

Provenance, power and place: Linked data and opaque digital geographies

Heather Ford
University of Leeds, UK

Mark Graham
University of Oxford, UK

Environment and Planning D: Society and
Space

2016, Vol. 34(6) 957–970

© The Author(s) 2016



Reprints and permissions:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/0263775816668857

epd.sagepub.com



Abstract

The ability of search engines to shape our understandings of the world by controlling what people discover when looking for information is well known. We argue that the power of search engines has become further entrenched in the wake of the current move to restructure the Web according to the logics of 'linked data' and the 'semantic Web'. With the goal of sharing information according to structured formats that computers (rather than humans) can easily process and analyse, linked data engineers are abstracting information from fact sharing websites like Wikipedia into short, uniform statements that can be more efficiently shared, compared and analysed. In response to this enhanced power by search engines and the corresponding loss of agency by ordinary users, some Wikipedians have challenged the ways in which data from the encyclopedia has been used (often without credit) by search engines like Google. Using the capabilities approach first developed by Amartya Sen, we interrogate exactly what some Wikipedians believe they are losing when they complain about how Google represents facts about the world obtained from Wikipedia and other sites.

Keywords

Semantic web, linked data, capabilities theory, geoweb

Digital places

Places are not only material but are also informational. Place is made up of memories, stories, information and histories. What is Johannesburg? It is a city of trees and buildings, concrete and sand. It is also constituted by a myriad statements made by multiple actors, some of which are represented by information in books, in census reports, in tourism leaflets and photographs. Today, much of that information is digital and available on the Internet. Spatial information is either digitized from analogue sources or in increasingly 'born digital' (created as digital data rather than scanned or translated into

Corresponding author:

Heather Ford, University of Leeds, Clothworkers North, Leeds LS2 9JT, UK.

Email: hfordsa@gmail.com

digital formats) and can take a range of forms such as geotagged images on Instagram, hashtags on Twitter, annotations on Google Maps and Wikipedia articles, in addition to official data from government and corporate sources.

In addition to the enhanced ability of ordinary people to contribute to the digital representations of cities (Goodchild, 2007; Graham, 2013; Haklay et al., 2008), we have also seen a growing centralisation in the control of platforms that mediate everyday life. Silicon Valley-based Google, Facebook, Twitter and Wikipedia have become the most used websites and digital platforms in most countries, and some scholars (Introna and Nissenbaum, 2000; König, 2014; Morozov, 2013) warn of the dangers of increasing centralisation and commercialisation of the guiding forces of the Internet.

The power yielded by search engines, in particular, has come under increased scrutiny by researchers in recent years. Introna and Nissenbaum (2000: 1), for example, have shown how search engines 'systematically exclude... certain sites, and certain types of sites, in favor of others'. Eli Pariser (2012) argues that search engines drive the construction of 'filter bubbles' that only show users information that they agree with. Our increasing reliance on search engines like Google constitutes what Siva Vaidhyanathan (2012) refers to as an 'outsourcing' of judgement to Google, particularly because search engines have become critical to the public health of the Internet (König, 2014). As place becomes increasingly digital and the digital becomes increasingly spatialized, Graham and Zook (2013) have shown that informational filter bubbles can manifest into material divisions and barriers.

The goal of this paper is to highlight a new problem. As digital data becomes increasingly abstracted into short data statements that can be shared and interconnected according to logics of 'the semantic web' or 'linked data', the concentration of power in the hands of search engines has been enhanced still further. We argue that the increased control of search engines over human knowledge has been garnered due to the loss of provenance, or source information, in data sharing algorithms. When the links between information and their sources are severed, users' capabilities to actively interrogate facts about the world are significantly diminished. Paul Groth (2013) has noted that the loss of provenance information in semantic web projects is a significant challenge (c.f. Groth, 2013), but we explore the socio-spatial implications of this technological change by focusing on what the loss of provenance information means for how people experience and represent place.

We highlight the origins and consequences of the loss of provenance information in the context of the contemporary moment in which the web is being significant re-engineered. What first appears to be merely a simple engineering problem turns out to be indicative of the growing commercialisation of the web, a problem that stems from the dominance of an epistemology that sees knowledge about the world as essentially reducible to depoliticised data that is natural and obvious, rather than what it actually is: a re-constructed representation that obscures the origins of information and, in so doing, reduces the ability of ordinary users to interrogate that data. Despite the promise of the move towards a more semantic web (Egenhofer, 2002) for more precise digital representations of place, there has been a parallel decrease in the capabilities of people to interrogate and control that data.

The semantic web and linked data revolution

The move to a web of linked data was catalysed when in 2001, Tim Berners-Lee (the inventor of the web), James Hendler and Ora Lassila first published their new vision for 'The Semantic Web' in an article for the *Scientific American*. The idea for the semantic web was that, instead of information on the web produced for human consumption, it was

being restructured so that machines could more efficiently deliver personalised results and services. The web was moving away from the ‘web of documents’ and the storing of data in flat, static formats, to the abstraction of information into short, modular statements that could be linked together and mined by algorithmic processes.

Berners-Lee et al. (2001) applauded and urged further development of semantic web principles, declaring that the semantic web was not just a new tool for conducting tasks but rather that, ‘the Semantic Web (could) assist the evolution of human knowledge as a whole’.

This structure will open up the knowledge and workings of humankind to meaningful analysis by software agents, providing a new class of tools by which we can live, work and learn together. (Berners-Lee et al.2001)

The authors noted that the ‘benefits (were) hard or impossible to predict in advance’, but the goal was that as engineers restructured websites and shared their ontologies (maps of relationships between different entities in their databases) this would better enable ‘computers and people to work in cooperation’.

The goal of the semantic web initiative, in other words, was to focus more keenly on formatting information in ways that computers – as opposed to humans – could more easily process. According to the authors, the semantic web would usher in a new era in which machines are able to ‘process and “understand” the data’ than to merely display that data.

Most of the Web’s content today is designed for humans to read, not for computer programs to manipulate meaningfully. Computers can adeptly parse Web pages for layout and routine processing—here a header, there a link to another page—but in general, computers have no reliable way to process the semantics. (Berners-Lee et al. 2001)

Although the particular application of linked data or semantic principles has differed in some respects, the foundation of all semantic web or linked data projects is founded in a particular, simple algorithmic representation of information: the key-value pair (or ‘triple’ as Berners-Lee et al. (2001) call it). Key-value pairs are a foundational element of computing systems and used in designing a variety of applications, from mapping applications to database systems and library metadata. Key-value pairs divide a statement into a subject and object, making assertions about a thing (a person, place, or any other subject) which has particular properties (‘is author of’, ‘belongs to’, etc.) with particular values (another person, place or thing). Each element of the statement is identified by a unique Universal Resource Identifier (URI) which then ‘enables anyone to define a new concept, a new verb, just by defining a URI for it somewhere on the Web’. (Berners-Lee et al., 2001)

Before data are structured in this way, information about a city might be contained within a 2000-word document about the city, with a variety of headings about its demographics, governance, culture and geography. Structuring information about the city means that the entire document is divided up into a series of hierarchically organised, short statements that are entirely made up of key-value pairs. According to this structure, two different objects are associated with one another in a data structure that can be extended without changing any of the underlying objects. This means that there can be a database comprising numerous cities whose attributes can be iteratively added to (when the city gets a new park, for example), or whose values can be edited (when the population figures increase) without having to change the entire entry.

Objects in key-value pairs are represented in one column by a person, a place, an event, etc. and a value in another column (a measurement, an amount, a description, quality or comparison). The objects are related or linked to one another using a qualifying label.

Johannesburg	
Instance of	city
Country	South Africa
Coordinates	26°8'42"S, 28°3'1"E
Inception	1886

Figure 1. A representation of Johannesburg as a series of key-value pairs.

The objects 'Johannesburg' and 'South Africa' are meaningless without the connecting label 'city in' that determines Johannesburg to be a city in South Africa (see Figure 1).

The two objects could, in turn, be connected using other labels such as 'largest city in', so that a limited number of objects can create a myriad of facts in different combinations. The relationships between different objects in the database are designed in what is called a data model which specifies what kind of data can be supported by a system and the types of relationships between different values that can be represented. The data model may require that distances can only be represented as miles and not kilometres or it could establish a rule (and an accompanying algorithm) that converts all mile values to kilometres automatically. The data model could specify only a limited set of sources for determining population figures (national government statistics agencies rather than corporate mapping companies, for example), or it could specify what national languages can be attributed to particular countries through a defined list.

In their vision for the semantic web, Berners-Lee et al. (2001) noted that the semantic web would be truly powerful when people created programmes that collected content from diverse sources and noted that agents could share their ontologies in directories 'analogous to the Yellow Pages'. This process of joining up different languages to reach a 'wider common language' was essential and could be achieved through sharing these dictionary-type structures 'even when the commonality of concept has not (yet) led to a commonality of terms'. (Berners-Lee et al., 2001)

Although the sources of data extracted and applied in data models have largely been obscured in current instantiations as we will show below, data provenance was actually a feature of Berners-Lee et al.'s (2001) semantic web vision statement. The authors declared that automated agents would collect data in the form of key-value pairs (or triples) from diverse sources on the web and present them to the user, along with evidence of the sources from which the agents derived their information.

An important facet of agents' functioning will be the exchange of "proofs" written in the Semantic Web's unifying language (the language that expresses logical inferences made using rules and information such as those specified by ontologies). For example, suppose [someone's] contact information has been located by an online service, and to your great surprise it places [them] in Johannesburg. Naturally, you want to check this, so your computer asks the service for a proof of its answer, which it promptly provides by translating its internal reasoning into the Semantic Web's unifying language. (Berners-Lee et al., 2001)

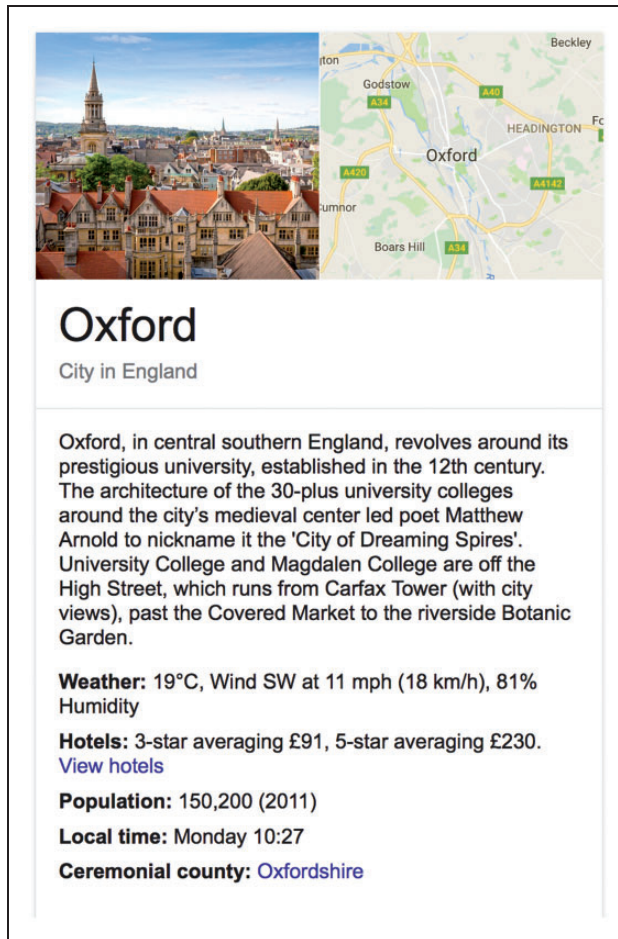


Figure 2. Google knowledge graph results for a search for 'Oxford' (22 August 2016).

Another key feature of the semantic web according to the authors would be digital signatures that would be used to verify that the attached information was being provided by a trusted source.

Agents should be skeptical of assertions that they read on the Semantic Web until they have checked the sources of information. (Berners-Lee et al., 2001)

Fifteen years later, one of the major semantic web initiatives has emerged in the work of search engines to extract semantic data from multiple sites (predominantly Wikipedia) and display a selection of that data to the user in the form of an 'infobox' containing key facts about a particular person, place or thing. However, many Internet users (particularly Wikipedia editors) have found it problematic that provenance data are missing in these large data extraction and structuring projects. In particular, some Wikipedia editors (who refer to themselves as 'Wikipedians') are concerned that Wikipedia data are being extracted by Google and presented in the form of a prominent infobox in search results (see Figure 2) as part of Google's 'Knowledge Graph' initiative (Singhal, 2012) but that the source of the

data is not always visible (NMaia, 2016; Kolbe, 2015). Furthermore, Google refuses to answer questions about how its results are garnered (Ford and Graham, 2016) and many believe that the introduction of infoboxes has led to a decline of visits to Wikipedia (Kloc, 2014; Kolbe, 2015, 2016; Orlikowski, 2014).

Wikipedia's own semantic web initiative, Wikidata, has also not escaped controversy. Established in 2012, Wikidata's goals were twofold: to support Wikipedia and other Wikimedia projects by enhancing consistency across different projects and language versions, and to support the many different (third party) services and applications that reuse Wikipedia data in a structured way (Vrandečić and Krotzsch, 2014). Information about Paris, for example, is distributed across Wikipedia articles in many of the 250+ language versions of Wikipedia, images labelled 'Paris' on Wikimedia Commons (commons.wikimedia.org) and quotes labelled 'Paris' in Wikiquote (wikiquote.org). Wikidata now stores links to all of the data about Paris from across Wikimedia projects (in addition to other sources of data from across the web) so that it becomes the central site for those wanting to reuse data.

Wikidata has been criticised by Wikipedians because the majority of its statements remain unsourced, because discussion of Wikidata entries can only take place in English rather than any of the 250+ language versions of Wikipedia, and because participation in Wikidata requires technical expertise that many Wikipedians do not possess. In June 2016, the Wikipedia Signpost (a newsletter focused on Wikipedia) published an op-ed calling for a change to the licensing of Wikidata – from a license that does not require attribution to one that does (NMaia, 2016). The authors wrote that corporations like Google were profiting from the labour of Wikipedia because they were not required to attribute the source on Wikipedia.

The capability approach

It has been difficult for opponents of Wikidata and Google's semantic web activities to articulate exactly what the problem with this loss of provenance is and why it is so important to integrate provenance data when websites share information. Isn't the point of Wikipedia that its information is shared as widely as possible? Do Wikidata's efforts not help, rather than hinder, such goals?

The difficulty in articulating the problem with semantic web 'remixing' is a result of the ways in which the provenance problem has been defined as a loss of information instead of what the problem represents more foundationally: a change in what we as users and as digital citizens are able to do and be. What does the unsourced extraction of data from collaboratively constructed information sources like Wikipedia mean for what we are able to do and to be? How does this shift in the context and containers in which information exists change our lived experience when it comes to everyday informational practices that have become so central to everyday digitally mediated life?

The capabilities approach developed by Indian economist and philosopher, Amartya Sen (2001), offers an important lens for answering these questions. According to Sen, the foundation for evaluating human development programs is the extent to which they enable people to actually do certain things should they choose to do so. The focus, for Sen, is on the ends rather than the means – what people are actually capable of doing rather than the predetermined functionings of a particular program (its affordances) – because this is what ultimately matters for human development.

The capabilities approach is particularly relevant in terms of users' participation in the representation of place. Much as people have long struggled for control over the physical

spaces of cities (including rights to public land, rights to public assembly, etc.), we might interpret a capability approach as a goal for people to be able shape the digital infrastructure of their cities. This is an important capability because it has an influence on the rights to control one's environment and to participate effectively in political choices (Nussbaum, 2011). Information about a city, country, street, monument, park, or neighbourhood affects how others view that place. For example, the ways in which Jerusalem is represented as either the capital city of Israel or Palestine (or both) have an impact on claims to international support during the ongoing conflict.

This loss of capability is reflected when tracing information about a city as it moves from Wikipedia to Google. Analysing how cities are represented in the many languages of Wikipedia and tracing a loss of provenance as data is extracted and positioned within Google and Wikidata, we notice the removal of key capabilities. On Wikipedia, readers and editors are able to individually and communally evaluate the accuracy of statements by interrogating the sources from which citation information was derived. This can be achieved individually by the user looking up the source in the citations provided and evaluating its accuracy according to their personal heuristics.¹

Evaluation can also be achieved socially on Wikipedia by the user engaging in a dialogue with other users about how the statement might be improved by adding a 'citation needed' tag to indicate that the statement requires evidence, by editing the statement directly to add or remove sources, or by discussing changes with other editors on the talk pages. Obviously not all of these actions are always available to all users because they depend on the ability of editors to apply and decipher the particular socio-technical language used by Wikipedia (Ford and Geiger, 2012). The range of possible actions, however, becomes significantly more limited when one compares them with what can be achieved by users when this same information is extracted from Wikipedia and presented without the available affordances on Google (and to a lesser extent on Wikidata).

Statements on Google and Wikidata, in contrast, are often unsourced. When they are sourced, the source information is so vague or general that it makes it difficult to determine where information was actually obtained from (e.g. searching for a city or place on Google results in the infobox that is mostly unsourced). In Figure 3, the population of Oxford, England is unsourced and when the user clicks on the 'population' link, they are shown an enlarged version of the population number followed by a link to the Wikipedia article of the same name. The English-language Wikipedia article represents a different (more recent) figure in its infobox, leading to confusion as to where Google's figure was obtained from. Because Google uses indexes of crawled data rather than accessing real-time data, it is most likely that the figure is an older figure obtained from earlier version of Wikipedia, but without this information, users are left in the dark about the actual provenance of the data being presented.

Statistics that track Wikidata's progress (see Figure 4) indicate that half of the statements in Wikidata lack any source reference, and only 30% indicate they come from Wikipedia (rather than a particular article within Wikipedia). The majority of Wikidata entries have been populated by the work of automated agents or 'bots' that have been written to extract data from Wikipedia entries. The lack of provenance information on Wikidata is thus the result of a combination of factors. First, there is confusion amongst editors regarding the legal implications of extracting data for linked databases on Wikidata. Facts are generally not copyrightable and therefore do not legally require attribution, and although Wikidata might have asserted a database right for its compilation, the original project leaders decided on a copyright license that is even less restrictive than Wikipedia's own license. This license (the Creative Commons Zero, or CC0 license) does not require that those who extract data

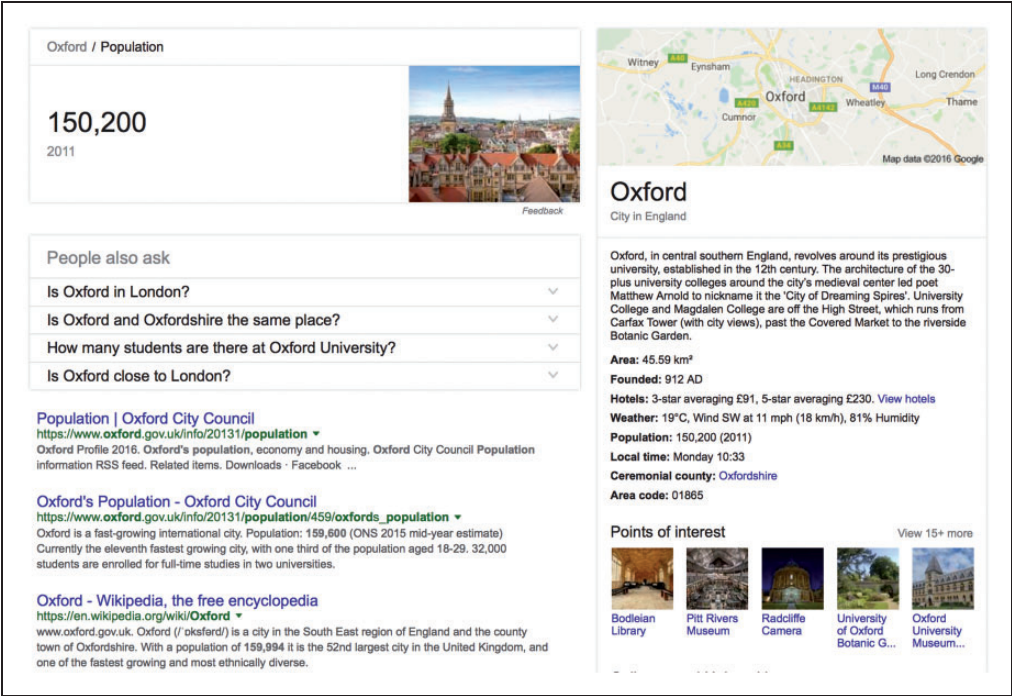


Figure 3. Screenshot from Google after the ‘population’ figure on the infobox is clicked (22 August 2016).

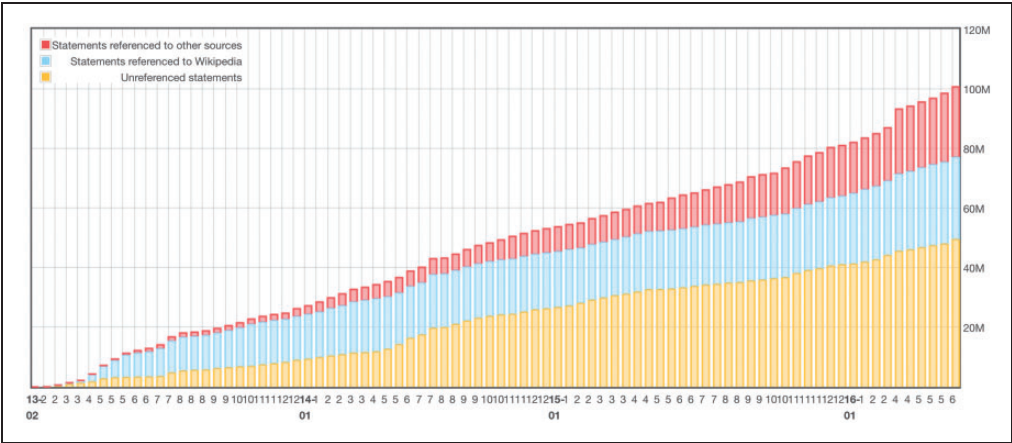


Figure 4. Statements in Wikidata that are referenced, referenced to Wikipedia and unreferenced By Wikimedia Foundation Labs – <https://tools.wmflabs.org/wikidata-todo/stats.php>? CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=45453804>.

from Wikidata attribute the source in any way, or release their own, enhanced versions of the data under similar terms, as Wikipedia's license does. In original discussions, the project lead for Wikidata, Denny Vrandečić wrote that there should be no wholesale extraction of Wikipedia data for Wikidata because this was against the terms of Wikipedia's own license, but (as Figure 4 indicates) there is still a significant proportion of statements in Wikidata that are extracted from Wikipedia.

The removal of provenance data from the facts represented in both Google and Wikidata's repositories thus leads to users losing their ability to effectively engage with the origins (and thus contexts and biases) of a statement. This loss of capability is compounded by the loss of accountability mechanisms on both Wikidata and Google. Figures 5 and 6, for example, show how a user is able to report that an error in the data, but the user receives no feedback on their complaint.

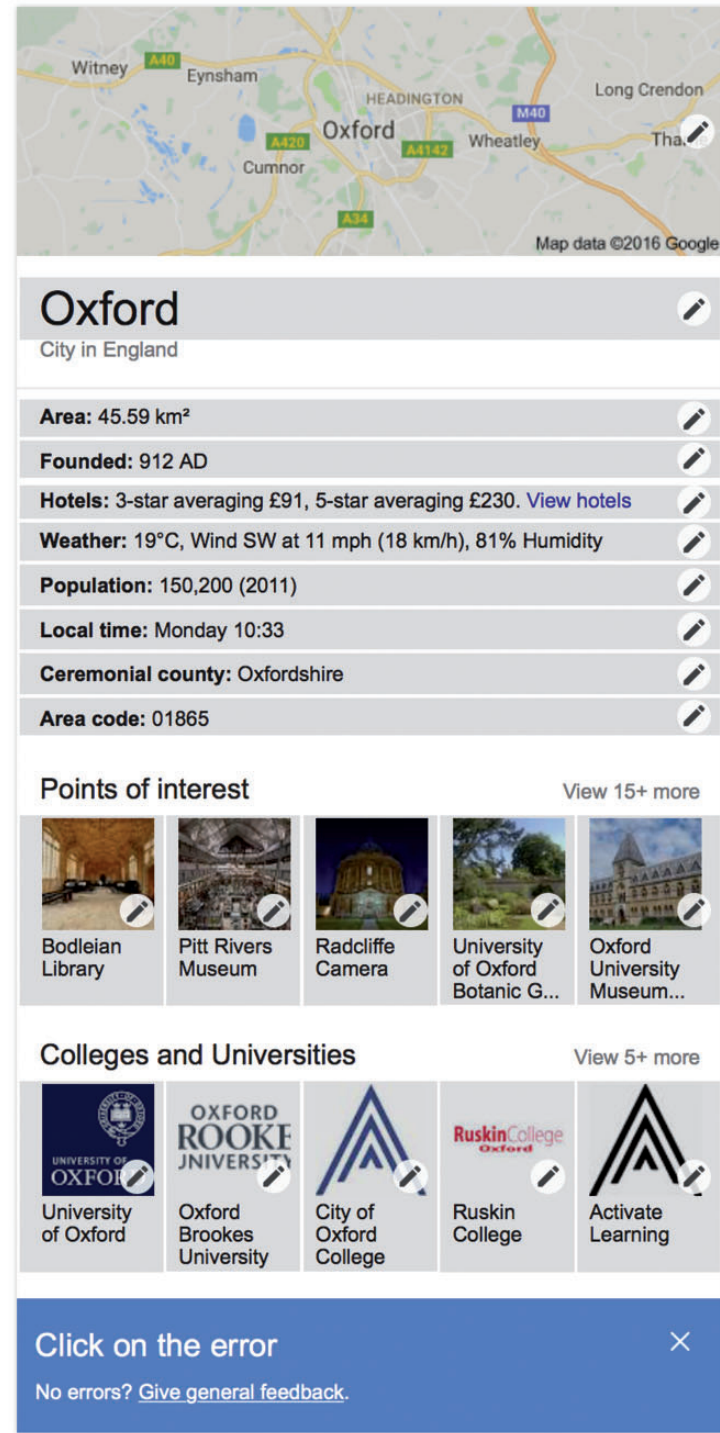
If users click on the 'feedback' link at the bottom of the infobox, they will have an option of detailing what is wrong with the results (see Figure 6). Google claims that 'input helps improve the Google Search experience' but a user's input 'won't directly influence the ranking of any single page'. A help page entitled 'Why we want your feedback' contains the following statement:

Search is constantly evolving. In a typical year, we experiment with tens of thousands of possible changes. Every change is tested in an experiment where some users see the change and others don't. By getting your feedback on our experiments, we learn which experiments are successful and should become part of Google Search for everyone. (<https://support.google.com/websearch/answer/3338405?hl=en>, as at 5 July 2016)

Users are not, however, able to tell Google whether the information that they are representing as fact, is actually accurate or not. Instead, users are informed that they are being experimented upon and that their feedback will have no real impact on how information is actually represented. Instead of being able to alert the company as to the accuracy or inaccuracy of their facts, users are only able to provide any feedback to 'improve' the search experience for others. No connection is made between an improved search experience and accurate information for the individual user.

Although Wikidata offers significantly greater opportunities for users to question the data being represented within it, such questioning is limited by the centralisation of data from all 250+ language versions of Wikipedia into a single page, where consensus is difficult to garner. Even though every language version of Wikipedia may choose to use different data from an item in Wikidata, the singularity of the representation has meant that conflict still regularly occurs between users from different language projects (for example, see Ford and Graham, 2016). Furthermore, because of the complexity of the data structures (in relation to Wikipedia), there are concerns by some Wikipedians that control of Wikipedia is moving away from average users and falling into the hands of those with technical prowess, which will only deepen problems of gender and geographic inequality (Graham, 2011; Eckert and Steiner, 2013; Collier and Bear, 2012; Reagle, 2013) on the platform.

In summary, the lack of provenance information and the inability of users to meaningfully question those who control and represent digital information have significant implications for the capabilities of people in relation to their spatial environments. Users lose the ability to effectively question statements that are reflected as singular and authoritative within Google and Wikidata's domain. Feedback has limited efficacy because there is no response to it from those with the power to control the representation. This change cuts people off from representing the places in which they live and constitutes a diminishing of their capabilities to be an active co-constructor of digital place.



Oxford
City in England

Area: 45.59 km²

Founded: 912 AD

Hotels: 3-star averaging £91, 5-star averaging £230. [View hotels](#)

Weather: 19°C, Wind SW at 11 mph (18 km/h), 81% Humidity

Population: 150,200 (2011)

Local time: Monday 10:33

Ceremonial county: Oxfordshire

Area code: 01865

Points of interest [View 15+ more](#)

- Bodleian Library
- Pitt Rivers Museum
- Radcliffe Camera
- University of Oxford Botanic G...
- Oxford University Museum...

Colleges and Universities [View 5+ more](#)

- University of Oxford
- Oxford Brookes University
- City of Oxford College
- Ruskin College
- Activate Learning

Click on the error [No errors? Give general feedback.](#)

Figure 5. Screenshots after clicking on 'feedback' and clicking the 'Wrong?' hyperlink above the word 'Oxford' in a Google.com results page (22 August 2016).

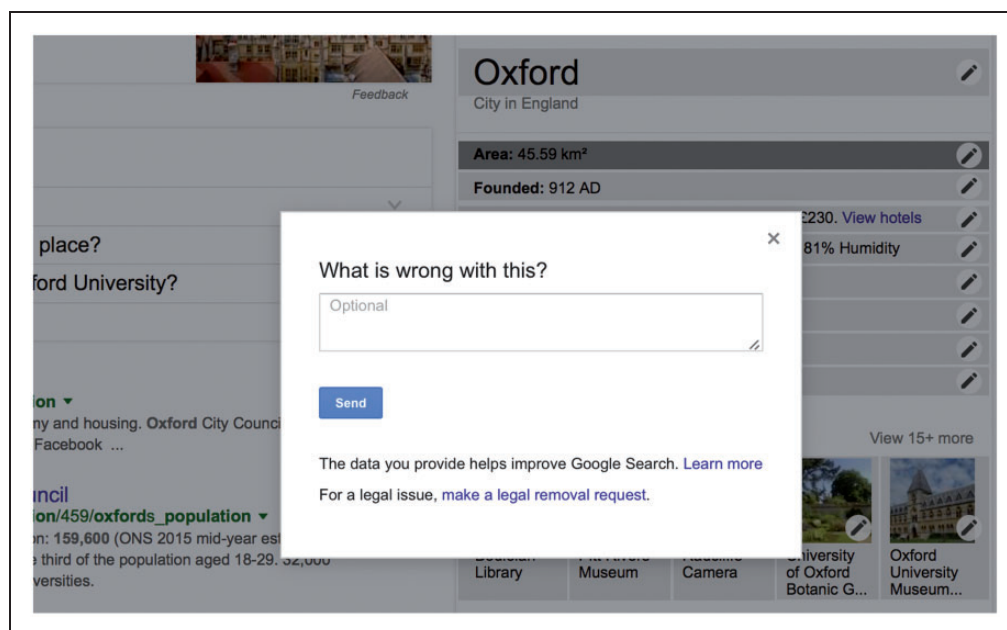


Figure 6. Screenshot after clicking the pen icon on the infobox for 'Oxford' (22 August 2016).

The implications of linked data for the representation of cities

The linked data revolution has resulted in new authorities of factual knowledge. Companies and projects like Google and Wikidata extract facts from Wikipedia using automated mechanisms and re-present them in new digital containers and new relational configurations. Some have rallied against this move because these organisations tend not to extract provenance information when they extract these facts.

Although there may be no legal barriers to reusing data created by users on other sites, we argue that there is an ethical argument to be made for reconnecting facts to the social contexts from which they are derived. Although spatial information might appear inherently geographically contextual, by virtue of it always being produced about a place, we have argued that a lack of provenance continues to strip the digital layers of place of important context. With the move towards a more semantic web, the increasing practice of extracting data about place and depositing it in decontextualized containers that pay little heed to the data's origins is difficult to reverse, but efforts are being made, at least on Wikidata, to try to improve the sourcing of facts.² For those working on such projects, provenance data should be made available so that users can at least trace the source of statements back to their origins. Doing so would afford the user the capability of investigating the sources of dominant facts and to question the authority of digital statements.

As data are structured and shared between different organisations and projects on the web, the digital layers of material places can become over-simplified over time. Data can lose connections to the contexts in which they were constructed, particularly through the loss of provenance information. In the case of cities, we see this in the way that there have been choices made about whose version of the status of a city like Jerusalem should

be represented as dominant and whose should be subordinate, and the biases inherent in any system where hierarchical choices need to be made about what must be shown to whom.

In this moment in which we are increasingly losing the power to control space and spatial representations, we find that Google's Knowledge Graph and a host of similar web initiatives represent both a continuation of the blackboxing of everyday urban life as well as a deepening of it. In sum, we argue that a change to the engineering of the web can have real-world implications on the cities that we live in. By allowing data to easily flow between different digital platforms, the move towards linked data and a more semantic web has resulted in a loss of provenance information in the digital layers of place. Consequently, our ability to see spatial information as always and already political is diminished. As our cities become increasingly digital, and the digital becomes ever more important in defining what a place is, it will be crucial to always be able to ask questions about who owns, controls, and can manipulate the informational layers of place. We therefore need to redouble our efforts to trace, track and follow the digital geographies that surround us.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Although not every statement on Wikipedia is cited, there tends to be a significant proportion of citations, especially to web sources and compared to citations on traditional encyclopedias.
2. See, for example, the WikiCite project https://meta.wikimedia.org/wiki/WikiCite_2016

References

- Berners-Lee T, Hendler J and Lassila O (2001) The semantic web. *Scientific American*. Available at: <http://www.scientificamerican.com/article.cfm?id=the-semantic-web> (accessed 24 October 2015).
- Collier B and Bear J (2012) Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12*. New York, NY, USA: ACM, pp.383–392. Available at: <http://doi.acm.org/10.1145/2145204.2145265> (accessed 24 October 2015).
- Eckert S and Steiner L (2013) (Re)triggering backlash: Responses to news about Wikipedia's gender gap. *Journal of Communication Inquiry* 37(4): 284–303.
- Egenhofer MJ (2002) Toward the semantic geospatial web. In: *Proceedings of the 10th ACM international symposium on advances in geographic information systems*, 4–9 November, pp.1–4. New York, NY: ACM.
- Ford H and Geiger RS (2012) Writing Up Rather Than Writing Down: Becoming Wikipedia Literate. In: *Proceedings of the eighth annual international symposium on wikis and open collaboration, WikiSym '12*, pp.16:1–16:4. New York, NY, USA: ACM. Available at: <http://doi.acm.org/10.1145/2462932.2462954> [accessed 7 March 2015].

- Ford H and Graham M (2016) Semantic Cities: Coded Geopolitics and the Rise of the Semantic Web. In: Kitchin R and Perng S-Y (eds) *Code and the City*. London: Routledge, pp. 200–214.
- Ford H, Sen S, Musicant DR, et al. (2013) Getting to the source: Where does Wikipedia get its information from? In: *Proceedings of the 9th international symposium on open collaboration*, Hong Kong, 5–7 August, pp.9:1–9:10. New York, NY, USA: ACM.
- Graham M (2011) Wiki Space: Palimpsests and the Politics of Exclusion. In: Lovink G and Tkacz N (eds) *Critical Point of View: A Wikipedia Reader*. Amsterdam: Institute of Network Cultures, pp. 269–282.
- Graham M (2013) The virtual dimension. In: Acuto M and Steele W (eds) *Global City Challenges: Debating a Concept, Improving the Practice*. London: Palgrave, pp. 117–139.
- Graham M and Zook M (2013) Augmented realities and uneven geographies: Exploring the geolinguistic contours of the web. *Environment and Planning A* 45(1): 77–99.
- Goodchild MF (2007) Citizens as sensors: The world of volunteered geography. *GeoJournal* 69(4): 211–221.
- Groth PT (2013) The knowledge-remixing bottleneck. *IEEE Intelligent Systems* 28(5): 44–48.
- Haklay M, Singleton A and Parker C (2008) Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass* 2(6): 2011–2039.
- Introna LD and Nissenbaum H (2000) Shaping the web: Why the politics of search engines matters. *The Information Society* 16(3): 169–185.
- Kloc J (2014) What do we make of Wikipedia's falling traffic? *The Daily Dot*. Available at: www.dailydot.com/news/wikipedia-falling-traffic-meaning/ (Accessed 11 July 2016).
- Kolbe A (2016) Wikimedia's Dario Taraborelli quoted on Google's knowledge graph in the Washington post. *Wikipedia Signpost*, 2016-05-17. Available at: https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedia_Signpost/2016-05-17/In_the_media&oldid=722552526 (accessed 9 June 2016).
- Kolbe A (2015) Wikipedia signpost Op-ed: Whither wikidata? *Wikipedia Signpost*, 2015-12-02. Available at: https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedia_Signpost/2015-12-02/Op-ed&oldid=694206756 (accessed 1 July 2016).
- Koünig R (2014) (ed) *Society of the Query Reader: Reflections on Web Search*. Amsterdam: Institute of Network Cultures.
- Morozov E (2013) The internet ideology: Why We are allowed to hate Silicon Valley. *Frankfurter Allgemeine*. Available at: www.faz.net/aktuell/feuilleton/debatten/the-internet-ideology-why-we-are-allowed-to-hate-silicon-valley-12658406.html (accessed 11 July 2016).
- NMaia (2016) Wikipedia signpost Op-Ed wikidata licensing. *Wikipedia Signpost*, 2016-06-15. Available at: https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedia_Signpost/2016-06-15/Op-ed&oldid=725724109 (accessed 11 July 2016).
- Nussbaum MC (2011) *Creating Capabilities: The Human Development Approach*. Cambridge, MA: Harvard University Press.
- Orlikowski A (2014) Google stabs Wikipedia in the front. *The Register*. Available at: http://www.theregister.co.uk/2014/01/13/google_stabs_wikipedia_in_the_front/ (accessed 11 July 2016).
- Pariser E (2012) *The Filter Bubble: What the Internet is hiding from You*. London: Penguin.
- Reagle J (2013) "Free as in sexist?" Free culture and the gender gap. *First Monday* 18(1). Available at: www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/4291 (accessed 11 March 2016).
- Sen A (2001) *Development as freedom*. New York: Knopf.
- Singhal A (2012) Introducing the knowledge graph: Things, not strings. *Google Official Blog*. Available at: <http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html> (accessed 23 December 2014).
- Vaidhyanathan S (2012) *The Googlization of Everything: (And Why We Should Worry)*. Berkeley, California: Univ of California Press.
- Vrandečić D and Krotzsch M (2014) Wikidata: A free collaborative knowledge base. *Communications of the ACM* 78–85.

Heather Ford is a University Fellow in Digital Research Methods at the University of Leeds School of Media and Communication. With a background in internet rights activism working with organisations including Creative Commons, Privacy International, the Association for Progressive Communications and Ushahidi, her research interests include issues around the governance of digital platforms, media power in networked information environments and the design and politics of software platforms. Ford has written for a variety of media and academic publications including the *International Journal of Communication*, the *Association for Computing Machinery* and *Big Data & Society*.

Mark Graham is the Professor of Internet Geography at the Oxford Internet Institute in the University of Oxford. He is also a Faculty Fellow at the Alan Turing Institute, a Visiting Fellow at the London School of Economics and Political Science, and an Associate in the University of Oxford's School of Geography and the Environment. He has published widely in major geography, communications and urban studies journals, and his work has been covered by the media in dozens of countries. He was recently awarded a European Research Council Starting Grant to lead a team to study 'knowledge economies' in Sub-Saharan Africa over five years. This entails looking at the geographies of information production, low-end (digital labour and microwork) knowledge work, and high-end knowledge work (e.g. bespoke programming and work done in innovation hubs) across Africa.