Legal and Ethical Considerations: Step 1b in Building a Health Web Observatory

Marie Joan Kristine T. Gloria
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
glorim@rpi.edu

John S. Erickson
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
erickj4@rpi.edu

Joanne S. Luciano
General Electric Global Res. & Dev. Center, Niskayuna, NY USA
joanne.luciano@ge.com

Dominic DiFranzo
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
difrad@rpi.edu

Deborah L. McGuinness
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
dlm@cs.rpi.edu

ABSTRACT

This paper explores the impact of health information technologies, including the Web, on society and advocates for the development of a Health Web Observatory (HWO) to collect, store and analyze new sources of health information. The paper begins with a high-level literature review from across domains to demonstrate the need for a multi-disciplinary pursuit when building web observatories. For as researchers in the social sciences and legal domains have highlighted, data carries assumptions of power, identity, governance, etc., which should not be overlooked. The paper then recommends example legal and ethical questions to consider when building any health web observatory. The goal is to insert social and regulatory concerns much earlier into the WO methodology.

Categories and Subject Descriptors

K.4.0 [**Computers and Society**]: General

General Terms

Web Observatory; Web Science; Health Web Science

Keywords

Web Observatory; Web Science; Health Web Science

1.INTRODUCTION

Information technologies, particularly the Internet, have rapidly transformed many aspects of our daily lives. Specifically, current research in multiple disciplines note the Web's impact on the health care industry, personal health care maintenance and health care policy. For example, communication scholars have highlighted the use of social media in health care indicating a growing trend in online participatory habits by U.S. adults [8] while science journals have turned their attention to covering novel methods such as crowdsourcing for drug discovery [5]. Given these advancements and discussions about the widespread impact of health information technologies in both research and clinical applications, a multi-disciplinary field has recently emerged to focus on these achievements and their future implications as they relate to the Web and health care issues.

The field of Health Web Science (HWS), a sub-discipline of Web Science [16], moves beyond discussions of the Web's impact on health care towards a deeper understanding of the design, structure, and the evolution of the Web itself. This is particularly important as the role of the consumer (e.g. user, patient, etc.) and producer converge and as medical experts, policymakers and researchers rely more on the Web to explore, share and analyze health related information for both medical application and research. HWS related activities include: semantic annotation and linking of health records and data; synthesis, curation, and discovery of health information on the Web; the structure and utilization of interactive social media sites, etc. [16]. In this paper, we further the HWS discussion by drawing attention towards the development of a Health Web Observatory (HWO) with a critical focus on recommendations to protect the privacy and authority of relevant, sensitive health data.

The paper proceeds as follows. First, we explore the growth of online health resources, their adoption by Internet users and the growth of personalized health data. We discuss that despite such new models for care, consumer reluctance and mistrust of health information online remains a common concern. The next section briefly explores current legal and technical solutions that help mediate issues of privacy and authority. We argue why Health Web scientists, in particular, must actively and deliberately protect the use of such sensitive personal data against misuse and security breaches. We assert that HWS researchers carry the unique burden to understand not just how the Web is used within the medical domain but to incorporate such learnings in the Web's continued structural evolution. We point to the emergence of Health Web Observatories (HWO) as a working case study for the development of best practices to address such concerns.

1.1E-Health v. Privacy

In the U.S. alone, 81% of adults use the Internet with 59% indicating they did so over the past year in search for health related information [9]. Of the same survey sample, 35% say they have gone online specifically to research a medical condition for themselves or on behalf of another person [9]. In addition, 34% of

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579210

U.S. online health seekers turn to the Web to read someone else's commentary or experience about health or medical issues on an online news group, website, or blog [11]. However, while many may turn to online sources, patients remain reluctant to share their own health information online and to trust the information found [3,7]. Bansal et. al (2007) seminal study on health information systems found that a user's intention to share health information online depends on their trust, privacy concern, and information sensitivity, which are "determined by personal dispositions like personality traits, information sensitivity, health status, prior privacy invasions, risk beliefs" [2]. Instead, clinicians continue to be the central resource for information or support (mostly offline) during serious health episodes for most U.S. adults [9]. Additional research and anecdotal references of key characteristics in the patient-physician relationship underscore the importance of privacy when disclosing personal data, such as medical history and diagnosis or sexual preference, etc. [1,17]. Legal scholarship also criticizes the expectation of privacy for personalized health data under the guise of intellectual property [19]. However, in the U.S., the law only protects data from unauthorized uses and disclosures only sometimes. Samuelson (1999) notes that "however intuitively powerful the notion of property rights in one's data may be, it is clear that in the U.S. the existence of some legally protectable interests in personal data in certain circumstances is not equivalent to a legal rule that a person has a property interest in her personal data" [19]. Moreover, a lack of adequate security measures coupled with heightened news coverage of medical data breaches over the past decade have fueled the distrust perceived by online health consumers [13].

More recently, the explosion of the "Internet of Things" (IOT)[1] supported devices, like the FitBit, RunKeeper, Mimo[2] or Google glasses, are reshaping our understanding of personal data management for health. Pew Research indicates that "seven in ten (69%) U.S. adults track a health indicator for themselves or a loved one" with over 50% of the respondents stipulating their use of some sort of record managing system such as paper or a web application [10]. Participants also noted that such tracking has changed their overall approach to health maintenance. This increase in personalized health data (also known as the quantified-self movement and the patient-driven health care model) allows for a deeper understanding of human behavior and individual health practices. It also, however, places enormous pressure on already ill-suited, inadequate regulatory and security measures that protect user privacy. To similar effect, science technology studies, media studies, surveillance and sociology scholars have all called for a reexamination of privacy that problematizes its relation to data, context and power. These perspectives provide a powerful understanding of how data may shape not just individual health practices, but system-level design, governance and policy.

1.2.Current Solutions

To address some of these concerns, both legal and technical solutions have since been introduced to extend protections for the consumer. Most notably, the U.S. Health Maintenance Organizations Act of 1973 (HMO) and the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA) provide Federal-level protection of a patient's privacy and detailed data use guidelines for the entire U.S. health care ecosystem. HIPAA protects "most 'individually identifiable health information' such as an individual's past, present, or future physical or mental health or condition whether held or transmitted on paper, electronically or oral" [12]. The increasing adoption of health information technologies as noted above motivated the inclusion of section 164.514(a) of the HIPAA Privacy Rule, which outlines the standard for de-identification of protected health information. Two de-identification methods are defined: "1) a formal determination by a qualified expert; or 2) the removal of specified individual identifiers as well as absence of actual knowledge by the covered entity that the remaining information could be used alone or in combination with other information to identify the individual" [12]. In summary, the HIPAA Privacy Rule sets limits and conditions on the uses and disclosures of sensitive patient information that may be made without the patient's authorization.

From the technology side, accidental disclosures, misuse of data, unauthorized intrusion of network systems etc., continue to dominate health information privacy research literature [1]. Unfortunately, advancements in health information systems security lags behind industry-driven technology development. For example, as IOT devices proliferate, we must formulate new questions such as: *whether the collection of health data is sufficiently protected from data breaches when stored by third-party organizations?*; *whether the data collected and stored is compliant with existing law?;* or *whether anonymity can be preserved by the system?* As such, new mechanisms of transparency and accountability of data use have been proposed as alternative solutions [22].

The flow of health information is complex, multi-layered and heavily regulated. And, while technology itself may resolve some security concerns through encryption, transparency or accountability mechanisms, privacy issues are too complex to be resolved by technology alone. Already, we see increasing debate over technology systems that could considerably conflict against current legal protections. For example, questions surrounding cloud service providers and compliance under the HIPAA privacy rule have yet to be resolved [6]. This is especially relevant given the accelerating production of IOT sensor data, which in large part are hosted by third-party cloud service providers. It is our position that the Health Web Science community must include a 360 perspective - technical infrastructure, public policy and social structure - in order to sufficiently understand health on and of the Web. One way to realize this is on the web observatory level where HWS researchers must make strides to ensure a trusted governance structure on both the technical and policy levels.

2.A CASE FOR THE WEB OBSERVATORY

The Web Science community is unique in its pursuit to understand the Web's complexity through a diverse set of semantically enriched tools and processes that enable collaboration across multiple disciplines. One tangible manifestation of this is the *Web Observatory*, defined as "a global data resource and distributed open analytics environment which enables live monitoring and

[1] For an extended discussion on IOTs, *please see* Swan, M. (2012).

[2] Mimo debuted at CES 2014 and is a baby monitoring "onesie." The outfit is tethered to the Internet and sends baby vital "updates" to an iPhone or Android device. More information can be found here <http://huff.to/1ghDvxm>

predictive modeling of the state of the Web and its evolution"[3] [21]. Concentrating on the medical domain, the potential contributions of a Health Web Observatory could be enormous. As the quantified-self literature suggests, our culture's recent shift towards a patient-driven health care model, facilitated by technologies like the Web, enables access to new data types that could improve personal health decisions, medical research and inform policy makers. The creation of a Health Web Observatory is the next critical step toward fulfilling what the IOT and quantified-self communities lack: a "data commons" that enables the community to collaborate and share data [20].

2.1Building a Health Web Observatory

The precise "recipe" for building a web observatory continues to be a topic of discussion for the Web Science community. Tiroponis et. al outline two activity streams for its development: 1) identify existing, related web observatories and archives, and 2) identify and share tools to visualize, analyze, and collect large distributed datasets [21]. Additional steps and tools may also be found on the TWC Web Observatory Portal.[4] We argue that the greatest challenge in building a WO is the design and implementation of a platform that will accommodate the future development of innovative and interesting, yet unknown, third-party applications. A WO must be structured and simple enough to replicate and be valuable, while also being flexible enough to encourage creativity and novel use. The TWC WSRC "template" is to build a web observatory for our initial purpose, then adopt or develop additional technologies such that our WO can be duplicated, understood, and reused for other users and different purposes. This iterative process draws from previous and current work developed by the TWC, which closely aligns with the Semantic eScience methodology [4].

To address legal and ethical concerns, Tiropanis et. al suggest the creation of a "forum to discuss an ethics framework on archiving and processing web data and relevant policies as well as the creation of a data licensing framework" [21]. At the time of this paper's submission, both suggestions remain largely unfulfilled by the Web Observatory community. Therefore, we see the development of a HWO, with its complex, sensitive and heavily regulated information flow, as an ideal use case to address legal and ethical questions right from the beginning. In the next section, we provide additional recommendations to insert within the WO framework. The need is framed into two segments - input and output - of the health information flow. This deliberate segmentation is simple to understand and better maps onto pre-existing public policy structures, which for the most part regulate either the flow-in or flow-out of health data. We limit our focus to the concerns surrounding privacy and the authority of data when building a HWO.

2.1.2What's your input and output?

The first step is to **determine the purpose** of this specific Health Web Observatory while defining types of users and the specific goals (e.g. will it be accepting and or propagating data?). Take for example, a HWO dedicated to breast cancer research. Potential users can vary widely from breast cancer survivors to academic researchers to medical practitioners; and can greatly impact the legal parameters of the inputs of your HWO. A HWO collecting and hosting public comment, such as tweets, comments and forum discussions about breast cancer, should make a conscientious effort to 1) understand the pre-existing privacy rules of the online community from which I am mining data; 2) know and understand acceptable governance practices of the community and to consider whether collecting the data violates these "unwritten" practices?; 3) understand whether the mining, hosting, and the propagation of the data violates any current privacy law (e.g. ECPA, CFAA); etc. Just as anthropologist strive to understand communities and cultures without much disruption, web observatory builders should also take into account the community practices from which they source data. This first step is critical in setting the tone for the observatory's own governance and infrastructure policies.

Following this, a researcher begins **gathering inputs and data sets (and creating data collection systems).** Here is where the first set of the legal and ethical considerations must be addressed by the researcher. Referring back to our example, a Web scientist collecting breast cancer survival stories to feature on the HWO should consider the following: 1) whether data re-published by the HWO contains user handle/identity, 2) the HWO can securely maintain anonymity, and 3) is the data from this community reliable and authoritative? In addition, a HWO should include defined rules for access to the inputs and who can use and produce outputs.

In addition to privacy concerns, studies point to a reluctance by online health consumers to trust and share personal health information online. One potential resolution can be found in the trust and authority literature. Specifically, the use of provenance within captured metadata can substantially mitigate such concerns (we discuss this further in detail below). We add that during a HWO build, one must take steps to **document where the data exists as well as the metadata/provenance for the dataset**. This includes understanding where the data originated, how it was produced, its unique identifiers, its units, etc. Not all collected data will explicitly contain such metadata; but it is equally important to capture such gaps in knowledge in order to inform others. Similarly, as one considers his/her outputs in a HWO, expressing post-collection metadata also needs to be incorporated. We provide Table 1 (see below) as a summary of the evaluations and considerations needed as one builds a HWO.

Table 1: A Review of Input and Output Evaluations

Input(s):	
Privacy:	• Is the data publicly accessed? or open government data? ◦ If mined from online social networks, will there be any personally identifiable information (PII)? ◦ Can anonymity of users be maintained? ◦ What are the pre-existing privacy rules for the online community from which I am mining data? ◦ What are the "unwritten" community governance practices; and will I be violating these practices if mining the data? • Who will have access to the HWO and available datasets?

[3] *See also*, International Open Government Dataset Search <http://logd.tw.rpi.edu/iogds_data_analytics>

[4] *See* TWC Web Observatory Portal < http://tw.rpi.edu/web/web_observatory> for more information

Input(s):	
Authority:	• Is the from an authorized and trustworthy source? • Does the dataset include metadata? If so, is it sufficient? Does it include: ◦ Original source(s) ◦ Regional information ◦ Notation(s) of modification(s) and timestamps? • Who will have access to the data? How will this process be managed?
Output(s)	
Privacy:	• Will the propagation of this data violate any pre-existing privacy laws? • Will the hosting of this data violate any pre-existing privacy laws? ◦ What is your home institution's current practice for compliance with government information requests? Will your WO follow the same practice? • Have all PIIs been removed from the data?
Authority:	• Have you captured all the necessary metadata? ◦ Who will have access to the metadata? • What mechanisms will be in place to manage third-party access to the data? ◦ Have you created a set of appropriate rules that comply with existing WO community standards? ◦ Have you created a set of appropriate rules that comply with existing public policy and regulations?

2.2Semantic Accountability Systems

Much work in semantic technologies has brought about alternative enforcement of both privacy and authority. One specific example are semantic accountability systems. Weitzner et. al (2006) formative work in the "Policy Aware Web" suggest a technical infrastructure that supports transparent and accountable data use on the Web as well as new mechanisms to support and enforce elements of public policy [22]. More importantly, we emphasize its capability to provide such protections on distributed platforms, much like that of a HWO. This approach "will provide users with accessible and understandable views of the policies associated with resources, enable agents to act in response to rules or on a user's behalf, thereby making compliance with stated rules easier" [22]. The original TAMI architecture featured general-purpose inferencing components - such as the Truth Maintenance System - to ensure privacy and provide accountability steps [22].

Once again turning to our breast cancer research HWO example, we consider the preliminary work of the TWC Mobile Health[5] project, which in the future could produce data shared with the breast cancer research HWO. The Mobile Health project focuses on gathering "requirements from health data and services providers to identify and refine requirements for representing health data in different systems" such as sensor-based health monitors [15]. More importantly, the project leverages semantic technologies to represent and reason over a variety of available sensor-data "to yield an integrated and easily re-purposable view of a user's health and activity state" [15]. Ideally, before we begin integrating associated data collected by the Mobile Health project, like one's heart rate readings after chemotherapy, it is essential that we review any appropriate considerations from above. For this case, it is imperative to anonymize the data and to make sure all personally-identifiable information have been deleted. In addition, since these are two distributed platforms, we would need to establish access permissions for both input and outputs for the breast cancer research HWO. The implementation of a TAMI architecture, or any semantic accountability system, can help manage and automate the necessary permissions, capture metadata and preserve anonymity.

Since the introduction of the TAMI architecture, additional technologies have been developed and refined to aide in similar situations. The Accountability in RDF (AIR) language is one example that "supports nested activation of rules, negation, closed world reasoning, scoped contextualized reasoning, and explanation of inferred facts" [14]. Additionally, a conceptual model created by Oberoi *et. al.* which leverages the Semantic Web and OWL, exploits control and customized access across multi-user, multi-database systems [18]. It does so by separating the flow into 1) data for processing the query and 2) data for end use as a result of the query [18]. Thus, when information is searched, authorization must be met for access while any violation of that access can be identified and isolated to the specific request without compromising the integrity of the entire dataset. It is clear that semantic mediation systems have the potential to address both privacy and authority concerns throughout the entire flow of information. Yet, it is under-utilized by many, including the WO community.

3.CONCLUSION

Health information technologies have dramatically changed the way most U.S. adults understand and approach their personal health decisions. The Web, in particular, has become an invaluable resource for new information that may reveal additional insight into human practices with regards to individual health decision making and social behaviors. More importantly, its recursive nature affords researchers, specifically Web Scientists, the opportunity to draw from these human insights in order to build a better, more responsive Web. To achieve this goal, the Web Science community, particularly Health Web Scientists, must strive to first, provide a distributed open data collection and analytics environment to facilitate cross-collaboration among researchers (aka: the Web Observatory). Second, leverage the Web Observatory platform to study socio-technical aspects of the Web. And, third, consider how these systems are themselves artifacts to be reviewed and or observed. This type of meta-analysis will ensure that the technical infrastructure and governance policy of the WO reflects lessons gleaned from the socio-technical observations. Throughout this paper, we have advocated for an additional layer of legal and ethical considerations that should be incorporated within the Web Observatory framework. In particular, we argue that given the sensitivity and highly-regulated nature of health information, Health Web Scientists are primed to take the lead in this space. We recognize that this is exploratory and preliminary work that will require further discussion and revision. For example, while there are technical and legal frameworks that may help guide the development of a HWO, privacy in health requires an understanding of multiple levels of complexity that include

[5] For more information the TWC Mobile Health project, please see <http://tw.rpi.edu/web/project/MobileHealth>

individual, social, and legal factors. One person's privacy practice may certainly not be the same as another's. How a WO may address these individualities while providing enough structure to enable sharing and collaboration will be an iterative process. Moreover, while the above recommendations can be used to help with the initial build, we urge that it eventually become an essential evaluation criteria for all future Web Observatories.

4.ACKNOWLEDGEMENTS

Our thanks to the Tetherless World Constellation lab for their support and help in the development of tools featured in the TWC Web Observatory Portal. Additional thanks to the entire Web Science Community for their recommendations and feedback on the schema.org extension vocabulary.

5.REFERENCES

[1] Appari, A. and M. Eric Johnson. Information security and privacy in healthcare: current state of research. *International journal of Internet and enterprise management* 6.4 (2010): 279-314.

[2] Bansal, G., Zaheid, F.M. and Gefen, D. The impact of personal dispositions on privacy and trust in disclosing health information online. *Americas Conference on Information Systems*. (2007). Keystone, CO, http://aisel.aisnet.org/amcis2007/57

[3] Bansal, G., Fatemeh Z, and David G. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems* 49.2 (2010): 138-150.

[4] Benedict, J., McGuinness, D.L, and Fox, P. A Semantic Web-based Methodology for Building Conceptual Models of Scientific Information. *In American Geophysical Union, Fall Meeting (AGU2006)*, San Francisco, Ca., December, 2007. Eos Trans. AGU 88(52), Fall Meet. Suppl., Abstract IN53A-0950

[5] Bradley, D. Crowdsourcing Drug Discovery: Antitumor compound identified. *SpectroscopyNow.* Website. (1 Jan. 2014). http://bit.ly/1cXE7rc

[6] Center for Democracy and Technology. FAQ HIPAA and Cloud Computing (v1.0). Website. (7 Aug. 2013). http://bit.ly/1ePEoMA

[7] Center for Democracy and Technology. Health Privacy. Website. http://bit.ly/1dkWO9q

[8] Chou, Wen-Ying Sylvia, et al. Social media use in the United States: implications for health communication. *Journal of medical Internet research.*11.4 (2009).

[9] Fox, S. and Duggan, M. Health Online 2013. Pew Internet Research. Website. (15 Jan 2013). http://bit.ly/1ePtPcx

[10] Fox, S. and Duggan, M. Tracking for Health. Pew Internet Research. Website. (28 Jan. 2013). http://bit.ly/19qKDIB

[11] Fox, S. The Social Life of Health Information. Pew Internet Research. Website. (12 May 2011). http://bit.ly/1hkBotR

[12] Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. U.S. Department of Heath and Human Services. Website. http://1.usa.gov/1azIDMV

[13] Hasan, R. and William Y. Beyond Media Hype: Empirical Analysis of Disclosed Privacy Breaches 2005-2006 and a DataSet/Database Foundation for Future Work. *The Workshop on the Economics of Securing the Information Infrastructure.* 2006.

[14] Khandelwal, A., et al. Analyzing the AIR language: a semantic web (production) rule language. *Web Reasoning and Rule Systems*. Springer Berlin Heidelberg, 2010. 58-72.

[15] Mobile Health. Tetherless World Constellation. Website. http://bit.ly/1kyQzEp

[16] Luciano, J., Cumming, G., Wilkinson, M., and Kahana, E. The Emergent Discipline of Health Web Science. *Journal of Medical Internet Research.* 15(8):e166 (2013). DOI: 10.2196/jmir.2499

[17] McGraw, D. Privacy Protections Must Accompany New Models of Health Care. Center for Democracy and Technology. Website. (4 Nov. 2013). http://bit.ly/1drcZDJ

[18] Oberoi, M., Jagtap, P., Joshi, A., Finin, T., and Kagal, L. Information Integration and Analysis: A Semantic Approach to Privacy. Proc. *Third IEEE International Conference on Information Privacy, Security, Risk and Trust.* (9 Oct. 2011).

[19] Samuelson, P. Privacy as intellectual property. *Stan. L. Rev.* 52 (1999): 1125.

[20] Swan, M. Sensor mania! The Internet of Things, wearable computing, objective metrics, and the Quantified Self 2.0. *Journal of Sensor and Actuator Networks* 1.3 (2012): 217-253.

[21] Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N., and Hendler, J. The Web Science Observatory. *IEEE Intelligent Systems*, vol. 28, no. 2. (March-April 2013) 100-104. DOI:10.1109/MIS.2013.5 SOTON: http://bit.ly/1crWAsP.

[22] Weitzner, D., Abelson, H., Berners-Lee, T., Hanson, C., Hendler, J., Kagal, L., McGuinness D.L., Sussman G., and Waterman, K. Transparent Accountable Data Mining: New Strategies for Privacy Protection. *Proceedings of AAAI Spring Symposium on The Semantic Web meets eGovernment.* AAAI Press, Stanford University, Stanford, CA., USA, March, 2006. Also available as MIT CSAIL Technical Report-2006-007 and Stanford KSL Technical Report KSL-06-03.