CHAPTER 6

Big Data, Bad Metadata

A Methodological Note on the Importance of Good Metadata in the Age of Digital History

Kimmo Elo

Introduction

During the past decade, digital humanities has emerged as a new paradigm seeking to gather scholars interested in applying computational methods on their research materials. This development has been supported by the almost exponential growth of either born-digital or digitised materials currently available for researchers. Further, the availability of computational research tools is much better today than, say, five or 10 years ago.

New terminology like big data, data mining and text mining well illustrate the massive growth of digital data available for research purposes. At the same time, the digital research agenda is filled with huge expectations regarding exploratory research, the growth of scientific and societal knowledge or new forms of data analysis. Some scholars have rather strong expectations about how digital humanities should change our whole understanding of knowledge and how knowledge is presented.[1]

This chapter supports the general understanding of digital humanities as an important, computational field of research for the Humanities and social

How to cite this book chapter:
Elo, K. (2020). Big data, bad metadata: A methodological note on the importance of good metadata in the age of digital history. In M. Fridlund, M. Oiva, & P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history* (pp. 103–111). Helsinki: Helsinki University Press. https://doi.org/10.33134/HUP-5-6

sciences in general, and for historical research in particular. The chapter stems from the deep conviction of a scholar rooted in the intersection of computational, historical and social scientific research that exploring digitised historical sources could help us to gain new insights and improve our understanding of the past.

At the same time, however, this chapter is motivated by my worry that, as regards historical research, thus far much attention has been paid to the creation of digital research material, but too little has been to paid to the creation of research data. To clarify my point, with *research material*, I refer to original, primary sources like documents, letters, photographs, etc. With *research data*, I refer to corpora consisting of both the original material and additional, descriptive information derived from the original material. To put it bluntly, we are almost over-flooded by the former, but there still is no shared or common strategy about how to cope with the latter. The importance of the latter is, however, reflected by the fact that many universities are developing research data management practices.[2]

The main thesis of this chapter is that more attention should be paid and more resources should be invested in metadata creation. The next section introduces the very concept of metadata and tackles the question of why metadata matters. The second section presents arguments about why metadata should be considered as an important part of digitising projects. The chapter is rounded up with concluding remarks related to the future work in digital history.

What Is Metadata and Why Do We Need It?

Due to the limited space available for this chapter, I refrain from a literature review and just point out some of the most important aspects related to metadata and discussed (mostly) by librarians or archivists. Metadata is widely understood and defined as 'data about data' and, thus, is expected to provide information about the content of the material it is linked with. In other words, metadata should summarise the most important content. According to *The metadata handbook*, metadata should be constructed in a way which 'fully supports findability and discovery'.[3]

According to Allen Benson, metadata is a descriptive model, a summary report to present the main content according to a formalised structure consisting of information-bearing entities.[4] Richard Pearce-Moses defines metadata creation as the 'process of creating a finding aid or other access tools that allow individuals to browse a surrogate of the collection to facilitate access and that improve security by creating a record of the collection and by minimizing the amount of handling of the original materials'.[5] Hence, metadata is an ontological model providing a structure for information arrangement. At the same time, metadata creation is a descriptive process aiming at filling in the ontological model with material-related, descriptive information.

I am quite convinced that the ontological side is not the core problem. Several well-developed models exists as to how metadata should be structured or what descriptive elements are available in order to guarantee a standardised, formalised metadata.[6] Further, as regards born-digital materials specialists have been discussing from the late 1990s onwards how this development affects ontological requirements for the metadata.[7]

Hence, the real problem is the metadata creation process, especially when this process must be started from scratch and/or with only limited previous knowledge about the full content of the material to be modelled and summarised into metadata. Although the metadata should fulfil a relatively straightforward task (namely, support findability and discovery), at least three main pitfalls should be taken seriously.

First, what or who determines the elements included in the metadata structure? The answer to this question widely determines the content described and formalised in the metadata. At the same time, however, it has a strong impact on both findability and discovery, since metadata queries are limited to the fields used in the model. A more complicated issue relates to hierarchies or sub-categories typical for historical sources (for example, 'building'–'house' or 'building'–'church'). Two examples should clarify the point. Let us first consider a novel. A standard metadata includes the author(s), the title, the publisher, the year of publication, the genre and a few keywords used to summarise the main content. In most cases, these elements suit well the needs of a reader looking for certain novels. But how about a researcher looking, for example, for novels with a certain type of protagonist or a certain person/figure? Or, second, a photograph collection. Once again, many elements to be included in the metadata are quite straightforward and obvious (timestamp, photographer, title), but how about persons, places or abstract elements like gestures, memes or visual effects? The answers depend on the supposed group of end-users and, thus, make the material unusable or unfindable for certain groups.

Second, what or who determines the terminology (for example, keywords, descriptions) used to describe content? Once we have determined what content should be summarised in the metadata, we need to determine how different content-related aspects are described. Once again, standardised dictionaries, keyword indices, etc. exist, so there is rarely a need to reinvent the wheel by creating a new vocabulary. The challenge is to maintain coherence; that is, to ensure that the same (or similar) content is described in the same terms. To use a simple example, if there are bunches of photographs all having different kinds of buildings in them, all of these photographs should be found if one searches for 'building'. But should the end-user be able to find buildings of the type 'church' as well? Once again, findability should guide the process of metadata creation.

And, third, who creates and maintains the metadata? Prior to the digital era, collection management and metadata creation have been almost solely in the hands of librarians and archivists, especially when it came to the creation

and maintenance of large document collections.[8] Today, many collections are created, maintained and made available by private organisations, institutions and companies. This is partly due to the limited resources of public institutions like state archives or libraries, but also thanks to the reduced costs of digitisation, the increase of easy-to-use solutions for data management and hosting, and to the growth of data-sharing platforms like cloud-based services. The other side of the coin is that a majority of these platforms is rather weak and underdeveloped in metadata creation and maintenance, especially as regards the content description. One solution enjoying growing popularity is 'crowdsourcing', a process where 'ordinary people' help the maintainers to create descriptive metadata. There are many examples ranging from 'tagging' over 'person identification' to 'linked data creation', all of them producing interesting and promising results, but also highlighting many problems mostly related to the heterogeneous quality of the resulting metadata and difficulties in ensuring the correctness of input.[9]

Why Digital History Should Take Metadata Seriously

A quick survey in recent literature around digital history reveals that questions related to metadata creation have rarely been debated among digital historians. Instead, historians seem to be educated to use metadata when searching for sources, not to question the metadata itself. In other words, we are used to relying on metadata created by archivists or librarians.[10] This was a good practice in the times when collections were mainly and dominantly housed by libraries and archives.

The digital era has already changed this division of labour, and there is no evidence whatsoever that this would change in the future. Quite the contrary, billions of gigabytes of born-digital textual and visual materials are produced and made available without any, or with only weak and incomplete, metadata. However, without a proper metadata, materials 'are simply a meaningless collection of files, values and characters'.[11] And as Edelstein and colleagues point out: 'Historians increasingly find themselves utilizing digital databases as the idea of the searchable document and the virtual archive reorganize how libraries, research institutes, teams of scholars, and even individual researchers present and share interesting sources.'[12]

Quite much effort, money and time have been invested in the digitising of historical textual materials like manuscripts, documents, letters, etc. As a result, historians have access to a vast number of digitised text and can view and query digitised indexed document collections and editions online. One of the most prominent examples is the 'Republic of Letters' project, focusing on historical networks of correspondence between scholars from all around the world.[13] Another similar project is the 'Letters of 1916 Digital Edition' project, one of the first crowdsourced humanities projects, as well as histoGraph, which also uses crowdsourcing for metadata creation.[14]

In their evaluation of the 'Letters of 1916' project, the authors note that '[t]he meaning of the term "metadata" was unclear for most participants'.[15] This seems to be linked to a wider aspect, namely that '[m]uch attention in the past fifteen years has been directed toward text digitization',[16] forcing 'scholars to access historical sources in a new way: through specific words'.[17] As a result, most digitised collections available online are 'focused on searching, not browsing'.[18] Hence, findability might be good (thanks to the power of full text search in digitised text documents), whereas discovery might be poor.

Modern text mining methods can be of help when historians are dealing with digitised textual corpora. Further, computational methods like (semi-) automated document classification or indexing can make the metadata creation process easier and more effective. However, the current tendency to make old documents available as PDF collections worsens the situation. The positive thing in using the so-called layered PDF format is that end-users can see the original document, but also use search and copying functionalities through the text layer. The negative side is that in most cases the text layer is an exact, character-based reconstruction of the page (mostly based on the corrected results from the optical character recognition (OCR) process), not a raw text laid out and paginated according to the original design. As a result, hyphened words, to give an example, on two lines are not understood as one, but as two separate words (of which the first ends with a hyphen!). My reader can imagine what kinds of limitations result from this kind of practice for document discovery, even if the research interface offers expanded search capabilities like regular expressions. This is because most search engines are based on pattern matching, whereas, for example, irregularly split words do not have a distinct pattern.

Another growing challenge is that sources relevant for historians and social scientists include not only textual collections, but also visual or audio materials like photographs, music, films and so on. Although the question of metadata creation is relevant for all digitised collections, the real challenge relates to non-textual materials. Since the share of information delivered in non-textual, mostly visual forms is steadily growing, the problem of findability and discovery of such materials is of increasing relevance also for historians. There exists already vast collections of such materials, but at the same time our tools to directly query visual or audio materials are very limited, yet slowly improving.[19] For example, many digitised historical photographs include non-recognised persons or places, but the problem is also relevant for today. According to de Figueirêdo and Feitosa '[a]pproximately 350 million photos are added to Facebook each day[, but most of them] are not annotated'.[20] The problem here is not just about forgetting, but also about findability and discovery. Non-annotated photographs cannot be queried, and they do not appear in search results, even if their content was relevant for the query. How are we expected to find, for example, photographs with 'Konrad Adeanuer' on them if we lack both techniques to identify (that is, to name) persons behind recognised faces and metadata containing information about persons shown on the photographs?

Many recent articles point out that digitised collections and online resources affect the way in which scholars discover and access historical sources. Instead of selecting research material from the sources by close reading, research material is increasingly selected by using search engines or by applying methods of computer-aided, distant reading. Two biasing consequences seem worth being noted. First, the use of search engines and other online resources might influence and steer scholars to favour materials available online and, consciously or unconsciously, to change their research questions to suit digitally available materials. Second, scholars might not be aware of missing or incomplete metadata possibly affecting and limiting research results. This second aspect is especially relevant for non-textual material collection, but has at least some relevance also in regard to textual data offered as simple, non-indexed PDF document collections. Another problem is that many collections do not provide any information about the completeness (or better: incompleteness) of their data.

Discussion

This chapter has tackled the question of the relevance of metadata for historical research. Metadata is understood as 'data about data', an ontological model summarising the main content of the data. The very idea of metadata is to make the source material findable and discoverable. In the current digital era characterised by the exponential growth of digitised materials and the availability of vast online resources, both goal-settings gain in importance also for historical research.

Based on the arguments presented above, I conclude that metadata is extremely relevant also for historians. On the one hand, historians increasingly use and explore online resources like historical document collections or photograph corpora. Most of these online portals offer search engines or other possibilities to query the collections. Instead of selecting material by the process of reading the material document by document, material selection is increasingly based on search results. Since there is no reason to believe that this will change in the future, historians should be interested in ensuring that all relevant aspects are searchable, findable and discoverable.

On the other hand, the whole collection management is in flux, as digitised collections are made available by a wide variety of actors. If there exist no standards for quality management of data collection, how can findability and discovery be guaranteed? Once again, the ontological side is not the problem. The problem is the process of creating annotations and metadata.

A third aspect should be added to the two points above. Historical digitisation projects often deal with materials of which only trained historians possess knowledge. With all respect to librarians and archivists, we cannot expect them to have an in-depth knowledge of historical persons, events or eras. Despite this, these two groups are still in charge when national, governmental and official collections are digitised and annotated with metadata.

Although there is no easy patent solution regarding how to ensure metadata quality for historical collections, historians should be encouraged to engage in digitisation projects in their own fields of expertise. As Reilly point out, libraries, but also archives, 'must ensure that they maximize the visibility of their collections—not just to the general public but to those in the education system'.[21] In this respect, historians should engage as mediators between the research community and libraries and archives.

Historians value original documents and are trained to source criticism and to work in archives. At the same time, they are quite reliable on what is involved in the quality of collection management and hosting in archives, and many archivists and librarians enjoy a high respect for their expertise. A good archivist can fill the gaps in a researcher's inquiry and, thus, find relevant and reliable sources.

The shift from this human-to-human interface towards a human-to-computer interface replaces the 'silent knowledge' of an archivist with algorithms run by the computer. The search process itself might be more effective and quicker, but the other side of the coin is that the user has only limited possibilities to explain her intentions. As pointed out above, a scholar is forced to figure out correct terms and words for his query, but still he cannot be sure whether he receives all (or even the most) relevant materials.

To round up my argument: it is by far not sufficient to digitise original sources if we cannot ensure findability and discovery. Digitised original sources must be processed into research data consisting of the original content plus descriptive metadata summarising the essential content of the material. Metadata creation should not be disparaged, nor should it be seen as a quick, dirty task to be completed as soon as and as inexpensively as possible. Research data is the most valuable content of a vast material collection, since it enables both findability and discovery. If scholars cannot rely on getting reliable results when committing searches in online collections, the digital leap manifested by proponents of digital humanities might end with a belly flop.

Notes

[1] See, e.g., Burdick et al. 2012.
[2] See, e.g., https://www.helsinki.fi/en/research/research-environment/research-data-management.
[3] Register & McIlroy 2015.
[4] Benson 2009: 161–162.
[5] Pearce-Moses 2005: 112–113.
[6] Benson 2009; Gonzales 2014; Valentino 2017.
[7] Langdon 2016.
[8] Edelstein 2017: 401.
[9] See, e.g., Stvilia 2009; Reilly 2012; Stvilia 2012; Turin 2015; Valentino 2017; Wusteman 2017.

[10] Edelstein 2017: 401.
[11] See https://www.fsd.uta.fi/aineistonhallinta/en/data-description-and-metadata.html.
[12] Edelstein 2017: 401.
[13] Stanford University 2013.
[14] Letters 1916–1923 Consortium 2016; University of Luxembourg 2018.
[15] Wusteman 2017: 133.
[16] Edelstein 2017: 417.
[17] Huistra 2016: 220.
[18] Ibid.: 222.
[19] See, e.g., Huang, Ma & Gong 2014; Ries & Lienhart 2014; Ko & Lee 2015; Vinyals et al. 2015; Li, Wang & Zhang 2016; Osadchy, Karen & Raviv 2016; Wang, Wang & Liu 2016; Zhong, Liu & Hua 2016; Li et al. 2017.
[20] de Figueirêdo & Feitosa 2015: 203.
[21] Reilly 2012: 39.

References

Benson, A. C. (2009). The archival photograph and its meaning: formalisms for modelling images. *Journal of Archival Organization, 7*(4), 148–187.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & **Schnapp, J.** (2012). *Digital_Humanities*. Cambridge, MA: MIT Press.

de Figueirêdo, X., & **Feitosa, H.** (2015). Semi-automatic photograph tagging by combining context with content-based information. *Expert Systems with Applications, 42*(1), 203–211.

Edelstein, D. (2017). Historical research in a digital age: reflections from the mapping the republic of letters project. *American Historical Review, 122*(2), 400–424.

Gonzales, B. (2014). The conversion of MARC metadata for online visual resource collections: a case study of tactics, challenges and results. *Library Philosophy and Practice (e-journal)*, 1–64.

Huang, M., Ma, Y., & **Gong, Q.** (2014). Image recognition using modified zernike moments. *Sensors & Transducers, 166*(3), 219–223.

Huistra, H. (2016). Phrasing history: selecting sources in digital repositories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 49*(4), 220–229.

Ko, C.-N., & **Lee, C.-M.** (2015). Image recognition using adaptive fuzzy neural network based on lifting scheme of wavelet. *Artificial Life and Robotics, 20*(4), 353–358.

Langdon, J. (2016). Describing the digital: the archival cataloguing of born-digital personal papers. *Archives and Records, 37*(1), 37–52.

Letters 1916–1923 Consortium. (2016). *Letters of 1916 digital edition*. Retrieved from http://letters1916.maynoothuniversity.ie/

Li, K., Wang, F., & **Zhang, L.** (2016). A new algorithm for image recognition and classification based on improved bag of features algorithm. *Optik—International Journal for Light and Electron Optics, 127*(11), 4736–4740.

Li, W., Chen, L., Xu, D., & **Gool, L.V.** (2017). Visual recognition in RGB images and videos by learning from rgb-d data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PP*(99), 1–1.

Osadchy, M., Keren, D., & **Raviv, D.** (2016). Recognition using hybrid classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(4), 759–771.

Pearce-Moses, R. (2005). *A glossary of archival and records terminology*. Chicago, IL: Society of American Archivists.

Register, R., & **McIlroy, T**. (2015). *The metadata handbook*. Retrieved from http://themetadatahandbook.com/wp-content/uploads/2015/01/Metadata-Handbook-Preview-Revised.pdf

Reilly, S. K. (2012). Collaboration to build a meaningful connection between library content and the researcher. *New Review of Information Networking, 17*(1), 34–42.

Ries, C. X., & **Lienhart, R.** (2014). A survey on visual adult image recognition. *Multimedia Tools and Applications, 69*(3), 661–688. Copyright: Springer Science+Business Media, New York, 2014; last updated 30 August 2014.

Stanford University (2013). *The republic of letters*. Retrieved from http://republicofletters.stanford.edu/

Stvilia, B. (2009). User-generated collection-level metadata in an online photo-sharing system. *Library and Information Science Research, 31*(1), 54–65.

Stvilia, B. (2012). Establishing the value of socially-created metadata to image indexing. *Library and Information Science Research, 34*(2), 99–109.

Turin, M. (2015). Devil in the digital: ambivalent results in an object-based teaching course. *Museum Anthropology, 38*(2), 123–132.

University of Luxembourg (2018). *histoGraph*. Retrieved from http://histograph.eu/

Valentino, M. (2017). Linked data metadata for digital clothing collections. *Journal of Web Librarianship, 11*(3–4), 231–240.

Vinyals, O., Toshev, A., Bengio, S., & **Erhan, D.** (2015, June). Show and tell: a neural image caption generator. In Computer Vision and Pattern Recognition conference (pp. 3156–3164).

Wang, Y., Wang, X., & **Liu, W.** (2016). Unsupervised local deep feature for image recognition. *Information Sciences, 351*, 67–75.

Wusteman, J. (2017). Usability testing of the letters of 1916 digital edition. *Library Hi Tech, 35*(1), 120–143.

Zhong, S.-H., Liu, Y., & **Hua, K. A.** (2016). Field effect deep networks for image recognition with incomplete data. *ACM Trans. Multimedia Comput. Commun. Appl., 12*(4), 52:1–52:22.