Trust and Understanding. The Value of metadata in a digitally joined-up world: Introduction

Rolande Depoortere, Tom Gheldof, Dorien Styven, Johan van Der Eycken

INTRODUCTION

Rolande DEPOORTERE, Tom GHELDOF,
Dorien STYVEN, Johan VAN DER EYCKEN

Contemporary researchers are beginning to explore the possibilities and opportunities of digital humanities, but encounter major obstacles regarding (meta)data[1]. Many archival institutions lack the technology or the skills to process (meta)data, let alone share it. Different types of standards coexist and sometimes conflict with each other, while implementing the same standards often leads to slight differences which decrease interoperability. Storing, securing and making born-digital or digitized information available in a sustainable manner is a major challenge. Concepts such as metadata, Meta Information, Linked Open Data (LOD), Big Data… are on the rise, but their meaning and content — let alone their implications in terms of social impact — are seldom questioned. Archival institutions thus face a myriad of challenges when sharing (meta)data with the scientific community and when planning to preserve information for future generations while maintaining data authenticity[2].

In order to tackle these challenges, the DARIAH EU *Working Group Sustainable Publishing of Metadata*[3] organized a workshop in collaboration with the State

---

[1] V. Vanden Daelen, J. Edmond, P. Links, M. Priddy, L. Reijnhoudt, et al., "Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives. Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives", Conference Paper, December 2015, Brussels, Belgium, 2016; https://www.kbr.be /en/events/colloquium-inside-the-users-mind (conference proceedings, to be published).

[2] R. Speck, P. Links, "The Missing Voice: Archivists and Infrastructures for Humanities Research", in *International Journal of Humanities and Arts Computing*, 7. 1-2 (2013); J. Wettlaufer, W.L. Sina, "Digital Humanities", in *Der Archivar*, 67. Heft 3 (2004), p. 270-277.

[3] https://www.dariah.eu/activities/working-groups/sustainable-publishing-of-archival-catalogues/; The working group was set up in 2016 and thrives on a network of institutions initially associated with the European Holocaust Research Infrastructure (EHRI). The following were among the founding members of the working group: State Archives (BE) and CEGESOMA (BE), Kazerne Dossin (BE), International Institute for Social History (NL), DANS-KNAW (NL), NIOD-KNAW (NL), Trinity College Dublin (IE), Digital Curation Unit (GR), National Institute for Research in Computer Science and Control (FR) and Cyprus University of Technology (CY). At the last DARIAH-EU meeting in Paris on 23 and 24 May 2018, the working group was joined by KU Leuven (Humanities Faculty), the members of DARIAH-FED and a number of major foreign institutions such as The Austrian Centre for Digital Humanities (ACDH-OeAW) (AT), University of Vienna (AT), Archeovision (Université-Bordeaux-Montaigne) (FR), Consortium MASA (Mémoires des Archéologues et des Sites Archéologiques) (FR), and Centre national de la Recherche scientifique (FR).

Archives[4] of Belgium, Kazerne Dossin[5], KU Leuven and the FWO[6] on the 14th and 15th of May 2018. The objectives of this workshop and the current publication are to address the various aspects of these challenges, to provide a state of affairs with the recent developments in this respect at international level, to share this information regarding sustainability and interoperability and to work collaborative on solutions for common problems. In this respect, this publication can be considered as a first step toward the creation of a lasting bond between institutions and research infrastructures that play a key role in this research domain. The various members of this working group are closely involved in one or more European research infrastructure projects such as *Archives Portal Europe* (APEF)[7], *European Holocaust Research Infrastructure* (EHRI)[8], *European Research Infrastructure for Language Resources and Technology* (CLARIN)[9], *Digital Research Infrastructure for the Arts and Humanities* (DARIAH)[10] and *Europeana*[11].

This enables us to concentrate know-how, tools and expertise from all of these projects in our working group and to develop common solutions. This involves the alignment of (especially archival) standards currently in use and the development of common tools, a continuous exchange and provision of knowledge and expertise and the creation of synergies in new developments.

In order to deal with the different challenges collection holding institutions and researchers are confronted with and to formulate mutual answers to common problems and challenges posed by modern society concerning sustainable and interoperable access to data, the presentations of the workshop and articles in this publication are divided into three themes with metadata as a central approach:

1. Metadata, a path to standardization
2. Metadata, a link to the world
3. Metadata, communication and interoperability

Metadata, a path to standardization

The first theme deals with the necessity of using standards for interoperability and the easy distribution of quality information. The use of standards is after all essential for the interoperability and exchange of data between collection holding institutions, research infrastructures and users. In an ideal world, it should be possible to exchange information between local, national and international research infrastructures with limited investment. Although important steps have

4 www.arch.be.
5 https://www.kazernedossin.eu.
6 https://www.fwo.be.
7 https://www.archivesportaleurope.net/; http://www.archivesportaleuropefoundation.eu.
8 https://ehri-project.eu.
9 https://www.clarin.eu.
10 https://www.dariah.eu.
11 https://www.europeana.eu.

already been taken, we are still far from this ideal scenario. General practice has shown that standardization is difficult. There are several reasons for this, both technically and practically.

The archiving sector was already aware of the importance of standards before it entered the digital era. In 1992 the *International Council on Archives* (ICA) presented the first version of *General International Standard Archival Description* (ISAD(G)). This provided the impetus for a whole series of new standards: 1995: *International Standard Archival Authority Record for Corporate Bodies, Persons, and Families* (ISAAR(CPF), 2008: *International Standard for Describing Institutions with Archival Holdings* (ISDIAH), 2008: *International Standard for Describing Function*s (ISDF) for the respective description of producers, collection holding institutions and functions. The last of the series is the standard *Records in Context* (RiC) which was published by the ICA in 2016 and lays the foundation for a new conceptual model for archive descriptions[12]. The practical implementation of *Records in Context* will be discussed in the contribution of Anamaria Lopez[13].

The computerization in the 90s created the need for a digital version of these 'analogue' standards. The first version *Encoded Archival Description*-XML (EAD) developed by the University of Berkeley was ready for use in 1998. A second version generally known as EAD2002 was published in 2002. Although a third generation is also available (EAD3) EAD2002 remains the most widely used variant[14]. Analogous to ISAAR and ISDIAH, XML digital standards were also designed, EAC-CPF[15] and EAG[16], respectively.

Through this history, the first problems associated with the use of standards are immediately clear. Evolution is evident in both the digital variant and the analogue version. As a result, differences in speed with which they are implemented have

---

[12] More information about these standards and their evolution can be found on the website of the ICA: www.ica.org.

[13] Cfr. p. 47.

[14] For more information please check the following websites: the official homepage (www.loc.gov/ead), the documentation provided by the Society of American Archivists (https://www2.archivists.org/groups/technical-subcommittee-on-encoded-archival-description-ead/encoded-archival-description-ead) and the information presented by the Dutch National Archives (https://web.archive.org/web/20130329171117/; http://www.nationaalarchief.nl/openbaarheid-toegankelijkheid/publieksbeleid-website/coderen-uitwisselen).

[15] The first version of this standard was developed in 2001. A second edition followed in 2003. More information can be found on the official homepage (https://eac.staatsbibliothek-berlin.de/), the documentation provided by the Society of American Archivists (https://www2.archivists.org/standards/encoded-archival-context-corporate-bodies-persons-and-families-eac-cpf) and the information presented by the Dutch National Archives https://web.archive.org/web/20130329171117/http://www.nationaalarchief.nl/openbaarheid-toegankelijkheid/publieksbeleid-website/coderen-uitwisselen).

[16] http://wiki.archivesportaleurope.net/index.php/EAG2012.

immediate consequences for communication and interoperability. A second difficulty is that there is a certain margin for interpretation as to how a standard is formed. As a result, two standards created according to the same principles are not always identical. To make this problem even worse it is worth noting that key players in the field, who are all faced with similar or identical problems, come up with differing, sometimes conflicting solutions. A good example of this is the development of Ape-EAD[17] and EHRI-EAD[18]. Both are simplified versions of EAD2002 respectively developed by APEnet (2009-2012)[19] and APEx (2012-2015)[20] on the one hand and EHRI (2017). Both simplified versions were intended to reduce the interpretation margins and thus allow better interoperability. Were it not for the fact that both standards developed by these key players differ from each other. As a result, data from both research infrastructures cannot simply be exchanged with each other. Tools such as the EHRI-EAD creation tool, which are publicly available and would provide a solution to APE users, need to be rewritten for universal usage. The practical application of the EHRI-tool is discussed through the lecture by Charlotte Hauwaert and Francesco Gelati[21].

The moment the EHRI network started with the spread of EHRI-EAD, APEF was already betting on EAD3 in order to enable the ingestion of so called 'additional finding aids'[22]. The technical team of APEF chose EAD3 as a solution in this way for the wealth of additional information available in addition to the brief archival description offered by EAD2002. In theory information available in a variety of formats, handwritten texts, card systems, databases, … can be included and made available to the general public. However, the practical implementation of EAD3 does not seem that simple, since on the one hand content providers must support EAD3 according to the model developed by APEF and on the other hand there still is a long way to go before the technology is ready[23]. Moreover, it seems that the objectives that APEF wished to achieve through the introduction of EAD3, can also be achieved in another way, e.g., by using the RiC-model, used among others by the UK National Archives and the Spanish State Archives[24] and the introduction of *LOD* in the archival sector, as Ettore Rizza, Anne Charodennens and Seth van Hooland proved with their contribution[25].

---

[17] http://www.apex-project.eu/index.php/en/outcomes/standards/apeead; http://wiki.archivesportaleurope.net/index.php/apeEAD.
[18] https://ehri-project.eu/ehri-for-institutions.
[19] http://www.apenet.eu.
[20] http://www.apex-project.eu/index.php/en.
[21] Cfr. p. 15.
[22] http://wiki.archivesportaleurope.net/index.php/EAD3.
[23] During the workshop in May 2018, Wim van Dongen presented the recent developments regarding EAD3 in the APEF consortium. However, due to circumstances he could not devote an article to the subject. We try to close this gap by paying more attention to these developments in the conclusions.
[24] Cfr. p. 47, 103.
[25] Cfr. p. 37.

So far, we have focused on archival standards. This brings us to a whole new problem: just about every discipline in the humanities uses its own standards. The MARC-format or *MAchine-Readable Cataloging format* is mainly used by the library-sector, the social sciences are familiar with *Data Documentation Initiative* (DDI), and so on. From the archivists' point of view there are two generally accepted methods to store and to make available this standardized data from the various scientific disciplines. The first option is to treat these files as digital objects. As a result, one should not consider the standard used, file formats and integrated metadata. Just like analogue archives, the digital objects can be described and made accessible via EAD. A practical application of this strategy is used for example in the context of the PROMISE project, which aims to develop a strategy for the preservation of the Belgian Web[26]. The harvested data will be stored using the *Web Archive* (WARC)-format[27], which functions as a container for digital objects[28].

A second promising strategy consists of mapping the different standards. This method allows to recover the metadata incorporated in one standard for reuse, which once automated promises to have a significant potential. In order to create a functional model a lot of time has to be invested into the analysis of the different standards and in the creation of technical infrastructure. The possibility and practical feasibility of mapping was investigated by Benjamin Peuch in the context of the SODA project[29].

Metadata, a link to the world

Data and metadata are an important part of research infrastructures in particular and of everyday life in general, although perhaps not all users are aware of the digital footprints they leave on a daily basis. The second theme addressed at the workshop in Brussels therefore focused on the role that metadata plays or can potentially play in modern society. Phenomena such as Big data, open data, etc. are steadily growing to become or can be already looked upon as the most important challenges in the field of digital humanities today. Datasets used by consumers in everyday life are growing exponentially in size and questions for open access are becoming more and more stringent. The rise of these types of data also has an effect on how cultural heritage and collection holding institutes – be it libraries, museums or archives – deal with their data and metadata, both on a technical and an organizational level. The presentations in this section therefore focused on possible solutions for the challenges digital humanities are currently faced with regarding users and research strategies. Participants presented their specific and innovating projects dealing with these challenges that are overall present in society today.

---

[26] https://www.kbr.be/en/promise-project.
[27] https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.
[28] Cfr. p. 63.
[29] Cfr. p. 23.

In a world where users heavily rely on search engines and digital reading rooms to find the items they are looking for, one of the biggest challenges is locating, identifying and accessing the huge amount of hidden data stored away in collection holding institutes. The digital age has given birth to users who often don't exactly know where to start looking when an item can't be retrieved immediately in a digital manner. Many archival institutes, however, hold huge collections of documents and photos which are not available online. Those that are digitally accessible are only the top of the iceberg. Many institutes don't have online descriptions on the collection level for all their assets, only for a small percentage. In his presentation Mike Priddy addressed the challenges of this hidden data. In his view strategies regarding knowledge complexity and making hidden data visible should be developed as a joint venture by the collection holding institutes. Such approaches will not only allow researchers to work more effectively, but will also diminish the workload of the staff members working at the institutes. Unfortunately, Mike Priddy's contribution could not be included in this publication. However, his call to improve visibility and make hidden data findable with respect for the complexity of knowledge needed to contextualize the items is more relevant than ever.

Apart from making their hidden data visible to users, collection holding institutes are also faced with another big challenge when providing researchers with information: web archiving. The world wide web changes constantly, almost as if it were a living organism. Online publications are therefore very flux, changing regularly overtime, erasing or changing information available on webpages that once were used as sources for research data and information. Once a content manager changes texts or structure, the old version of the item – in this case the webpage – is lost for users. How then can institutes preserve websites for future research and reference? How can they store and curate different versions of webpages? How to make these available to the public so the references in published research are still valid? And most importantly, how can we capture the metadata of these webpages and store them together with the pages for future use? Libraries and archives worldwide are today looking for procedures to tackle this challenge. The strategies developed by the Royal Library and the State Archives of Belgium during the PROMISE project are discussed in the contribution of Emanuel di Pretero, Friedel Geeraert and Sébastien Soyez.[30]

Today many transaction in society are organized in a digital manner. Belgium became a pioneer in digital identification by introducing a smartcard instead of a paper ID in 2004. Other countries quickly followed and the number of trust services as well as the European online market grew steadily. In 2014 the European Union installed the e-IDAS regulation in order to harmonise the use of electronic identification and other trust services. The main question to be

[30] Cfr. p. 63.

addressed were: How can the trust of citizens in electronic transactions and services be enlarged? And how can the integrity, authenticity and readability of digitised or born-digital documents be guaranteed? Unfortunately, the e-IDAS regulation did not address digital archiving as an element of digital services. However, due to different privacy regulations in the member states, every member state addressed the implication of the e-IDAS regulation differently. In his contribution, Sébastien Soyez focused on the translation of the European regulation to Belgian law, known as the Digital Act, and on solutions for the void in the e-IDAS regulation regarding digital archiving[31].

Digital humanities are increasingly implemented in other fields of research. Initiatives are being taken to promote multi- and interdisciplinary approaches, especially within the social sciences and humanities (SSH). The European Cooperation in Science and Technology (COST) for example launched ENRESSH to accord social sciences and humanities a more central role in the scientific spectrum. Challenges addressed in this field of research include, among others, open research data, scientific and societal interactions in the different disciplines and patterns of dissemination in order to support evidence-based policy making and evaluation processes. In their case study, Marc Vanholsbeeck, Tim Engels and Andreja Istenic Starcic address the struggle which social sciences and humanities experience today regarding questions on data publication and data citation as markers for open research data.[32] The authors focus on the challenges of citation practices as well as on data provision, data sharing and data policies and guidelines in scientific journals in the field of social sciences and humanities, taking also into account stakeholders such as the European policy makers, researchers and publishers in order to develop a strategy regarding guidelines towards open data policies.

Last but not least, over the last years both researchers and cultural heritage institutions have been requesting more and more urgently a standardized and sustainable access policy to cultural heritage data for digital research in both social sciences and humanities. The cultural sector envisioned a tool thanks to which principles and mechanisms to use and re-use cultural data could be communicated clearly and uniformly to researchers, but which would also allow them to in fact reuse data which until today remains difficult due to copyrights and other legal issues. Recently this vision took the form of a Cultural Heritage Data Reuse Charter. In her presentation Sally Chambers focused on the development of the Charter, the ways in which this charter will be rolled out in the cultural heritage sector and the challenges met during the process. Although Chambers'

[31] Cfr. p. 63.
[32] Cfr. p. 83.

contribution could unfortunately not be included in this publication, more information regarding the project is available via the DARIAH website[33].

Metadata, communication and interoperability

The third and final theme of the workshop focused on the challenges of communication and interoperability from a user's perspective. Interoperability between metadata providers, platforms and standards was already outlined as a major challenge and (technical) solutions and methods to overcome this challenge were offered and discussed throughout the previous workshop sessions. From a researchers' point of view, communication of metadata, alongside interoperability and reusability, mostly applicable to findability and accessibility, is of even great importance to facilitate the user's work. Hence, advanced search strategies and (semi-) automatic enrichment of metadata are two main strategies used by cultural heritage institutions to disseminate metadata and make data sets better discoverable.

Cultural heritage institutions mostly approach the discoverability of their resources by collecting associated metadata and descriptive records. Aggregating these distributed resources, e.g., by using different aggregation technologies, has been a solution to problems of both discoverability and interoperability. Projects like Europeana Photography have invested in sustainable aggregation initiatives such as MINT to make their structured (meta)data more widely available for reuse in digital environments. Another way of achieving reuse is to provide content to initiatives such as the Archives Portal Europe (APE)[34], like Jane Stevenson explained in her description of the transformation process and the creation of a new automated workflow for the UK Archives Hub.

The UK National Archives describes its approach to digital description in a position paper, written by Jone Garmendia. She introduced us during the workshop in Brussels to the problems and challenges that born-digital records (especially those of the second generation) pose to archival institutes and their respective researchers. One of the proposed ways of dealing with this is using metadata to develop a different style of archival digital description. For instance, with the so-called contextual description we could attempt to derive any of this new contextual metadata from external (and non-archival) sources such as DBpedia. Annelies Van Nispen from EHRI/NIOD Institute for War, Holocaust and Genocide Studies, elaborated on how external sources could be used to enrich or to better retrieve archival records. She demonstrated how experiments with using controlled vocabularies such as EAD, Wikidata, Geonames, VIAF to enrich the existing EHRI Vocabularies proved (un)successful and how Linked Open

[33] https://www.dariah.eu/tools-services/data-re-use/
[34] https://www.archivesportaleurope.net

Holocaust Data could serve as an efficient cataloguing and integration tool for the end users of the EHRI portal.

Eric De Ruijter from the International Institute of Social History (IISH) focused on large-scale and (semi-)automated description of metadata methods. First, linked data (also linked to external vocabularies) served as a way to improve the IISH website by making its data more re-usable and interoperable. Secondly, collaboration between heritage institutions and with (digital humanities) students and researchers allows for more large-scale research tools and methods to be used, not only on the level of the IISH repository but also beyond. Similarly, Fred Truyen, president of the Photoconsortium, the aggregator of Photography for Europeana, advocated during his talk the need of developing advanced search strategies to improve the accessibility, searchability and findability of Europeana records for photography. Automated tagging of images is then one of the proposed solutions, but curation and photographic expertise are equally important in developing an enhanced search experience.

With the contributions in this book, we wish to provide an overview of the problems that both researchers and collection holding institutions currently face on three different but related themes. This publication serves as a first step towards a rapprochement of mayor players in the field. It is extremely important that we constantly inform each other of the difficulties we all have to face, the way we deal with them and which solutions are being proposed. In this context, we wish you a lot of reading pleasure.