Article

ALEXANDRIA
The Journal of National and International Library and Information Issues

Alexandria: The Journal of National and International Library and Information Issues
1–9
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0955749018778984
journals.sagepub.com/home/ala
SAGE

Metadata: The political dimension

David Haynes
Department of Library & Information Science,
City, University of London, London, UK

Abstract

The use and management of metadata raises many ethical and eventually political issues. The security revelations by Edward Snowden in 2014 demonstrate the key role that metadata plays in surveillance. Privacy has become a particularly hot topic in recent months and much of the debate has centred on the misuse of metadata from social media and the potentially invasive effect this has on individuals. Metadata also has a key role in delivering reliable sources of information, although this has to go hand in hand with user education and improved information literacy. Access to information is a fundamental right and the appropriate use of metadata can help improve access to health, agriculture and education as well as contributing to economic development. Metadata is a route to good governance, but it must also be handled appropriately to maintain information privacy, a fundamental human right.

Keywords

data protection, intellectual property, metadata, privacy

Introduction

If anyone wondered about the importance of metadata, the Snowden revelations about the US government data-gathering activities should leave no one in any doubt (Greenwald, 2014). Stuart Baker, the National Security Agency (NSA) General Counsel, is quoted as saying, ‘Metadata tells you everything about somebody’s life. If you have enough metadata you don’t really need content’ (Schneier, 2015, p. 23). Although some may say that ‘metadata is content’, in many situations, it is actually better than content.

Corresponding author:
David Haynes, Department of Library & Information Science, City, University of London, Northampton Square, London EC1V 0HB, UK.
Email: david.haynes@city.ac.uk

- Metadata is structured. Unlike free text or audio recordings, metadata has fields or some kind of markup that allows investigators to focus on one particular area. For instance, phone records will have details about duration of call, originating number, destination number, time of the call and so on. This makes it relatively easy to search the appropriate column and to collate different types of data to pinpoint the information required.
- The metadata may have some form of vocabulary control. This may be a code from a formal classification scheme or a preferred term from a thesaurus. These types of control enhance the quality of the sets retrieved during a search. For instance, searching the time column of the extracted metadata about phone calls would normally interrogate a system date rather than being dependent on interpreting a potentially ambiguous representation of date (e.g. 03/02/2018 could be 3 February or 2 March depending on whether it is in the British or American date format).
- Metadata is generally less extensive than the content. Security agencies are often plagued by the very large volume of content that they have collected and which is not so easily analysed, whereas a structured file with key information in it is much easier to query.
- It is searchable. It is easier to search the metadata associated with telephone calls than it would be to search the sound recordings of those conversations. Even with the advent of artificial intelligence and speech recognition, reliably interpreting the content of a good voice sound file can be very challenging and may require a lot of processing power.
- It has a context. Metadata also helps to ensure that the term searched is in the right context (e.g. ‘Green’ a noun meaning field rather than someone’s surname). In effect, it provides a form of semantic indexing.

Susan Landau in an interview with *The New Yorker* and speaking about metadata gathered by the US security services from telephone calls put it very succinctly: ‘The public doesn’t understand . . . It’s much more intrusive than content’ (Mayer, 2013). The intrusiveness partly lies in the utility of metadata. If it is easier to analyse individual activity through the metadata, it is likely to have more impact on the individual.

Metadata can be used to track patterns of activity in groups as well as individual online transactions. The advantage from the surveillance point of view is that metadata is structured data, which is more tractable to analysis of the type required to reveal activity. For instance, it is much easier to monitor phone records (i.e. the metadata about phone calls) rather than the phone calls themselves. It is possible to identify who you are calling, when you called, how long you spoke for, your location at the time of the conversation and indeed to track current location through the mobile networks.

This debate about surveillance continues with the UK’s Investigatory Powers Act 2016 that allows considerable latitude in the metadata and personal data that public agencies and security services can collect. Groups such as Electronic Frontier Foundation (www.eff.org) and the Open Rights Group (https://www.openrightsgroup.org/) have expressed concerns about the wide scope of data that can be gathered and the relative lack of safeguards against invasions of privacy. There is a particular concern that large volumes of personal data of individuals who are not under suspicion of any crime are routinely gathered. There is also lack of transparency and accountability because these agencies hide behind a screen of ‘national security’.

Daniel Solove (2011) directly addresses the conflict between surveillance agencies and human rights activists. He makes the point that in actively monitoring the population to keep it safe from terrorism, we are undermining the very values that make for a free society. Anyway, it has not proved to be a very effective means of identifying potential terrorists and thereby keeping us safe. He is challenging the view that privacy has to be sacrificed to improve security. Angwin (2014) has also discussed this in detail.

Big data

So far we have looked at metadata from a single source. However, it is more realistic to consider the effects of aggregating data from a variety of sources. This is the basis of many apps and online services. These systems depend to some degree on metadata. Knowledge graphs are a good example of the way in which metadata is used to construct new datasets.

‘Metadata in aggregate is content’ as Jacob Appelbaum observed when the WikiLeaks controversy first blew up (Democracy Now, 2013). In other words, when metadata from different sources is aggregated, it can be used to reconstruct the information content of individual communications.

Invasion of privacy or personal benefit?

The social media giants prosper by exploiting personal data and targeting digital advertising. Personal profiles of targeted individuals are based on metadata about online use and are the basis of online behavioural advertising. Cookies and other tracking technologies can monitor the online activity of an individual to predict future behaviour. Metadata about online sessions reveals a great deal about an individual and his or her life. This may extend to gathering information about friends, family, colleagues and other contacts.

In a way it is surprising that the Facebook/Cambridge Analytica scandal did not emerge earlier (Cadwalladr and Graham-Harrison, 2018). We will continue to see new revelations in the heightened public awareness of the General Data Protection Regulation and individual privacy rights. In April 2018, the BBC reported that Grindr, a social media site widely used by gay men, was sharing information about HIV status with third parties (Lee, 2018). On the face of it, this is a serious breach of very sensitive personal data. On the other hand, users may have consented to that data being used by apps geared to reminding them to take antiviral medication at specified times. If the UK Information Commissioner’s Office (ICO) decides to investigate, digital forensics will examine the metadata to track online data transactions.

Risks

Metadata describes an information object whether that be raw data or more descriptive information about an individual. This is important because the treatment of metadata has become a political issue. Personal data, especially data that reveal opinions, attitudes and beliefs, is potentially very sensitive. Use of this personal data by service providers or by third parties can expose users to risks such as *nuisance* from unwanted ads, *harassment* from internet trolls or *fraud* through identity theft. Many digital advertisers would say that because these data is aggregated, it is not possible to identify individuals – that is, these data is anonymised. However, there is no protection against privacy breaches as has been demonstrated by Narayanan and Shmatikov (2009) and others. It is relatively easy to *de-anonymise* data by combining it with other available sources. We should talk about pseudonymisation in these cases. On a practical level, keeping a record of the anonymisation applied to sets of personal data is probably a good idea – metadata.

Fact-free content

Daniel Rosenberg (2013) makes a nice distinction between 'data', 'facts' and 'evidence'. Data comes from the Latin verb *dare*, to give as in something given in an argument. Fact comes from *facere* to do, as in that which was done or occurred or exists. Evidence comes from *videre*, to see as in something witnessed. As Rosenberg said: 'When a fact is proven false, it ceases to be a fact. False data is data nonetheless'. Let me illustrate that with an example: Data suggested that there were nine planets in the solar system. That was a fact from 1930 until 2006 when Pluto was reclassified as a dwarf planet, and not even the largest one at that (International Astonomical Union, 2018). Suddenly we were back to eight planets, even though no one was disputing the existence of Pluto. The data is still valid.

Samuel Arbesman (2012) in his book *The Half Life of Facts* introduced the idea that in a given period, half the certainties that we had are shown to be false or are superseded by new understandings and that they cease to be 'facts'. Data, whether it is true or not, continues to be data but is only factual if true. Perhaps there is some way of recording the reliability of information or data, so that it can be exploited appropriately. Many of the arguments and counterarguments on climate change, for instance, centre on the quality and veracity of the evidence used by each side of the debate. This idea is not new, as medical researchers have for some time evaluated the quality of research used to make clinical decisions (The Cochrane Collaboration, 2018). This information about the quality and reliability of data is another example of metadata.

This is to suggest that although the 'Fake news' debate is important, the issues are more nuanced than some of commentators would have us believe. Many news stories occupy a grey area between fact and fiction. Services such as the Post-Truth Forum and CILIP's (The library and information association) 'Facts Matter' campaign provide a useful exploration of the issues and help individuals to navigate this complex landscape (Clarke, 2018; CILIP, 2017). Cooper (2017) describes some of the emerging services and initiatives that help to protect against fake news, such as 'Full Fact', 'First

Draft' and dminr. She emphasises the role of journalists both as purveyors of fake news and as protectors against it. At the BCS IRSG/ISKO UK Search Solutions 2017 meeting, Mark Harwood described techniques for identifying which news stories on social media had been manipulated. This was based on the metadata associated with the stories themselves and the resulting comments.

Of course, ethics is wider than privacy or fake news, and it is worth exploring some of the other ethical issues that arise when people talk about metadata. These issues can be put in the context of metadata's role in managing the 'information communication chain' (Robinson, 2015).

Ownership

Intellectual property (IP) and ownership of data raises the question: 'Who actually owns information and how is that reflected in the metadata?' This is a particularly active debate in the academic world at the moment. If research is publicly funded, why are the results only available behind a paywall? Some research funders are getting around this by paying publishers an open access fee to allow general access to research results. There is also the Creative Commons (2016) movement where people are encouraged to share content more openly. Well-established standards such as Dublin Core acknowledge the importance of IP and make some provision for this, with its 'dc:rights' data element. This is an example of metadata providing a means of recording IP ownership, although not necessarily facilitating wider access. The metadata itself is IP, and this does raise issues about copying bibliographic records. There are mechanisms for acknowledging the IP of an original publication but less so for acknowledging the ownership of the cataloguing data – generated by skilled professionals. Some of this hides behind paywalls, but more often it is seen as a route into the richer territory of paying for content.

Information inequality

Information inequality is a longer term problem. Although there have been various international initiatives, they give the impression of being divorced from the concerns of ordinary people. They are often high-level gatherings of government functionaries or senior executives of large enterprises. Looking at specific factors that affect the digital divide, Pick and Nishida (2015, p. 15) found that:

> Overall, the most significant determinants of technology utilization and availability are tertiary education, capacity for innovation, judicial independence, and foreign direct investment. The findings imply that influences of technology utilization are distinctive between developed and developing regions and between major continental regions.

Perhaps more interesting is the work done by academics and by campaigning groups to improve information literacy. The Prague Declaration describes information literacy as: 'key to social, cultural, and economic development of nations and communities, institutions and individuals in the 21st century' (UNESCO, 2003). One aspect of information poverty is lack of exposure to and knowledge of reliable and relevant sources of information. Information poverty and poor information literacy are both barriers to access to reliable information on health and nutrition, agriculture, education and social welfare.

Describing resources in a way that is relevant to the lives of individuals is nothing new. The UK government attempted this with its metadata strategy, starting with the Modernising Government White Paper (Cabinet Office, 1999). This initiative was at a time when the internet was emerging as a way of delivering access to government services and there was a recognition of the need to use structured information to describe these resources. The e-Government Metadata Standard based on Dublin Core was a direct result of this (e-Government Unit, 2006).

The growing sophistication of online search engines has to some extent bypassed the use of applied metadata, but it is making a comeback with linked data. In my book, information retrieval is the second purpose of metadata, the first being identifying and describing resources (Haynes, 2018). An understanding of the way in which metadata works is essential for those responsible for making information resources available over the internet and is probably a good idea for those acting as intermediaries. Understanding user needs is essential for anyone involved in the design of systems or development of taxonomies or thesauri to enhance retrieval. This is all metadata and needs to be delivered in a way that can be interrogated by search engines or search services.

Giving control back to individuals

A common theme that has arisen in the development of information services in the last 30 years is the empowerment of individuals, the people who use these systems. At one time, searching bibliographic databases was a highly specialised role and required skilled professionals to act as intermediaries. Batch searching overnight meant that a lot of care had to be taken to get the search strategy right first time. Otherwise it meant having to wait until the next day to resubmit a modified search query. Even with the emergence of online searching, familiarity with systems and understanding of the codes used and the controlled vocabulary were necessary. Now the emphasis of information professionals is on acting as enablers and trainers. The customer wants to do the search for herself or himself. This democratising of information access, perhaps creates greater responsibilities for the information professional. Not only does it mean greater emphasis on teaching information literacy skills, it also means ensuring the quality of the metadata associated with information resources which will make the information accessible to the users.

Information governance

So far I have talked about the ethics of metadata – the issues that arise and some of the dilemmas we face when we manage and use information. I would now like to consider ways in which metadata plays a role in information governance. This is another aspect of ethics and underpins many aspects of corporate governance. Consider how important it is, for instance, to be able to secure and protect confidential or personal data from unauthorised access. Being able to examine an audit trail of a document or access to a record provides one level of protection against wrongful or unlawful access to data as has happened in stalking cases. Metadata is an essential part of digital forensics covering not only stalking but also money laundering and fraud. Audit trails based on metadata can also be used in heavily regulated industries where it may be necessary to demonstrate that staff have followed procedures or have had access to up-to-date information when making decisions.

Conclusion

To conclude, I make the case that metadata has become a political issue. Its use by the security agencies, the privacy concerns about online behavioural advertising and the power relationships that emerge from the management of information resources through its metadata are all indications of its relevance. Anyone who had asked the question 'What does metadata matter?' prior to 2013 will realise now just how important a bearing it has on current political issues. The Fourth Amendment to the US Constitution protects 'The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures'. That is being challenged by the way in which government agencies have been exploiting metadata. Metadata is here for good or ill – as citizens and information users, we need to be vigilant about what data is out there, how it is managed and what controls are in place to ensure good governance.

Author's note

Based on a talk given at the launch of the second edition of *Metadata for Information Management and Retrieval* at City, University of London on 4 April 2018.

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 acknowledges the support of the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Fellowship Programme.

References

Angwin J (2014). *Dragnet Nation: a quest for privacy, security, and freedom in a world of relentless surveillance*. New York: Times Books, Henry Holt and Company.

Arbesman S (2012) *The Half-Life of Facts: Why Everything We Know Has an Expiration Date*. New York: Current.

Cabinet Office (1999) Modernising Government White Paper. Cm 4310, London. Available at: http://webarchive.nationalarchives.gov.uk/20060802203310/http://archive.cabinetoffice.gov.uk/moderngov/whtpaper/index.htm (Accessed 23 May 2018).

Cadwalladr C and Graham-Harrison E (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Observer*. (London) 17 March 2018. Available at: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (Accessed 23 May 2018).

CILIP (2017) Facts matter. Available at: https://archive.cilip.org.uk/advocacy/facts-matter (accessed 16 March 2018).

Clarke D (2018) Post-Truth Forum. *Synaptica*. Available at: https://www.posttruthforum.org/ (accessed 16 April 2018).

Cooper G (2017) False news: a journalist's perspective. In: Knowledge Organization: What's the Story. ISKO UK 2017. London: ISKO UK. Available at: http://www.iskouk.org/content/false-news-—-journalist's-perspective (Accessed 23 May 2018).

Creative Commons (2016) Creative Commons. Available at: http://creativecommons.org (accessed 5 February 2016).

Democracy Now (2013) Court: Gov't can secretly obtain email, Twitter info from Ex-WikiLeaks volunteer Jacob Appelbaum. Available at: https://www.democracynow.org/2013/2/5/court_govt_can_secretly_obtain_email (accessed 21 March 2017).

e-Government Unit (2006) e-Government Metadata Standard ver. 3.1, p.59. Available at: http://www.nationalarchives.gov.uk/documents/information-management/egms-metadata-standard.pdf (accessed 24 March 2016).

Greenwald G (2014) *No Place to Hide*. London: Hamish Hamilton.

Haynes D (2018) *Metadata for Information Management and Retrieval: Understanding Metadata and Its Use*, 2nd ed. London: Facet Publishing.

International Astonomical Union (2018) Pluto and the developing landscape of our solar system. Available at: https://www.iau.org/public/themes/pluto/ (accessed 13 April 2018).

Lee D (2018) BBC News: Grindr defends HIV-related data sharing. *BBC News Online*. Available at: http://www.bbc.co.uk/news/technology-43624328 (accessed 4 April 2018).

Mayer J (2013) What's the matter with metadata? *The New Yorker* 6 June 2013. Available at: https://www.newyorker.com/news/news-desk/whats-the-matter-with-metadata (Accessed 23 May 2018).

Narayanan A and Shmatikov V (2009) De-anonymizing social networks. In: *2009 30th IEEE symposium on security and privacy*, Oakland, CA, 17–20 May 2009, pp. 173–187. Washington DC: IEEE Computer Society.

Pick JB and Nishida T (2015) Digital divides in the world and its regions: a spatial and multivariate analysis of technological utilization. *Technological Forecasting and Social Change* 91: 1–17. Available at: http://www.sciencedirect.com/science/article/pii/S0040162514000079.

Robinson L (2015) Multisensory, pervasive, immersive: towards a new generation of documents. *Journal of the Association for Information Science and Technology* 66(8): 1734–1737.

Rosenberg D (2013) Data before the fact. In: Gitelman L (ed) *'Raw Data' is an Oxymoron*. Cambridge, MA: MIT Press, pp. 15–40.

Schneier B (2015) *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. New York: W.W.Norton.

Solove DJ (2011) *Nothing to Hide: The False Tradeoff Between Privacy and Security*. New Haven: Yale University Press.

The Cochrane Collaboration (2018) Cochrane. Available at: http://www.cochrane.org/ (accessed 13 April 2018).

UNESCO (2003) *Towards an information literate society – Prague Declaration*. Available at: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/PragueDeclaration.pdf (Accessed 23 May 2018).

U.S. Const. amend. IV.

Author biography

David Haynes is an intelligence community postdoctoral research fellow at City, University of London where he is currently conducting research into the concept of risk in the context of online privacy. He teaches ‘Information Management and Policy’ on City’s innovative CityLIS master’s programme in library and information science. He is also an honorary tutor at the Centre for Archives and Information Services (CAIS) at the University of Dundee where he teaches a module on ‘Metadata Standards and Information Taxonomies’. During his career, David has had extensive experience in all sectors of the library and information domain, including public, academic, special and national libraries. He is particularly interested in the way in which individuals interact with online systems and in the way in which human rights are affected by regulation of information systems. David is a Chartered Fellow of CILIP and chairs the UK chapter of International Society for Knowledge Organization (ISKO). The second edition of his book *Metadata for Information Management and Retrieval* was published by Facet Publishing in 2018.