# Trust and Understanding. The value of metadata in a digitally joined-up world: Conclusion, a Vision for the Future

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

# CONCLUSION, A VISION FOR THE FUTURE

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

New technologies offer researchers possibilities that not long ago existed only in fiction. Social phenomena such as the open data movement and the newfound availability of large amounts of data (so-called "Big Data"), made possible by automation, present researchers with new opportunities. There has never been so much data and metadata available as today, but so far we have failed to make the most of these new opportunities. There is no unique explanation for this phenomenon, let alone an unambiguous easy-to-implement solution. Last year's *Trust and understanding*-workshop offered us the opportunity to bring together international expertise from different international networks (DARIAH, EHRI, APEF, EUROPEANA, etc.) about metadata in all its aspects with the aim of finding a solution to this question or at least initiating it. This leads us to the unavoidable but also necessary question: to what extent have we succeeded in this approach?

To be able to answer this question objectively, it is best to start with a number of general assessments which we previously suspected to be true but which are clearly expressed through this workshop and the accompanying publication. The first and possibly the most important is that we all face the same difficulties seen from our own research background and perspective. The second conclusion that can be drawn, is that we are all working on or thinking about solutions for these same problems. At first glance, this does not seem to be an exciting novelty. Nothing is less true since every scientific research, every form of progress starts with assessing what we don't know and identifying the problems. Only in this way can we tread new paths and achieve progress. This means that we can only answer positively to the question "*Have we succeeded*?". The resources the community invested in this project were not spent in vain, but there is still a lot to do to reach the finish line. Based on the three main themes evocated, we wish to define the way to achieve this goal.

**Metadata, a path to standardization**

The first articles were dedicated to specific challenges regarding sustainable archiving and availability of digital information, such as the usage and implementation of international standards. In the introduction we noted the existence of multiple and ever evolving standards and standardised file formats, but also the difficulties to implement them. It is interesting to note that – in an

ideal situation - if we could pull the achievements and resources of the different projects together, all the projects would be a few steps further.

One of the most important challenges facing the *country managers* of Archives Portal Europe is the availability of standardized EAD files. These are necessary to make the portal a success. We can see that only larger institutions have the resources and knowledge to create these files. Many smaller institutions such as municipal archives, private archives or institutions with a different objective such as museums, but who dispose of an important archive collection, fail to make their valuable collections available to the public, simply because they do not have a valid EAD. The European Holocaust Research Infrastructure dealt with the same problem to improve the EHRI Online Portal[1]. In order to tackle this problem two tools were developed of which the *EHRI Conversion Tool*. This tool allows to convert XML, JSON, XML-EAD1, CSV, TSV metadata into EAD 2002 format, by mapping, correcting and validating in accordance to the standard guidelines[2]. Such a tool would be an asset both for APEF and for many collection holding institutions.

EHRI chose to use the most commonly used variant of EAD: EAD 2002. The use of EAD3, which offers more possibilities, was abandoned because of technical difficulties and practical problems to implement the new standard. APEF on the other hand made the choice to implement EAD3 with the aim of providing extra depth and more specific results when conducting a search query online. EAD3 is suitable for tagging all sorts of additional finding aids describing events, persons, places and subjects in detail, such as notary records, birth, marriage and death records, records of courts of law, etc. The possibility to tag this information in a more semantic way, will make the step towards Linked Open Data easier[3]. If the implementation of EAD3 is properly addressed, APEF can play a pioneering role in Europe by establishing standards and uniform procedures. Ideally, other research infrastructures would support the realization of the European archives portal (APE). EHRI could call on this expertise to solve the problems they face and to augment the possibilities of their own portal.

What the contributions in this publication clearly demonstrate is that even big players in the field cannot rely only on past achievements. People must be constantly on the lookout for new developments, even if they come from unexpected sources. One of the examples is the work of Benjamin Peuch. He discovered that DDI-files contain partly the same information as EAD-files and that it is possible to extract the information contained in DDI and to export it to

[1] https://portal.ehri-project.eu.
[2] https://www.slideshare.net/petradrenth/intro-ehri-conversion-tool-82363264.
[3] http://www.archivesportaleuropefoundation.eu/index.php/news/37-apef-starts-on-implementing-ead3.

EAD[4]. Although work still needs to be done to automate this process, the question must be asked whether it is possible to modify existing tools such as the EHRI-EAD tool to achieve this goal. DDI would be added to the long list of metadata formats (XML, JSON, XML-EAD1, CSV, TSV) the tool already deals with.

Ettore Rizza, Anne Chardonnens and Seth van Hooland demonstrated that Linked Data principles offer opportunities to improve the consultation of research data. For example, in a portal such as Archives Portal Europe it is not yet possible to combine data from multiple countries in a meaningful way due to the absence of a unique identification system. For instance, the archives of Charles V, King of Spain (Carlos I) and German Emperor (1500-1558) which are located in Belgium are not automatically linked to those in Spain, France, and Germany. Linked Data seems promising and may offer a solution to this problem. On the other hand, there are still many hurdles to overcome. The experiment demonstrated that automatic extraction of information out of knowledge bases such as Wikidata, DBpedia, etc. delivered poor results. These cloud-databases do not use uniform data-schemes which complicates the process. The manual triplification process was promising but time-consuming and requires skills. A solution for the problems this project copes with can –at least partly – be found in archival science. The conceptual model *Records in context of NEDA CM* keeps all the existing standards and achievements intact but creates the possibility for a Linked Open Data-system[5], without archivists having to fundamentally change their way of working. This also guarantees a certain reliability of the metadata collected, as the basis was laid by work processes that have already proven their reliability, and are supplemented with automatic ingest via API's from other research institutions.

In practice, collection holding institutes and researchers are not sufficiently informed about the existence and potential of new technologies. Pan-European research infrastructures such as APEF, EHRI, EUROPEANA, etc. can play a role in this and are ideally placed to take the lead in these developments. Only in this way can uniform procedures and standardization be achieved. Other initiatives also deserve a greater role in this, as the *Standard Survival Kit* (SSK), which is an open tool that supports researchers in choosing standards and best practices. Developed around the idea of providing research scenarios, it establishes a low-barrier entry point to get an overview about the standards used in different research fields[6]. Of course, there is again the danger that these initiatives will run side by side, without mutual consultation. The workshop last year and this publication are there first step to look in the same direction and to promote 'cross-pollination'. Only in this way can a future be built together efficiently.

---

[4] Cfr. p. 23.
[5] Cfr. p. 49.
[6]https://www.dariah.eu/2019/01/25/standardization-survival-kit-workshop-1-2019-textual-data-scenarios; http; https://ssk.huma-num.fr/#; cfr. p. 57.[7] https://www.cessda.eu

**Metadata, a link to the world**

The second theme of the conference focused on the needs of researchers today and the answers to their questions which are currently being developed by numerous projects. Especially in this digital day and age the role of archives, libraries and research infrastructures remains, first and foremost, to collect and provide access to data then used by researchers to create knowledge. The way in which researchers request, find and use the data and metadata, however, is changing. The challenges to be addressed regarding this topic were set out in the introduction: If users can't find the information or collections relevant for their research, if sources such as web pages used by other researchers are no longer accessible to them too, if published research data is not sufficiently exploited or if the data collected by previous researchers can't be reused, what then would be the role of collection holding institutes today?

During the past few years the Dutch Data Archiving and Networked Services, DANS-KNIAW, has invested heavily in the development of methodological and analytical frameworks to analyse the way in which researchers move around in the digital sphere and to investigate how these users deal with challenges such as big data. The results of the K-PLEX project were presented in 2017. Continuing the research, DANS now focusses on cultural heritage institutes and their methods to increase online visibility – and thus findability – of collections, data and metadata within the context of knowledge complexity. Although the contribution regarding this topic unfortunately could not be added to this publication, presenter Mike Priddy's call for collaboration between institutes, e.g., via portals and research infrastructures such as EHRI or APE, in order to broaden the spectrum of visibility, and share and exchange data, metadata, information and knowledge for the benefit of users is an important message to the cultural heritage sector.

Another challenge when fine-tuning the interaction between institutes and the research community is to store and preserve sources used for existing research so this research can be retraced or verified in the future. A specific puzzling item to preserve are the constantly appearing, changing and disappearing web pages which are often used as research material in the social sciences and humanities. From a research and archival perspective web pages are no longer flux entities but real publications that need to be preserved in an equal manner as books and articles. The PROMISE project at the Royal Library of Belgium focusses on how to capture Belgian web pages and their metadata for future reference and research purposes on the long term. As illustrated in the contribution, several issues had to be addressed while creating a preservation strategy. During the selection phase, criteria as well as legislation regarding e.g., illegal content had to be developed. The capture and quality control phase as well as the preservation phase included issues such as copyrights, authenticity and integrity addressed via the implementation of the WARC file format as the standard. Taking into account the need for accessibility, a new type of storage facility needs to be developed which combines

standardized metadata guidelines and search options adapted to user needs and behaviour. Again, in this case, standardization is key.

Society today is oriented towards a digital environment. Citizens rely on electronic transactions and services to interact with their government and the economic markets. With the creation of the e-IDAS regulation introduced by the European Union in 2014, member states were obliged to subsequently create national laws on topics such as e-signatures, e-seals, e-time-stamps and e-delivery. However, the regulation missed out on including digital archiving in its list of trust services. Sébastien Soyez in his contribution illustrated how Belgium filled this void by in 2016 creating the national Digital Act as well as a completely new trust service called the Electronic Archiving Trust Service. This service will focus both on the digitisation of analogue sources and on the preservation of digitised or digital-born documents in both the private and public sector. By providing a list of mandatory metadata in the Digital Act, standardization of the trust services is achieved in Belgium. Apart from metadata Belgian law also stressed that authenticity, integrity and readability of the digitised or digital-born items should be guaranteed. With his case study, Soyez showed how the Electronic Archiving Trust Service thus created a general regulatory environment for digitisation and digital preservation.

Researchers in social sciences and humanities today apply a growing multidisciplinary approach and thus produce interdisciplinary data. However, not much research has been conducted about how these researchers themselves then transfer their data via publication or how their data is cited in publications, both books and journals. COST ENRESSH in 2017 therefore decided to focus on analysing the field of social sciences and humanities publications, in order to develop general policies regarding exchange of data and data evaluation processes. The case study presented at the conference dealt with data publication and citation as markers for open research data (ORD) and the role of European policy makers, researchers and publishers in creating ORD. The introduction of ORD as a part of Horizon Europe (2021-2027) by the European Commission can, among other initiatives, be seen as a positive sign regarding the stimulation of ORD in the future. There remains, however, room for researchers to increase the sharing of data among each other. Increasing the visibility of such datasets created by researchers might be a central role to play for publishers.

Last but not least, practical and legal issues were addressed regarding the reuse of data. Researchers can often benefit by applying new techniques to data collected by their predecessors or stored at archives and libraries. However, too often the process of obtaining a copy of the relevant data can be a long and tiresome procedure in which not many researchers wish to invest time and money. Supra-national research infrastructures have the power to facilitate interactions between the involved actors and would be an ideal setting to launch a common online environment: the Cultural Heritage Data Reuse Charter. A user could, via the

environment be informed about contacts regarding reuse, the access procedure, reuse conditions and citation model. Although the contribution of presenter Sally Chambers was not included in this publication, the message of the presentation stresses the need for streamlining procedures and promoting reuse as well as exchange.

Across all these projects presented at the workshop, one general line prevails. In order to assist researchers, make data available and reusable, standardization is key: a standard way to exchange data and metadata to make it findable, a standardized environment for web archiving, a standard reference method for data, a standard reuse procedure… However, standardization and collaboration within the cultural heritage sector have to be in accordance with broader social evolutions. From a broader perspective, it is imperative to map out in a consistent way the legal framework and its latest updates, e.g., the implementation of the European General Data Protection Regulation (GDPR) for cultural heritage (EU) as well as eIDAS which influence the opportunities and challenges met when combining archival and research needs. The accessibility and increased commercialization of (often publicly funded) data constitutes a new issue. There is a need for international consultation to address these problems.

**Metadata, communication and interoperability**

Making large amounts of data available to researchers but also to commercial companies and to citizens is a critical point of concern. Big data requires up-to-date research methods, frameworks and infrastructures. There is a need to develop new, cross-border and cross-language software solutions, such as the use of uniform thesauri, automatic translation and indexation. Jane Stevenson shows how it is essential to think about the potential of data and metadata before taking decisions and starting projects. Several authors underline the difficulty to standardize metadata inherited from previous practices and professional methodologies. The full automatization of (meta)data processing remains at this moment a dream as human intervention is still indispensable to manage and monitor the process, to (at least partly) control the quality of the metadata, and to collect missing metadata. Though, according to Jone Garmendia and Eric de Ruijter, this human contribution is not bearable with huge data collections and with born-digital archives.

That is why new ways to supply to human intervention are tested: Handwritten Text Recognition (HTR), Named Entity Recognition (NER), Linked Open Data (LOD), automatic image/photograph/object annotation, tools for geolocalisation, probabilistic description … According to the experience of EHRI, results are sometimes disappointing as existing thesauri are not always adapted to the needs of historical description. There is still a gap to fill in the development of multilingual databases for historical geolocalisation and identification of persons.

Issues as unique and persistent identifiers and disambiguation of content are key factors of success in order to guarantee the interoperability of systems and platforms. But archivists must also change their paradigms: till now metadata were generally produced by specialists for specialized categories of users. Photo collections (not only in Europeana) teach us that we have to think about a larger public, anticipate new reuses of the (meta)data and reconsider our description practices. Frederik Truyen emphasizes the necessity to add metadata responding to users needs like metadata on photographical techniques and attributes. We must also rethink the channels for the dissemination of the metadata and evaluate the advantages of collaborate with other platforms and generic search engines. The National Archives of Great-Britain explores original automatized methodologies in order to facilitate the contextual description of records and to face the problems that are put by temporally aware descriptions.

**Between cooperation and competition**

One of the objectives of research infrastructures in general and DARIAH in particular is to share knowledge and know how to avoid that the wheel must be invented twice. No matter how noble this goal may be, putting it into practice is less evident. Resources and staff are a scarce commodity and are distributed between different research infrastructures at European level. There is also some competition at national level. Research Infrastructures have to be recognized and have to share their scarce resources. In several countries, initiatives have already been taken to address this, either by merging infrastructures at the national level, e.g. *Common Lab Research Infrastructure for the Arts and Humanities* (CLARIAH) which is a cooperation between DARIAH and Common Language Resources and Technology Infrastructure (CLARIN), or by taking smaller, targeted initiatives at project level. In Belgium, the *Belgian Science Policy Office* (BELSPO) funds the *Belgian Infrastructure for Social Sciences and Humanities Open Science* (BISHOPS) project, which aims to create a science cloud for the *Consortium of European Social Science Data Archives* (CESSDA)[7].

Such cooperation not only provides economies of scale but also prevents the fragmentation of scarce resources. However, cooperation is not as obvious as it seems. Even at national level, one must also take into account phenomena such as regionalization and the distribution of powers, which can lead to different political choices and different priorities, with major implications for the financing of joint projects. This argument did not yet take into account competition between the various scientific institutions for the acquisition of funds for their own projects. All this makes the organization of collaboration within research infrastructures a political fact. People need to become aware of the usefulness of collaboration and of sharing knowledge and know-how, without losing sight of the individual interest or the interest of one's own institution.

[7] https://www.cessda.eu

This publication and the workshop were a first step to overcome these differences. We brought together - as already mentioned - the major players in the field, to make them aware of their respective projects and to stimulate 'cross-pollination' and cooperation. We showed clearly through concrete examples that we all can learn from our respective projects and that collaboration works. In Belgium, science policy is a competence divided between the language communities and the federal government, represented respectively by DARIAH-Flanders, DARIAH-Wallonia and DARIAH-FED. Each community has its own priorities, its own working methods and depends on its own financing. This makes cooperation within the DARIAH-BE consortium particularly difficult. Collaboration between the federal partner and one of the communities is feasible, but getting the entire consortium in line or having one project carried out together is extremely difficult. However, this does not mean that there is no will to cooperate. This project is the first project with which we have overcome these difficulties, by involving all partners. We have found and created willingness to overcome these challenges and to take steps together in the right direction.

**Beyond the workshop**

As stated above, the workshop last year was only a first step. At a certain moment, however, it is necessary to exceed the level of a workshop and online meetings. In other words, there is a need for a concrete project to perpetuate the steps taken. For this reason, when the call was submitted in 2017, it was decided to perpetuate the results of the workshop, available online – in the cloud - for the entire world as well as on paper as a symbol of durability, via the Belgian learned journal "Archive and Library", which has an international distribution.

We promised the participants of the Workshop and the members of the DARIAH-EU Working Group Sustainable Publishing of Metadata to continue the work already done and to put the achievements into practice. Our objective is to further develop the various topics that were discussed, in more detail and in a more practical way. This makes it easier to achieve visible results in the form of concrete applications and realistic projects. This can be done, for example, through the organization of a hackathon, the joint testing of tools, or the development of common solutions. In that regard, there are plans to test various AI applications and their possibilities. At the moment, unfortunately, the financial means are still missing.

Nevertheless, we keep looking for opportunities to continue the good work. We participated with the *French national research institute for digital sciences* (INRIA) in the SSK-project, which was the subject of the DARIAH Theme Funding Call 2018[8]

[8]https://www.dariah.eu/2018/11/05/dariah-theme-funding-call-2018-2019-meet-the-winning-projects/).

The same partners are also involved in the already mentioned BISHOPS-project. And other interesting projects may soon appear.

A direct result of the workshop was the expansion of the working group by adding new members from the museum sector (KIK-IRPA (http://www.kikirpa.be/), BOZAR (https://www.fine-arts-museum.be/nl), KU Leuven (http://www.kuleuven.be). The official introduction of these institutions took place during the DARIAH EU meeting in Paris. Klaus Illmayer, (OEAW-ACDH), Sara Tiefenbacher (project researcher, University of Vienna), Olivier Marlet (Consortium MASA) and Emmanuelle Morlock (CNRS) also joined the working group after the evaluation of the workshop at the DARIAH-EU meeting. This way, we achieved a positive impact on the operational aspect of DARIAH-EU.

**Conclusion**
This book shows that through collaborative work, it is possible to pool solutions and to establish relationships of cooperation, both at the level of practical tool development and with regard to sharing and creating knowledge and know-how. To establish concrete results however, substantial effort is required both at human level and practical level. There is a need to constantly think and look ahead despite sporadic setbacks. This is how we can achieve our objectives.

**About the 'invisible' collaborators**
Finally, we would like to thank everyone who made the workshop in 2018 and this publication possible, in particular the silent 'unnamed' employees who were doing their work diligently.

We would also like to add the biographies of people who have contributed but who could not make a 'visible' contribution to this publication due to various circumstances.

Mike Priddy is a Senior Information Systems Engineer at Data Archiving and Networked Services (DANS), an institute of the KNAW & NWO. Mike works across the fields of Social Sciences and Humanities on a range of European research infrastructures and development projects, specialising in architectural, process and quality modelling as well as project management. He has been involved in specifying and creating research infrastructures since 2005, including in that time, Digital Research Infrastructure for the Arts and Humanities (DARIAH preparatory phase), Data Services Infrastructure for the Social Sciences and Humanities (DASISH), Data without Boundaries (DwB), European Holocaust Research Infrastructure (EHRI phases 1 & 2), CESSDA Strengthening and Widening (CESSDA-SaW), Humanities at Scale (HaS-DARIAH), CESSDA (Consortium of European Social Science Data Archives) and Knowledge Complexity (K-PLEX). Mike has an academic background in computer science and visual communication.

Charlotte Hauwaert is based at the Belgian State Archives/CegeSoma (Centre for Historical Research and Documentation on War and Society) in Brussels and works for the European Holocaust Research Infrastructure (EHRI). She holds a master's degree in History (Universiteit Gent), in European Literatures and Cultures (Université Lille 3- Charles de Gaulle) and in Literary Studies (KU Leuven). Charlotte joined the EHRI project in September 2016, just after finishing her Advanced Master's program.

Wim Van Dongen graduated in History at Radboud University Nijmegen, developed ICT skills as an employee of a software development company and works for the Dutch National Archives since 2003. He was involved in several archive-related software developing projects in the Netherlands before he got involved in the Archives Portal Europe project. In the APEnet project he was Technical Coordinator and, as WP3 leader, responsible for the interoperability between Archives Portal Europe and Europeana. Within the APEx project he acted as Country Manager Coordinator.

Veerle Vanden Daelen is deputy general director and curator at Kazerne Dossin - Memorial, Museum and Documentation Centre on Holocaust and Human Rights, and holds a PhD from the University of Antwerp. Her doctoral research focused on the return of Jews and reconstruction of their life in Antwerp after the Second World War (1944-1960). She held fellowships at the University of Michigan (Frankel Institute for Advanced Judaic Studies, 2007-2008 fellowship "Jews and the City") and at the University of Pennsylvania (Herbert D. Katz Centre for Advanced Judaic Studies, 2008-2009 fellowship "Jews, Commerce, and Culture"). She is coordinator of the work package "Data Identification and Integration" for the project European Holocaust Research Infrastructure (EHRI). She is affiliated with the University of Antwerp, where she has taught "Migration History", "Jewish History" and other classes and where she organizes, together with Karin Hofmeester, the annual "Contact Day Jewish Studies on the Low Countries" at the Institute of Jewish Studies. Veerle is also a member of the Belgian delegation to the International Holocaust Remembrance Alliance (IHRA).

Sally Chambers is Digital Humanities Research Coordinator at Ghent Centre for Digital Humanities, Ghent University, where she coordinates Flemish and Belgian participation in DARIAH, the Digital Research Infrastructure for the Arts and Humanities. She is Vice-Chair of the DARIAH-EU National Coordinator Committee and member of the Senior Management Team. From 2011-2015, Sally was Secretary-General of DARIAH-EU, based in the Göttingen Centre for Digital Humanities, Germany. Previously Sally worked for The European Library, based in the National Library of the Netherlands, focusing on interoperability, metadata and technical project coordination. She has a first degree in Literature with Psychology and postgraduate qualifications in Cultural Studies and Information Services Management.