Second-order interoperability in the datafication of public health

MARTIN STOJANOV

This paper investigates the practical work of repurposing web search data for public health surveillance and highlights challenges related to interoperability.

Keywords: datafication, interoperability, public health, second-order friction

INTRODUCTION

With internet use becoming an integral part of many people's everyday lives, using the internet to obtain health information and engage in self-care is by now an established practice (Nettleton, Burrows & O'Malley 2005). Searching for health information online produces a data trace that consists of search terms and information on their distribution over time, which can be stored in the user's web browser as well as in the search engine's databases. Health-information seeking online and other online activities produce data traces of our actions, in a process that can be referred to as datafication (van Dijck 2014). This continually produces updating data sets, which can then be repurposed in different contexts (Newell & Marabelli 2015). Within the realm of public health, datafication is being adopted in syndromic surveillance – the practice of monitoring population health in order to formulate appropriate public health responses. Infodemiology exemplifies how web search traces, which often consist of search terms related to symptoms, are being used to identify a potential outbreak that could require a public health intervention (Eysenbach 2006).

Datafication, and with it the repurposing of data, is an example of a second-order system, meaning a system that relies on a network of (first-order) systems built for purposes other than those of the second-order system (Boyce 2016; Van Der Vleuten 2003). Second-order systems borrow from and are dependent on existing infrastructure, but they still maintain their own databases (Boyce 2016). Boyce (2016) has examined the public health infrastructure for surveilling foodborne disease outbreaks in terms of second-order systems, as it relies on data and materials repurposed from the food and health sector. She articulates some of the challenges they face through the notion of second-order friction, which brings to the fore the efforts involved in overcoming obstacles "when actors in the second-order system repurpose materials and data from other systems and infrastructures (Boyce 2016, 56)." One such challenge when repurposing diagnostic tests performed by different laboratories pertains to variations in testing approaches between the laboratories as well as variations over time, which can threaten interoperability (Boyce 2016).

DOI: 10.5281/zenodo.3297139

Interoperability is understood broadly as making "heterogeneous data work with each other (Ribes 2017 1515) " and entails the work of making data comparable according to a common metric or commensurable; it is an important factor for enabling reuse of data for disease surveillance in public health (Dixon, Vreeman, & Grannis 2014). Differences in what measures are used to diagnose patients could have an impact on whether repurposed diagnostic test data are comparable.

Interoperability in the context of the datafication of public health remains underexplored. The present paper adds to this discussion through a case study of the repurposing of data traces from health information seeking online for infodemiology. It examines the dependency of a second-order system in the context of the datafication of public health and looks at the implications for interoperability. It highlights second-order frictions with respect to the maintenance and development of the systems that the second-order system relies on in order to shed light on interoperability challenges for datafication in public health.

RESEARCH APPROACH AND RESEARCH SETTING

This paper comes from a larger study examining datafication in public health by following the development and use of an infodemiological system (hereinafter InfoSurv), which relies on data traces produced when individuals seek health information online. An automated analysis in InfoSurv of the proportion of searches that are related to a set of symptoms associated with seasonal illnesses allows epidemiologists to follow the incidence of illness in the population. When the relative number of search terms for symptoms related to influenza rise, epidemiologists consider it to be a signal that the incidence of influenza in the population has increased. InfoSurv is maintained by a Nordic public health organisation. It uses search data from a prominent health information portal (hereinafter HIP) to estimate illness in populations. The HIP website has a search bar for locating content, but many HIP visitors arrive at their desired webpage via a prominent search engine (hereinafter SE).

The study's methodology consisted of a participant observation of three developers who worked with maintaining and developing InfoSurv and of weekly meetings between five public health professionals whom InfoSurv factors into decision making about current influenza incidence. Informal interviews and document studies also supplemented the participant observation. While the larger study was initiated in September 2017, this paper focuses on a series of events surrounding changes to the HIP and how it affects InfoSurv; these unfolded between January and April 2019. As InfoSurv relies on stable search behaviour via the HIP search bar to find patterns indicating changes in the incidence of illness, a change in the HIP website risks creating a discontinuity with historical data and compromising its use. As of April 2019, the full implications of the changes to HIP were still unfolding.

EXPLORING ALTERNATIVE DATA SOURCES TO THE FIRST-ORDER SYSTEM

The extent of the planned changes to the HIP came to the attention of the developers in January 2019 and raised concerns that InfoSurv would no longer be able to operate. The HIP was to receive a new search engine, with a different way of providing search suggestions to users; this development could alter search patterns among future HIP visitors. Furthermore, the HIP website was to undergo design changes that could further alter the search patterns. Hence, the developers explored ways of mitigating the impact of the first-order changes on the second-order system.

The developers explored the possibility of using search data as an alternative data source for InfoSurv, which was originally developed using sentinel data on influenza-like illnesses and HIP data from the corresponding time period. The developers investigated whether the two datasets could be combined:

> **D1:** One of the problems is that we don't have enough historical data.
> **D2:** But perhaps we could adjust the historical data. But it's troubling that there is such a big difference in the search pattern for pregnancy.
> **D3:** Also fever and influenza.
> **D1:** So you're thinking that we could transform it?
> **D2:** It would have been easier if it wasn't so inconsistent. It's difficult to be confident with this difference.
>
> *(Adapted from fieldnotes)*

The difference between the HIP data and the SE data was too substantial and the inconsistencies too great for them to be reliably combined. Hence, the developers were not confident that the data sources were similar enough for the SE data to provide a solution.

A further limitation of the SE dataset pertained to the timing of its availability (see Figure 1 and Figure 2). Data for the full preceding week only became retrievable when there was a three-day delay instead of one, as had been the case with HIP data. If implemented with SE data, this delay would have affected existing influenza surveillance routines, which produced an analysis on Mondays based on the previous week and made an extrapolation for the current week on Wednesdays based on data from Monday and Tuesday of the same week. These analyses were discussed on Wednesdays in conjunction with other surveillance systems to assess influenza incidence. Furthermore, SE data was logged in Pacific Standard Time, which shifted the week in time due to the time differences. Basing InfoSurv on SE data would mean InfoSurv would no longer align with the definition of a week used in other surveillance systems.

The second-order system was developed from the data production of the first-order system. While the data from the SE could be acquired and the search terms mostly coincided with the HIP data, they could not be made to work together. To be a viable solution, the SE dataset needed to be made commensurable with the dataset from the HIP search bar. However, the model underpinning the surveillance system had been developed around a data stream constituted in the interaction between users and a particular HIP design. It also relied on the means of navigating that website with a particular search bar, such that one data source could not easily be aligned with another seemingly similar data source. This echoes the claims by Ribes (2017, 1516) that data interoperability is an arduously and historically inscribed accomplishment "still carrying the consequences of that interoperation to future uses."

Figure 1. Work practices based on HIP data

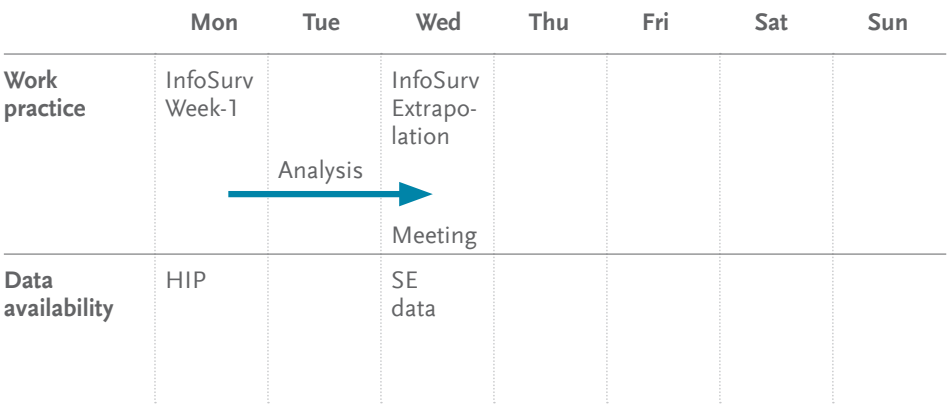
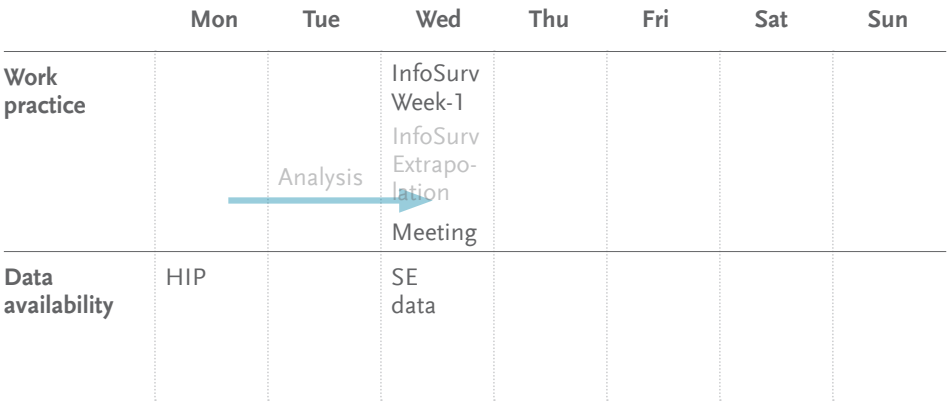


Figure 2. Work practices based on SE data



The findings also highlight the significance of spatio-temporality for the interoperability of data with respect to work routines and other surveillance systems. As work practices structure time (Orlikowski & Yates 2002), and in this case are co-constituted together with a range of other surveillance systems with an existing definition of "a week", it becomes difficult to integrate a new data source if it does not align with the existing temporal structures. Both of these vignettes speak to the sensitivity of datafication to consistency in the way data is generated. In this regard, InfoSurv is similar to lab-based surveillance systems. Boyce (2016) shows how changes in lab equipment can threaten interoperability in disease outbreak surveillance. The examples suggest that a new data source can be challenging with respect to interoperability and that there could be a tension between the continuation of InfoSurv and innovation in the systems from which data is repurposed.

MAINTAINING A CONNECTION IN THE FACE OF FIRST-ORDER CHANGE

In addition to exploring alternative solutions, the developers worked to adapt the second-order system to ensure that InfoSurv would be able to continue receiving HIP data after the HIP changes. Partly, this process meant revising the way in which the second-order system managed first-order maintenance routines. The search data that InfoSurv receives from HIP needs to be filtered to remove test searches done as part of HIP maintenance. Depending on when the HIP data was generated, InfoSurv parses it differently to account for the HIP maintenance routines that were in place at the time when the data was generated. The consultants working with the changes to HIP were forthcoming on how to identify maintenance routines in the upcoming version of HIP, but this had not always been the case. For certain parts of the historical data, the developer had to infer which search terms stemmed from maintenance testing based on patterns in the data, as the organisation responsible for maintaining the HIP at the time was unable to provide reliable information about who could be doing testing and when.

The new version HIP has also changed the way in which the HIP produces and delivers data about search behaviour: the new search engine distributes the computing workload between several servers, dynamically adapting depending on workload and creating one search log file for each active server instead of one search log file. This has created uncertainty about how to distinguish an unusually small number of log files, a problem that arises due to an error related to unusually low HIP use. During a meeting with the consultants, the developer raised the question of how to reliably determine that InfoSurv had acquired all the data for the previous day:

> **D1:** I'd like us to define how we ascertain the number of files so that we know if there is something missing. It's hard to look for something that does not exist.
> **Consultant:** Our focus is on delivering what we need to. We have quite a lot to do. We can deliver the files and if it turns out to be a problem we'll have to re-evaluate. I need to give you an answer that is a bit dissatisfying.
>
> *(Adapted from fieldnotes)*

At the time, the first-order system and second-order system were connected without a guarantee that all relevant data had been transferred. While there were agreements regulating the transfer of data from the first-order system to the second-order system, the means of enforcement were such that it was done on a voluntary basis.

The findings suggest that the first-order maintenance practices can influence the interoperability of data in the second-order system. Since the system maintenance routines of the first-order system can influence the production of the search data that are repurposed in the second-order system, changes in these routines can impact what data patterns emerge. Hence, datafication is continuously being repaired in order to maintain interoperation according to what first-order system maintenance routines are in place. The focus on repair work articulates how data interoperation in datafication is a continuous accomplishment (Jarzabkowski & Pinch 2013), one that is repeatedly and constitutively linked to second-order maintenance practices that (re-)align algorithms and data to compensate for first-order changes.

Much like Boyce (2016), the study finds that the impact of first-order changes on interoperability can vary. While the maintenance routines could be compensated for, the developers had more difficulty in managing the dynamic load balancing between different servers. This did not remain unresolved due to an inherent technical difficulty but rather due to differing priorities in the face of time pressure. Interoperability in the second-order system was prioritised only in so far as it did not interfere with first-order priorities. This suggests that, due to the asymmetric dependence of second-order systems (Boyce 2016), a careful examination is needed of how responsibilities and accountabilities for data transfer are arranged in order to ensure interoperability in the datafication of public health and a sensitivity to what agreements are put in place. Furthermore, both the example of the maintenance routines and the dynamic server loading highlight the significance of working relations in making data interoperate across the first- and second-order systems.

CONCLUSION

This paper has sought to unpack some of the second-order frictions in making datafication in public health work based on repurposed web search data. In particular, it highlights challenges of making data commensurable. The case suggests that interoperability in datafication is a continuous accomplishment that entails compensating for first-order modifications to the system that produces the repurposed web search data. Furthermore, changes in the system generating the data that is being repurposed can threaten data interoperation in ways that can be difficult to mitigate. It highlights that first-order maintenance routines, the spatio-temporality of data and its compatibility with work practices as well as the alignment between first-order and second-order priorities can have an impact on interoperability.

BIBLIOGRAPHY

Boyce, A. M. (2016). Outbreaks and the management of 'second-order friction': Repurposing materials and data from the health care and food systems for public health surveillance. *Science & Technology Studies*. Retrieved from https://sciencetechnologystudies.journal.fi/article/view/55409

Dixon, B. E., Vreeman, D. J., & Grannis, S. J. (2014). The long road to semantic interoperability in support of public health: Experiences from two states. *Journal of Biomedical Informatics, 49,* 3–8. https://doi.org/10.1016/j.jbi.2014.03.011

Eysenbach, G. (2006). Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. *AMIA Annual Symposium Proceedings, 2006,* 244–248.

Jarzabkowski, P., & Pinch, T. (2013). Sociomateriality is 'the New Black': accomplishing repurposing, reinscripting and repairing in context. *M@n@gement, 16*(5), 579–592. https://doi.org/10.3917/mana.165.0579

Nettleton, S., Burrows, R., & O'Malley, L. (2005). The mundane realities of the everyday lay use of the internet for health, and their consequences for media convergence. *Sociology of Health & Illness, 27*(7), 972–992. https://doi.org/10.1111/j.1467-9566.2005.00466.x

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification.' *The Journal of Strategic Information Systems, 24*(1), 3–14. https://doi.org/10.1016/j.jsis.2015.02.001

Orlikowski, W. J., & Yates, J. (2002). It's about Time: Temporal Structuring in Organizations. *Organization Science, 13*(6), 684–700.

Ribes, D. (2017). Notes on the Concept of Data Interoperability: Cases from an Ecology of AIDS Research Infrastructures. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17,* 1514–1526. https://doi.org/10.1145/2998181.2998344

Van Der Vleuten, E. (2003). In Search of the Networked Nation: Transforming Technology, Society and Nature in the Netherlands during the Twentieth Century 1. *European Review of History: Revue Européenne d'histoire, 10*(1), 59–78. https://doi.org/10.1080/13507480303665

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society; Newcastle upon Tyne, 12*(2), 197–208.