
Article:
Tattersall, A. orcid.org/0000-0002-2842-9576 and Grant, M.J. (2016) Big Data - What is it and why it matters. Health Information and Libraries Journal, 33 (2). pp. 89-91. ISSN 1471-1834

https://doi.org/10.1111/hir.12147

Big Data - What is it and why it matters to Library and Information Professionals

Abstract
Big data, like MOOCs, altmetrics and open access are all terms that have been widely banded about the library community for some time. Whilst some are unsure what these things are, despite all being around for at least five years, many in the library and information sector remain confused as to the relationship between these terms and their roles. Whilst all of these developments do indeed have something to offer to the library and information community, big data perhaps remains the most ambiguous.

Keywords: Big Data, Google, Twitter, Microsoft, Librarian, Library, Health, Research Data

What is Big Data and where did it come from?
(1) Dan Ariely crudely, but accurately compared to teenage sex: “Everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.” Big data is certainly making a big noise but for many in the library community they might not realise how they can can get involved, it can seem to the outsider that this is only for researchers and IT experts. We have long been able to host large sets of data, but with the decrease in cost for data storage and increase in computing power it has become much more commonplace. What big data can do is give us new insights into problems we have previously struggled to solve or attend questions we have not yet asked. So much so that it can come down to what an organisation will do with a large dataset.

Google Flu Trends
For the best early example of big data in practice is that of Google Flu Trends which first appeared publicly in 2008. This was a good example of having a large set of data and not immediately spotting its potential application. Researchers at Google realised that it was receiving a high amount of search requests for flu and by analysing geographical data from user’s IP addresses behind these searches they could predict flu epidemics in parts of the U.S. Flu Trends then extended to other countries and was regarded as a better measure for predicting influenza epidemics than the Centers for Disease Control and Prevention (CDC). Sadly however the model was flawed as news of Google Flu Trends spread across the media, more and more people searched the web for news of flu trends in their area, therefore biasing the data. Google Flu Trends is no longer regarded the accurate measure for predicting epidemics as it once was, but still opened a lot of people’s eyes to big data’s potential. Google Flu Trends and the big data it generated proved to be a very valuable and useful tool in predicting epidemics and showed many what could be achieved by applying new uses on existing large sets of data. In addition it helped opened up the debate on big data and showed some of the flaws associated with it (2). Since then various studies (3) (4) (5) have looked at the vast generated by Twitter’s as a way of forecasting flu epidemics.

The Ethics of Big Data
More recently one of Google’s big competitors Microsoft - who are no stranger themselves to dealing with huge sets of data - suggested it was possible to predict whether a woman could be suffering with postnatal depression before giving birth based on her use of language in Twitter (6). The micro blogging site generates a tremendous amount of data and by looking for certain keywords it can potentially help identify problems such as postnatal depression.

The researchers at Microsoft Labs looked for changes in Tweets after the birth of a user's baby for any changes in the words they used.

The Google and Microsoft Labs stories have showcased the potential for big data in a healthcare setting. Yet there are still questions around ethics and privacy in using the public's data, although publicly available still used without consent. The Google data can be traced by IP address to a geographical location, the Twitter data points to a profile and other possible personal data including location. Most people using the web and social media regard their content as private, often unaware that it is often out there for everyone to see.

Where do librarians and information professionals fit in?

Big data has great potential in the healthcare setting, and as we have seen it is something that the technology giants are very much interested in. A lot of big data being generated is by the public sector, local government, healthcare and education. Areas where librarians have important parts to play. In my sector, higher education, institutions have been creating new roles for research data managers to help support those creating their own data sets. Whilst in the healthcare setting, research funders have increasingly placed more requirements on anyone using their data. As a result there has been an emergence of the research data manager role (7). Often these roles are located in the library and facilitated by a professional with a library or information background. The reason being more than just that librarians are good at organising and storing data but that they are often aware of the implications of such tasks. These include security, access, storage and governance. As with the aforementioned MOOCs, altmetrics and open access; big data is something that librarians can get involved in at a strategic and functional level. The research data manager role is not exclusive to big data, but given the overlap and shared interest between the two it should follow that they will become increasingly intertwined.

Ethics, information governance and data application are still playing catch up with the ever-changing web and technology world. The creation of more research data management posts is inevitable so it should matter to library and information professionals to keep an interested eye on developments in this area. At a time when parts of the library and information world is in state of fight or flight due to cutbacks across various public and private sectors, the growth in big data, as with the other initiatives mentioned should be seen as positives. There are still emerging areas and ones where LIS professionals could move into as part of a career progression.

Big data may not answer all of society's questions, especially in a health context. On the other hand, it could give us many of the same answers, just with a bigger data set. What it can do is answer some of the questions we have not yet fully attended, it has the potential to open up new areas of research, especially in a health. Nevertheless we will still have questions about ethics and privacy and the use of public data from such as social media is certain to lead to more criticisms of its use.

1. Ariely D. No Title [Internet]. Facebook. 2013 [cited 2016 Jan 18]. Available from: https://www.facebook.com/dan.ariely/posts/904383595868

2. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science (80-) [Internet]. 2014;343(6167):1203–5. Available from: http://www.sciencemag.org/content/343/6176/1203

3. Paul MJ, Dredze M, Broniatowski D. Twitter Improves Influenza Forecasting. PLoS Curr [Internet]. San Francisco, USA: Public Library of Science; 2014 Oct 28;6:ecurrents.outbreaks.90b9ed0f59bae4ccaa683a39865d91. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4234396/

4. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B. Twitter Improves Seasonal Influenza Prediction. HEALTHINF [Internet]. Vilamoura; 2012 [cited 2012 Jan 25]. Available from: http://www.cs.uml.edu/~hachreka/SNEFT/images/healthinf_2012.pdf

5. Santos JC, Matos S. Analysing Twitter and Web Queries for Flu Trend Prediction. Theor Biol Med Model [Internet]. BioMed Central; 2014 May 7;11(Suppl 1):S6–S6. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4108891/

6. De Choudhury M, Counts S, Horvitz E. Predicting Postpartum Changes in Emotion and Behavior via Social Media. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: ACM; 2013. p. 3267–76. Available from: http://doi.acm.org/10.1145/2470654.2466447

7. Surkis A, Read K. Research data management. J Med Libr Assoc [Internet]. Medical Library Association; 2015 Jul;103(3):154–6. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4511058/