like India that have yet to define privacy and countries like China where paternalism plays a larger part in policy than supporting the rights of autonomous individuals. In a data infrastructure, individual units are likely to align to local policies and laws, which may reduce legal concerns but at the same time raises the concern of interoperability across the network (O'Hara and Hall 2018).

Data, communications metadata, or data about networks provide a lot of actionable and personally identifiable information, especially when augmented with auxiliary data. The routinization of extensive metadata collection as well as contextual content analysis is a fundamental departure from the Internet's original end-to-end design (of locating intelligence at end points, technical neutrality toward packet contents, and using IP addresses simply as virtual identifiers) (DeNardis 2014). Moreover, identification technologies also collect information on the hardware for using the platform (device information) and software information, including browser type (software footprint) (DeNardis 2014). To address concerns about this, which will be essential to foster wider trust of the Web Observatory infrastructure, each resource on the infrastructure would need to be annotated with appropriate metadata to restrict access and use of the dataset and have the ability to notify the data owners and publishers in real time about possible privacy breaches and would need to be protected by a clear and understandable set of terms and conditions to prevent disclosure of personal information.

Finally, we note that if people are to have any control over their personal data, they need rights over the data and transparency about what is happening to it. But the exercise of these individual rights is truly effective only if an organization's technology is fully responsive to them and has the right functionality embedded in it. The core individual data protection rights in the General Data Protection Regulation are the "right of access," "right to rectification," "right to erasure" (or the right to be forgotten), "right to

restrict processing," "right to data portability," and "right to object" (GDPR Individual Rights [n.d.]). In a functional sense, these rights require the technology to connect individuals to their personal data, the technology to classify the data with respect to the purpose of use or processing and for mapping the full data life cycle, the technology to make it searchable, and the technology to rectify it and perform erasure and anonymization. These are complex requirements and may not easily be incorporated, but at a minimum one would expect the implementation of provenance-tracking mechanisms to track data and application use to bring some measure of accountability to the system.

## The Web Observatory

Clearly, the development of such a global infrastructure for data sharing is a major research question. In this section, we describe the concept of a web observatory that meets the infrastructure requirements set out in the previous section, as a distributed global resource in which datasets, analytical tools, and cross-disciplinary methodologies can be shared and combined to foster web science: the emerging interdisciplinary field of the study and engineering of large-scale distributed information systems, particularly the World Wide Web, with a focus on their cocreation by human users and participants (Hall et al. 2014; Shneiderman 2007; Tiropanis et al. 2013). The aim of the Web Observatory is to provide a locus for subject-centric management of data, to complement the current paradigm of organization-centric data management (Van Kleek et al. 2014) and support longitudinal web-science analysis.

Hall et al. (2013) proposed the idea of the Web Observatory[3] to bootstrap the analytics to create evidence of newer concerns emerging from collated and archived data on the web. It provides the capability to curate datasets and aggregate interaction data (web clickstreams, dialogue from crowdsourcing platforms), descriptive data (demographics and geospatial data), behavioral data (usage history), and attitudinal data (opinions, preferences of stakeholders) (De Roure 2014; Hall et al. 2013; Tiropanis et al. 2013). It provides infrastructural support in the form of analytical tools and datasets to investigate methods and mechanisms by which people, as a collective society, could be effectively studied in academic research settings through analysis of the archival information traces they created online (Van Kleek

et al. 2014). It enables users to share data with each other, while retaining control over who can view, access, query, and download their data (Tinati et al. 2015). Web Observatory infrastructure would certainly enable the evolution of social machines as means of bottom-up coordination of problem-solving at scale (in terms of both providing the data to enable the academic study of social machines and supporting their operation where data are needed as input [Shadbolt et al. 2019]).

A web observatory infrastructure can be a generic set of desiderata. However, in this section we report on an attempt to implement these ideas in a specific system, *the* Web Observatory, with a definite article. The effort to create the Web Observatory has involved many different communities and organizations, including the Web Science Trust network of laboratories (WSTNet),[4] other major research groups in this area, government agencies, public sector institutions, and industry (Hall et al. 2016; Price et al. 2017). The Web Observatory has formed partnerships with the World Wide Web Consortium (W3C), the Open Data Institute (ODI) in the United Kingdom, the Fraunhofer Institute for Open Communication Systems (FOKUS), the Web Foundation, and a growing list of industrial collaborators (Tiropanis et al. 2013). Also, researchers affiliated with the Web Observatory at the University of Southampton have cooperated with the University of South Australia to support data-driven government policy making for the Adelaide government (Fry et al. 2015). The Web Observatory, which is based at Southampton, has provided access to multiple government and academic datasets to answer questions concerning the provision of public services on the basis of the demographic landscape of Adelaide's neighborhoods.

The Web Observatory contains metadata, catalogs of data, visualizations, analyses, and tools. Hence, if standards were established to express metadata (e.g., using linked data formalisms), the Web Observatory could be distributed in a network, each node of which could be located remotely, because of the interoperability that would result. The Web Observatory therefore operates as a decentralized, distributed infrastructure for sharing data and analysis (Tinati et al. 2015; Tiropanis et al. 2013) on the Web. It is, in other words, a network of web observatories (we refer to observatories within the network as *nodes*). The Southampton University Web Observatory (SUWO) mentioned in the previous paragraph is one node of the network.

Appropriate standards to enable the discovery, use, combination, and persistence of datasets and tools are being developed in the W3C Web

Observatory community group[5] (Hall et al. 2014), established to foster discussions on standardization requirements essential to enable interoperability between available resources and on identifying the opportunities for industry and global government agencies to contribute large-scale systems, expertise, and datasets.

### Architecture of the Web Observatory

Figure 6.1 describes the different architectural components of the Web Observatory. The Web Observatory harmonizes data sources from, for example, social media, open data repositories, and Internet archives. An interoperable catalog makes the metadata available in compliant standards. These data are then available to researchers through the visualization and analytical tools and methodologies available on Web Observatory nodes. In making this web of observatories possible, the following principles are required in the deployment of Web Observatory nodes or instances (Tiropanis et al. 2014):

- Resources related to the Web Observatory (projects, datasets, analytical applications, and people) need to have unique identifiers, preferably uniform resource identifiers or application programming interface (API) access points.
- There need to be explicit links between analytical applications and the datasets that they use.
- There need to be explicit links between Web Observatory resources and related use, scholarship, and discourse.
- Metadata should be published for all available resources in a Web Observatory instance to enable search.
- Datasets and analytical applications hosted or listed on a Web Observatory node can be public or private; the publisher needs to control who gains access to them.
- It should be possible to enable access for identified individuals (or to applications using their credentials) to specific datasets or applications hosted in local or remote datasets.
- It should be possible to support distributed queries across Web Observatory instances and to make computational resources on each instance available to that end.

The Web Observatory architecture is implemented as a reverse proxy server[6] enabling the data owners to host datasets at their local sites but also

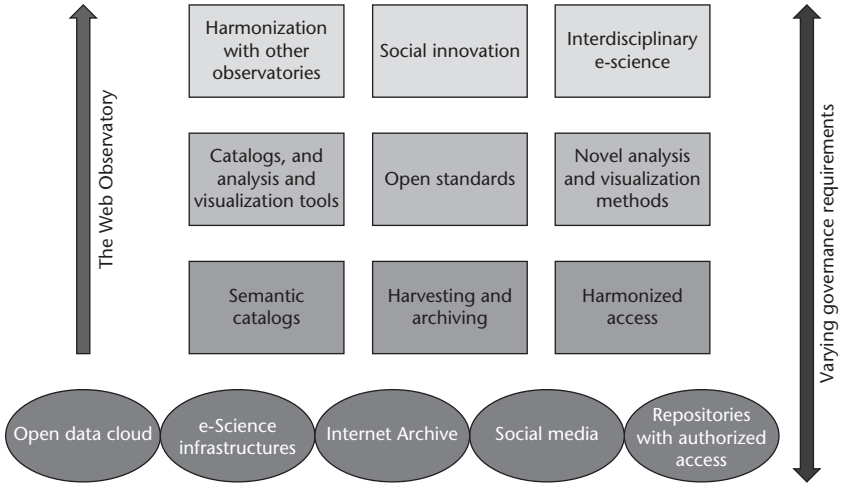


**Figure 6.1**
Web Observatory components and governance requirements.

allowing the observatory users to discover it through metadata search (Price et al. 2017; Wang et al. 2017). It adopts a decentralized architecture in which resources can be accessed in situ without the additional need of gathering them centrally. Users can publish to the Web Observatory resources from outside, and data generated within the Web Observatory such as system logs and user information are made available as datasets and listed in the Web Observatory itself (Siow, Wang, and Tiropanis 2016). The discoverability of the metadata is defined by the data owner or publisher on the Observatory as private or public. A Web Observatory node may be initiated by individuals for use as a personal observatory (Siow, Tiropanis, and Hall 2016). Different observatory nodes are not obliged to host the same software but should comply with the underlying principles of the Web Observatory. A distributed search tool is available to allow the end users to search for datasets and tools across the distributed Web Observatory using the metadata attributes.

The following core functionalities of the Web Observatory infrastructure support engagement across communities and emphasize ethical and legal rectitude.

**Metadata of datasets and analytical tools** Metadata describing the listed resources and projects is published. In this way, descriptions of resources can be harvested and listed in other Web Observatory nodes or web-based

resources. If a listed application (or visualization) uses a listed dataset, the link is explicitly mentioned in the metadata for the application. Metadata not only facilitate accurate discovery but also provide links between related resources. For example, datasets and the analytics produced from them are linked at the metadata level. By traversing these links, users can browse through a large network of analytical resources more effectively (Siow, Wang, and Tiropanis 2016). Once related resources are identified, users and applications can access them via the Web Observatory in a secure way.

**Access control for sharing datasets** The Web Observatory allows users to access datasets and analytics via a web interface. It also provides an API for applications to programmatically access analytical resources (Madaan et al. 2016; Siow, Wang, and Tiropanis 2016; Tinati et al. 2015; Tiropanis et al. 2014). The API is protected by OAuth 2.0[7] to ensure that access to any resource is controlled by the resource owner and can be delegated to applications without having to reveal the underlying credentials. Applications can also access multiple datasets simultaneously, through the API, to perform complementary analytics and information fusion. It is also possible to access live data streams and build real-time analytics (Madaan et al. 2016; Siow, Wang, and Tiropanis 2016; Tinati et al. 2015; Tiropanis et al. 2014). The Web Observatory lets users list or host datasets that are public or private. Access to private datasets is managed by the data owner, who hosts them on a Web Observatory node. Because access to datasets can be restricted, access to applications that make use of those datasets must be restricted as well (Tiropanis et al. 2013; Tiropanis et al. 2014). Resources can be queried using the Web Observatory API, which uses a JavaScript Object Notation (JSON) structured query language, and the mappings to the types of data stores are handled by the Web Observatory API and processed on the server side. This API acts as a secure middle layer between dataset locations and the end-user connections (Wang et al. 2017). The Web Observatory deploys an access control mechanism so that private datasets and analytics containing sensitive information are protected and accessible only to authorized users and applications.

**Provenance for data quality** The Web Observatory supports provenance of datasets and also their derivatives using the W3C PROV standard. This is achieved by adding provenance documents and linking datasets to provenance uniform resource identifiers in the Web Observatory administration interface. PROV-AQ[8] provides a pingback mechanism to discover provenance

information and has been integrated into the Web Observatory infrastructure. This helps data publishers understand how the dataset has been used once it has been created and further supports usage tracking, transparency, and ultimately, users' control over how their datasets are used.

**Terms and conditions for innovation on the Web Observatory** The PROSENT model (Wilson et al. 2016) considers whether the downstream use of datasets is in agreement with the data consent given by the data publishers on the Web Observatory platform. However, Wilson et al. (2016) also highlight the complexity of determining the consent violation and safeguarding the resources available on the Web Observatory owing to its distributed nature.

### The Web Observatory and Data Governance Challenges

The distributed Web Observatory architecture is designed to address the governance challenges raised in the previous section. The important ingredient is the distributed nature of the resource, which allows data controllers to retain control, to determine how discoverable their data are, and ultimately to decide who gets access and under what conditions. In particular, we draw attention to the following aspects.

**Privacy** The data owner's specified metadata description is available on the Web Observatory, and any access request is evaluated against the visibility of the resource. The data owners and publishers define partial or complete metadata for the discoverability of their resources on the Web Observatory platform.

**Openness** Each observatory node defines its own sharing attributes on specific datasets. The Observatory uses reverse proxy, and the datasets are stored in the jurisdiction where they originated. The datasets available can be open or private as specified by the data owner or publisher.

**Integrity** The provenance of analytical tools and datasets and the history of use provide a basis for trust for end users because they can see the derivations of their datasets, their transformation, and the context in which they have been used. Access patterns and usage tracking can be used to quantify the liability and accountability of stakeholders for data use, clarifying legal issues.

**Sovereignty** The Observatory connects individual and organizational stores. It does not mandate that end users store the data within the Observatory node; rather, the datasets can be hosted at the end user's site.

**Uncertainty** The support for access controls, provenance, and terms and conditions for innovation all help mitigate the uncertainty about liability and so forth. In the future, one could imagine the Web Observatory producing pro forma agreements for sharing and a set of best practices (cf. O'Hara 2019). However, given the international and cross-jurisdictional nature of the exercise, it is unlikely that this problem can be solved entirely by infrastructure.

## Conclusion

As the cyber and real worlds become harder to treat separately with every new technology adoption, data are an increasingly important input to evidence-based policy making and to Internet governance questions that often arise because of the innovative ways people interact with the web (a 2019 project explores the infrastructural, technological, and legal issues involved with creating the Web Observatory to share data from the IoT[9]). This chapter describes how the diverse communities of academia, industry, and the public sector can be brought together to understand the evolution of the web and cocreate methodological evidence using the Web Observatory ecosystem to inform web and Internet governance policy development. The Web Observatory methodology itself, however, raises many governance issues of its own, and this chapter describes how its distinctive set of principles and freedom of data governance supports safe, ethical, and legal access to Observatory datasets and tools to support longitudinal data analysis and policy making across time, geopolitical boundaries, and topics.

Knowledge inferred from data in the Web Observatory about security, surveillance, and the promotion of state propaganda through social media platforms could provide evidence to policy makers to understand the effects of different national and supranational views of how the Internet should be controlled and developed. It has been argued that some governments are warping the Internet by attempting to align information flows with their jurisdictional boundaries (Chander and Lê 2014–2015; Mueller 2017; and see Mueller and Badiei's chapter 2). Beyond that, there is evidence of four different visions of the Internet coming to the fore, which are affecting its governance, driven by ideals such as openness, privacy, paternalism, and commercial imperatives. We also need to take account of government-sponsored hacking behavior occurring at scale, an ideological view that

counterbalances the state-centric view (O'Hara and Hall 2018, 2020); these are the conflicting values mentioned specifically as a challenge in chapter 1. Resolving these conflicts is critical in the current fragmented scene of Internet governance, where we need to contrast national lenses with the multistakeholder view and with views such as that of the EU, which is using legislation such as the General Data Protection Regulation to protect individuals' online data while simultaneously projecting political power beyond its boundaries. Of course, the matter is complicated by the possibility that the Web Observatory itself requires certain ideological assumptions (for instance, openness) as a precondition for operating at all.

We have mentioned en passant a number of the challenges to Internet governance research set out in chapter 1; it should be clear by now that an architecture like the Web Observatory, by making the invisible visible through furnishing data, will help address most—although not all—of those challenges to some degree at least. We are left only with the concluding suggestion that the Web Observatory itself is one of the tools that need to be created, supported, and honed to enable Internet governance research, as cited in the final challenge. However, the Observatory is not only a tool; it demands a community of institutions prepared to make their data visible if not open, to cooperate in the research task, and to develop sufficient trust to facilitate the use and reuse of data. To that end, the status of the Web Observatory as a tool, as with any practicable technology, depends crucially both on architecture and on practice.

## Acknowledgments

This work is supported by SOCIAM: The Theory and Practice of Social Machines. The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1 and comprises the Universities of Southampton, Oxford, and Edinburgh.

## Notes

1. "Hypertext Transfer Protocol—HTTP/1.1," IETF Network Working Group, Request for Comments 2616, June 1999, https://www.w3.org/Protocols/rfc2616/rfc2616.html.

2. "Hypertext Markup Language—2.0," IETF Network Working Group, Request for Comments 1866, November 1995, https://tools.ietf.org/html/rfc1866.

3. "Web Observatory," accessed October 6, 2019, https://webobservatory.soton.ac.uk/.

4. "WSTNet Laboratories," WebScience Trust, accessed October 6, 2019, http://webscience.org/wstnet-laboratories/.

5. "Web Observatory Community Group," World Wide Web Consortium, accessed October 6, 2019, https://www.w3.org/community/webobservatory/.

6. "What Is a Reverse Proxy Server?," NGINX, accessed October 6, 2019, https://www.nginx.com/resources/glossary/reverse-proxy-server/.

7. "OAuth 2.0," OAuth.net, accessed October 6, 2019, https://oauth.net/2/.

8. "PROV-AQ: Provenance Access and Query," W3C Working Group Note, April 30, 2013, https://www.w3.org/TR/prov-aq/.

9. "IoT Observatory," Themes, PETRAS, accessed October 6, 2019, https://www.petrashub.org/portfolio-item/iot-observatory/.

**References**

Banyan. (2017, October 26). Is the world getting Myanmar wrong? *The Economist*. Retrieved from https://www.economist.com/news/asia/21730684-future-not-long-ago-deemed-bright-now-feels-bleak-world-getting-myanmar-wrong

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. New Haven, CT: Yale University Press.

boyd, d. (2008). Facebook's privacy trainwreck: Exposure, invasion, and social convergence. *Convergence, 14*(1), 13–20.

Chander, A., & Lê, U. (2014–2015). Data nationalism. *Emory Law Journal, 64*(3), 677–739.

Davies, T. (2013, October 9). Web observatories: The governance dimensions [Blog post]. Retrieved from http://www.opendataimpacts.net/2013/10/web-observatories-the-governance-dimensions/

DeNardis, L. (2010). The emerging field of Internet governance. *Yale Information Society Project Working Paper Series*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1678343

DeNardis, L. (2014). *The global war for Internet governance*. New Haven, CT: Yale University Press.

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy, 39*(9), 761–770.

De Roure, D. (2014, March 20). Web Observatories, e-research and the importance of collaboration. WST Webinar Series. Retrieved from https://www.slideshare.net/davidderoure/web-observatories-and-eresearch

Dutton, W. H., & Peltu, M. (2005). The emerging Internet governance mosaic: Connecting the pieces. *SSRN*. Retrieved from https://ssrn.com/abstract=1295330

Fry, L., Hall, W., Koronios, A., Mayer, W., O'Hara, K., Rowland-Campbell, A., Stumptner, M., Tinati, R., Thanassis, T., & Wang, X. (2015). Governance in the age of social machines: The web observatory. The Australia and New Zealand School of Government. Retrieved from https://eprints.soton.ac.uk/378417/

GDPR Individual rights. (n.d.). Retrieved December 6, 2017, from https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/

Hall, W., Hendler, J., & Staab, S. (2017). A manifesto for web science@ 10. *arXiv*. Retrieved from https://arxiv.org/abs/1702.08291

Hall, W., & Pesenti, J. (2017). *Growing the artificial intelligence industry in the UK*. Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. Part of the Industrial Strategy UK and the Commonwealth. Retrieved from https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

Hall, W., Tiropanis, T., Tinati, R., Booth, P., Gaskell, P., Hare, J., & Carr, L. (2013, May 1–3). *The Southampton University Web Observatory*. Paper presented at the First International Workshop on Building Web Observatories. *ACM Web Science*. Retrieved from https://eprints.soton.ac.uk/352287/

Hall, W., Tiropanis, T., Tinati, R., & Wang, X. (2016). Building a global network of web observatories to study the web: A case study in integrated health management. *Qatar Foundation Annual Research Conference Proceedings, 2016*(1), 3092.

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, 2014*(96), 29–30. Retrieved from https://eprints.soton.ac.uk/361437/1/ERCIM-69_p29-30.pdf

Hargittai, E. (2010). Digital na(t)ives? Variation in Internet skills and uses among members of the "net generation." *Sociological Inquiry, 80*(1), 92–113.

Hildebrandt, M. (2016). *Smart technologies and the end(s) of law novel entanglements of law and technology*. Cheltenham, UK: Edward Elgar.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web, 1*(2), 19.

Lanier, J. (2010). *You are not a gadget*. New York, NY: Knopf.

Madaan, A., Tiropanis, T., Srinivasa, S., & Hall, W. (2016). Observlets: Empowering analytical observations on Web Observatory. In *Proceedings of the 25th International*

*Conference Companion on World Wide Web* (pp. 775–780). Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from https://doi.org/10.1145/2872518.2890593

Mueller, M. (2017). *Will the Internet fragment?* Cambridge, UK: Polity Press.

Ngamassi, L., Ramakrishnan, T., & Rahman, S. (2016). Use of social media for disaster management: A prescriptive framework. *Journal of Organizational and End User Computing, 28*(3), 122–140. http://dx.doi.org/10.4018/JOEUC.2016070108

O'Hara, K. (2015). Data, legibility, creativity…and power. *IEEE Internet Computing, 19*(2), 88–91.

O'Hara, K. (2019). Data trusts: Ethics, architecture and governance for trustworthy data stewardship. Retrieved from http://dx.doi.org/10.5258/SOTON/WSI-WP001.

O'Hara, K., & Hall, W. (2018). *Four Internets: The geopolitics of Internet governance* (Paper no. 206). Centre for International Governance Innovation. Retrieved from https://www.cigionline.org/publications/four-internets-geopolitics-digital-governance

O'Hara, K., & Hall, W. (2020). Four Internets. *Communications of the ACM, 63*(3), 28–30. http://dx.doi.org/10.1145/3341722

O'Hara, K., & Shadbolt, N. (2015). The right to be forgotten: Its potential role in a coherent privacy regime. *European Data Protection Law Review, 1*(3), 178–189.

O'Hara, K., Shadbolt, N., & Hall, W. (2016). *A pragmatic approach to the right to be forgotten*. Global Commission on Internet Governance Paper Series (Paper no. 26). Centre for International Governance Innovation/Chatham House. Retrieved from https://www.cigionline.org/publications/pragmatic-approach-right-be-forgotten

Price, S., Hall, W., Earl, G., Tiropanis, T., Tinati, R., Wang, X., … & Groflin, A. (2017, April). Worldwide Universities Network (WUN) Web Observatory: Applying lessons from the web to transform the research data ecosystem. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1665–1667). Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Shadbolt, N., O'Hara, K., De Roure, D., & Hall, W. (2019). *The theory and practice of social machines*. Cham, Switzerland: Springer.

Shneiderman, B. (2007). Web science: A provocative invitation to computer science. *Communications of the ACM, 50*(6), 25–27.

Singh, J., & Haziq, M. (2016). *Myanmar's Rohingya conflict: Foreign jihadi brewing*. S. Rajaratnam School of International Studies commentary CO16259. Retrieved from https://www.rsis.edu.sg/rsis-publication/icpvtr/co16259-myanmars-rohingya-conflict-foreign-jihadi-brewing/#.Xi7QmzKgLIU

Siow, E., Tiropanis, T., & Hall, W. (2016, October). *PIOTRe: Personal Internet of Things repository*. Poster presented at the International Semantic Web, Demos, Japan.

Siow, E., Wang, X., & Tiropanis, T. (2016, May). Facilitating data-driven innovation using VOICE observatory infrastructure. In *Proceedings of the Workshop on Data-Driven Innovation on the Web* (p. 5). New York, NY: ACM.

Stevens, D., & O'Hara, K. (2015). *The devil's long tail: Religious and other radicals in the Internet marketplace*. London, UK: Hurst.

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277–1291.

Tinati, R., Wang, X., Tiropanis, T., & Hall, W. (2015). Building a real-time web observatory. *IEEE Internet Computing, 19*(6), 36–45.

Tiropanis, T., Hall, W., Hendler, J., & de Larrinaga, C. (2014). The Web Observatory: A middle layer for broad data. *Big Data, 2*(3), 129–133.

Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N., & Hendler, J. (2013). The web science observatory. *IEEE Intelligent Systems, 28*(2), 100–104.

Turkle, S. (2005). *The second self: Computers and the human spirit*. Cambridge, MA: MIT Press.

Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2018). Increases in depressive symptoms, suicide-related outcomes, and suicide rates among US adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science, 6*(1), 3–17.

Vaidhyanathan, S. (2007). The Googlization of everything and the future of copyright. *University of California Davis Law Review, 40*, 1207–1231.

van Kleek, M., Smith, D. A., Tinati, R., O'Hara, K., Hall, W., & Shadbolt, N. R. (2014, April). 7 billion home telescopes: Observing social machines through personal data stores. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 915–920). New York, NY: ACM.

Verhulst, S., Noveck, B. S., Raines, J., & Declerq, A. (2014). *Innovations in global governance: Toward a distributed Internet governance ecosystem*. Global Commission on Internet Governance Paper Series (Paper no. 5). Centre for International Governance Innovation/Chatham House. Retrieved from https://www.cigionline.org/publications/innovations-global-governance-toward-distributed-internet-governance-ecosystem

Wagner, F. R. (2016, November). The Internet governance ecosystem: Where we are and the path ahead. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web* (pp. 5–6). New York, NY: ACM.

Wang, X., Madaan, A., Siow, E., & Tiropanis, T. (2017, April). Sharing databases on the web with Porter Proxy. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1673–1676). Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Wilson, C., Tiropanis, T., Rowland-Campbell, A., & Fry, L. (2016). Ethical and legal support for innovation on web observatories. In *Proceedings of the Workshop on Data-Driven Innovation on the Web* (p. 1). New York, NY: ACM.

Wong, J. C., Safi, M., & Rahman, S. A. (2017, September 20). Facebook bans Rohingya group's posts as minority faces "ethnic cleansing." *The Guardian*. Retrieved from https://www.theguardian.com/technology/2017/sep/20/facebook-rohingya-muslims-myanmar

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. London, UK: Profile.