

Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives

Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, Václav Tollar, Annelies van Nispen

▶ **To cite this version:**

Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda 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", Dec 2015, Brussels, Belgium. hal-01281442v1

HAL Id: hal-01281442

https://inria.hal.science/hal-01281442v1

Submitted on 2 Mar 2016 (v1), last revised 10 Mar 2016 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives

Veerle Vanden Daelen (main author), Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, Václav Tollar and Annelies van Nispen[1]

Foreword (Veerle Vanden Daelen)

Since I began working on the data integration for the European Holocaust Research Infrastructure (EHRI) in 2011, I have worked with a wide variety of twentieth-century history archives, from classic archives to memorial sites, libraries and private archives. Integrating the descriptions – when they were already available – for the Holocaust sources held at these various institutions has proven to be a major challenge. Along the way, my colleagues and I have met with many highly dedicated people – from professional archivists to volunteers – who are preserving and describing the sources within the limits of their available time, infrastructures and staff. None of the archives EHRI encountered could export their archival descriptions without some degree of limited to extensive pre-processing or other preparatory work. Similar experiences were noted by the Collaborative EuropeaN Digital Archival Research Infrastructure (CENDARI). Hence, we initiated a proposal for the DARIAH call Open Humanities 2015, entitled: "Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives". The core group for this project consisted of Jennifer Edmond (Trinity College Dublin), Petra Links (NIOD Institute for War, Holocaust and Genocide Studies, Amsterdam), Mike Priddy (Data Archiving and Networked Services, DANS, The Hague), Linda Reijnhoudt (DANS, The Hague), Václav Tollar (Institute for the Study of Totalitarian Regimes and Security Services Archive, USTR and ABS, Prague), Annelies van Nispen (NIOD, Amsterdam) and myself (Centre for Historical Research and Documentation on War and Contemporary Society – CEGESOMA, Brussels).

The main goal of the "Open History" project was to enhance the dialogue between (meta-)data providers and research infrastructures. To this end we have been working on two tools: an easily accessible and general article on why the practice of standardization and sharing is important and how this can be achieved; and a model which provides checklists for self-analyses of archival institutions. The text that follows is the article we have developed. It intentionally remains at a quite general level, without much jargon, so that it can be easily read by those who are non-archivists or non-IT. Hence, we hope it will be easy to understand for both those who are describing the sources at various archives (with or without IT or archival sciences degrees), as well as decision-makers (directors and advisory boards) who wish to understand the benefits of investing in standardization and sharing of data.

[1] We would like to thank wholeheartedly all the participants of the DARIAH workshop at CEGESOMA on 9-10 December 2015.

It is important to note is that this text is a first step, not a static, final result. Not all aspects about standardization and publication of (meta-)data are discussed, nor are updates or feedback mechanisms for annotations and comments discussed. The idea is that this text can be used in full or in part and that it will receive further chapters and section updates as time goes by and as other communities begin using it. Some archives will read through much of these and see confirmation of what they have already been implementing; others – especially the smaller institutions, such as private memory institutions – will hopefully find this a low-key and hands-on introduction to help them in their efforts.

Framework

Introduction 3

The need for standardization in a digital world: Humans can live with (very) dirty data, computers cannot 3

No panic – nobody's perfect 4

What do research infrastructures like CENDARI and EHRI do? 5

Everybody's different (and tailor-made solutions are expensive and often not sustainable) 5

Understanding your organisation and its data management via the Capability Maturity Model 7

Stating the obvious? Providing your institution with its own unique ID 9

From the filing cabinets and printed catalogues to digital metadata 10

How to find the needle in the haystack: Describing your holdings 11

- Following archival standards 12
- Uniquely identifying your holdings 13
- Consider which levels of descriptions you want and need 14
- Standardizing the input 14
- Keeping the pieces together 15
- Publishing about your holdings on your website 16

"In our family we share" - export of information from your institution to the outside world 16

Bad news for the anarchists: the importance of documented and well-implemented guidelines and policies 18

Further reading? 19

Introduction

Generally, when a research infrastructure team contacts archival institutions, they are interested in digital data and/or digital metadata and receiving the (meta-)data in a specific format (specific software, following the project's guidelines, etc.). They ask the archival institution many questions concerning how its holdings are described and how to publish and share this information with researchers and the general public. Often, these questions are not easy to answer and require contacting multiple people within the institution. This can lead people at the archival institution to question why on earth their institution – which has its own established way of doing things – should now invest time and effort in doing things the way that the contacting team does. There are several very good reasons why this can be a key benefit to your institution; moreover, offering these large infrastructures what they are looking for certainly enhances institutions' visibility!

Before anyone can start sharing information in an effective way, standardization of information enters onto the agenda. Not only is standardization a prerequisite for efficient and sustainable sharing, it also brings in many benefits for your institution, as it will force your institution to make the work processes explicit and documented. The knowledge of how your holdings are described is no longer limited to one or a couple persons. This information will now be out there for everyone.

The need for standardization in a digital world: Humans can live with (very) dirty data, computers cannot

We've come a long way in recent decades in how computers and digital tools impact access to information and methods and tools for organizing and processing data. One still finds archives with cardboard indexes and stencilled finding aids, all of which can be tremendously useful for opening up the data being preserved. But these methods have gradually moved to a digital work environment. Pioneers, and maybe you are one of them, have invested much time and effort in developing custom-made solutions for the specific needs and requirements of their institutions, long before publicly and freely accessible tools were available. Maybe your institution looked for help from outside the archive and purchased software and tools from a vendor. You may have even published your finding aids on your website, only to find that when the research infrastructures come by that they cannot easily provide your information to their research community.

This problem arises because, even though the data and tools are easily read and interpreted by a human reader, the computers who need to integrate all the information cannot recognize and process it. Something described by your institution in its own way does not necessarily correspond to how another institution works. For example, when you read in a finding aid that certain materials are "in Duch and in Frenche", you perfectly understand this, despite the typos. If you are familiar with internally agreed-on abbreviations, you may equally understand that the languages are Dutch and French from the

abbreviations “D” and “F”. People who have little or no information about the internal codes for languages, however, would have trouble with the latter identifications, as they may be unsure whether the materials could just as well be in German (Deutsch) or Danish, Finnish or Flemish. A computer, though, is at loss in both cases, as it looks for standardized formats (for example, using standard language codes of NL and FR) with which it can organize information from different sources. Basically, computers make us standardize our working methods. And this is also what these research infrastructures knocking on our doors make us do. That’s a good thing; first, because it will allow the infrastructures to reach their goals; and second, because reflecting on and working towards standardizing your institution’s working methods and documenting internal work processes will afford you not only a much deeper understanding and overview of your own holdings (searching for all documents with the standard FR as a language code without missing the “Frenche” materials) but also further ways of preserving them and opening them to the outside world (for example, the materials in “Nederlands en Frans” or “Dutch and French” can be presented to the user depending upon their needs). If we believe in our mission to preserve our analogue and digital cultural heritage, we need to adapt to the new world and recognize what these new methods and tools offer us.

No panic – nobody’s perfect

While you may feel that you still have a long way to go, it is important to realize that, in general, most of us are still trying to find our way and a best-practice method that actually fits the institution, not only from a content perspective but also in terms of investment (including both financial and human resources). Archives which appear to be perfect to the outside world often still have elements that need improvement when looked at in a bit more detail. Having a long history of activity does not necessarily turn out to be advantageous. Quite often different data management systems have been added to existing ones, resulting in a unique but complex and difficult to disentangle system. They come with a history, with accreted knowledge and embedded practices and often face more difficulties to adjust to new digital opportunities. There is no “one-size-fits-all” method by which all the different archives out there can standardize their holdings’ descriptions. Similarly, not all archives need to reach the same level of sophistication and complexity in their tools and methods. However, what we all would like to do is to let the outside world know about us and what we do.

This analysis will share experiences from two pan-European research infrastructure projects who tried to merge information on twentieth-century collection-holding institutions and their holdings (on a top-level, such as collection, fonds, or record group, for example) into virtual research environments. Their experiences are instrumental in learning what the current status, or situation, in these archives actually is. How ready (or not) the institutions are to step into these projects is relevant to these research infrastructures themselves; moreover, and maybe more importantly in a long-term perspective, what they learned can now be translated to and used by the archives themselves.

What do research infrastructures like CENDARI and EHRI do?

The two projects who have merged their knowledge and experiences to write this article are CENDARI and EHRI. The acronyms stand, respectively, for the Collaborative EuropeaN Digital Archival Research Infrastructure and the European Holocaust Research Infrastructure. Both projects have both been funded by the European Commission under its program for Research and Innovation. Subject-wise both have a focus on twentieth-century history archives. In order to build a virtual research environment (a VRE), the projects wished to share information on archival holdings spread over a multitude of institutions across multiple countries. The aims of merging this information into one portal are multiple and can be summarized as follows: 1. Providing a centralized access point for the information; and 2. Enabling methodologies and tools for research that otherwise would not be possible. Hence, it is not about making the archival institutions superfluous; on the contrary, having their information included in the portals increases the visibility and knowledge about the participating institutions and their holdings. Research infrastructure projects should be viewed not as attempts to replace cultural heritage institutions, but rather as a very useful class of users, albeit with specific needs and on a grand scale. Moreover, by participating, the institutions become part of a community on medieval sources and sources on the First World War (in the case of CENDARI) or on sources on the Holocaust (in EHRI).

How did the infrastructures actually plan to integrate (meta-)data into these portals? CENDARI and EHRI each had an international consortium staff which included historians, archivists and IT/digital humanists. The work process involved all disciplines, to ensure that the right content came in the right format – right content meaning within the interest of the research topic (First World War for CENDARI, the Holocaust for EHRI), right format meaning in a structured manner (standards, see below). The work of these interdisciplinary teams seemed very straightforward in theory – contacting all the archives, receiving their metadata (the data describing the institution and its holdings) and putting them into the system. The reality, however, was quite different. Often the metadata were not available. And even if they were, the metadata were generally not following a standardized format valid beyond the institution itself. Reorganization and clean-up was necessary to make the metadata inter-operable and shareable. This does not mean that all metadata had to become completely uniform, but rather that the metadata, which institutions were going to share with the infrastructures, needed to be somehow mapped to a common denominator, so that they could be imported into this system.

Everybody's different (and tailor-made solutions are expensive and often not sustainable)

Getting there meant change, and change meant work. And as we all know, there is no archive – at least not to our knowledge – with a room where "reserve teams" are sitting and waiting for an assignment, much less such a room with a budget in it. The reality is that most institutions that are preserving our

cultural heritage and historical records are understaffed and underfunded. They are lucky to have a motivated staff doing their utmost to preserve the archives and open them up for research. What did CENDARI and EHRI then do to get the data into their portals? On CENDARI's website, it states that "integrating the institutional data with collection and item-level descriptions requires a custom workflow using XML tools and a customized underlying schema". Note the use of the words "custom" and "customized": they already give away that there was no universally applicable method. Both projects had to excel in tailor-made approaches. Among other things, CENDARI used tools to "scrape" websites, so that the information could be brought in as such. In EHRI, working methods ranged from fully manually entered descriptions to direct import. Most cases, however, were somewhere in-between and required thorough mapping, conversions and pre-processing to get the metadata into the portal, which is now online. And all this is not too terribly shocking and difficult if only the methods used would have been applicable from one institution to another. In 2011, in the first year of the EHRI project, the phrase "every archive is unique" was recited somewhat laughingly. By the end of the first phase of the project, four years later, the data ingest teams saying this phrase sounded much more desperate, as the number of custom-made approaches almost equalled the number of integrated institutions. The phrase "every archive is unique" is actually very true: each of the archival institutions has its own history, specific holdings and mission, and this has led to custom-made unique systems and ways of describing and opening up of the holdings. The same frustrations led CENDARI to create a flowchart for assessing how the project could work with any given institution, with the hope of ensuring that any possible reuse of process or tools could be easily identified and implemented within the archive liaison team.

Despite the wide variety among the archives, the good news is that both CENDARI and EHRI have managed to bring in content into the portals and in doing so, they already have revealed and exposed many "hidden archives" to the research community. The bad news, though, is that none of this is fully sustainable at this moment. Ideally, bringing in this information should be a repeatable process, without having to adjust working methods to each and every single case. Both the institutions themselves and the projects should be able to have easy access to the same data (at this moment the descriptions are not necessarily accessible both at the institutions and in the projects). Updates at one place are not directly reflected in the other place.

One institution EHRI worked with, for example, has all of its holdings in high-quality digital format available for internal use in the form of thousands of high-quality scans, but its website shows only basic information about its holdings. Thus these thousands of scans are not only hidden for research purposes from the outside world, but also to a high degree from the institution itself. The digitization process did not provide a solution for the "needle-in-the-haystack" problem, meaning there are still no standardized, easily accessible descriptions that identify where you can find what. If, for example, you would be looking for correspondence in private archives donated to this institution, the website would indicate only that private archives have been donated to the institution. No names or further detailed descriptions of the private archives are available on the institution's website. Moreover, even at the archive itself, there is no search tool available in the reading room. The researcher would have to ask the archivist, who then would – based on his or her expertise – know which folders to open, find a text

document with the descriptions and present the scans to further inform the researcher about what is in the requested source.

What is further frustrating is that, since the basic descriptions on this institution's website are written as a non-standardized flat text, they cannot be integrated as such into the research infrastructure's databases. An extra complicating factor is that the institution did not attribute stable and unique identifiers to its holdings: the identification numbers of the collections changed with every website update. All of this causes not only trouble for the research infrastructures, it also makes the organization of the (meta-)data in the institution chaotic and difficult to work with. And here is why the activity of the research infrastructures makes for an enhanced awareness and momentum within an institution to think about the following things, which we will frame more generally into a capability maturity model.

Understanding your organisation and its data management via the Capability Maturity Model

The term "Capability Maturity Model" (CMM) refers to a system developed in a study requested in the 1980s of the US Military. Since the use of computers had grown increasingly widespread since the 1960s, the military had begun to adopt computerized information systems. These systems developed and changed over time, and more and more new systems were launched and added. All these developments together created at best a more efficient and usable system, but they were equally accompanied by chaos and failure, lack of clarity on what was being managed in each system and how this was done. Having more tools and the fanciest software program clearly did not guarantee higher performance and better information management. On the contrary, the application of multiple programs, methods and tools that are insufficiently integrated within an organization results in chaos, and high costs for training and alignment to improve the situation. The CMM was developed as a tool to help organisations bring order into the chaos. It documents and structures the behaviour of an organisation and in doing so helps people understand the organization and its data management.

The CMM distinguishes five maturity levels. The first level – which is unfortunately the level most seen in twentieth-century history archives – is the initial level. Institutions at this level typically do not document their data management processes. The system is an "ad hoc" system. It is an uncontrolled way of working, which is under dynamic change. The problem is that this chaotic and unstable work environment makes it difficult, even for people who belong to the institution, to comprehend the data management system. For example: Sophie may be the person who describes the holdings into the database of the institution. As long as there is no document which explains how she does this work, none of Sophie's colleagues can replace her or work in completely a similar method. Sophie's institution would move to a level-2 process (the repeatable level) in the maturity model if she would put together a basic document with (and for) her colleagues, which at least documents the process in such way that Sophie's actions can be repeated. Even though this document would not be carved in stone, nor known by everyone in the institution, it would at least create a degree of documentation on the processes so that they become potentially repeatable. If the colleagues wish to achieve a level-3 process (defined

level), they would need further definition and standardization. The methods to describe the institution's holdings would need to be defined as a standard business process, referring to standards that are valid and consistent across the organisation. The next two steps, level 4 (managed) or level 5 (optimizing), bring the activities to a yet higher degree of capability and maturity. They respectively refer to quantitatively managed and the real best-practice, a level reached by very few worldwide.

What is important, though, is to become much more aware of the need to have transparent and reproducible workflows. These are in theory a requirement for any scholarly work, and for historical research should ideally also be applied on the input-side. Moreover, in digital ecosystems, it is a "conditio sine qua non" to enable communication. The proof that well-documented, transparent and reproducible workflows are absent at most institutions – hence leaving the core group sadly at the first, lowest level in the maturity model – became clear when CENDARI or EHRI wished to obtain more information about the software and archival standards used by archival institutions to describe their holdings. Even those institutions with many digital tools and utilization of archival standards had trouble answering the questions, as answers to these questions could not be found in a document and involved talking to different people from an institution. Often the persons describing the holdings were not archivists by training, and the databases had been bought from an external company without anyone in the institution having full permission (or knowledge) to handle them. Most institutions did not really have a clearly documented overview of their workflows, and engaging with a research infrastructure involved talking with at least three people: the one who could tell you "what" (was described and relevant for the portal), the one who could tell you "how" (software and standards or mapping) and the one who could tell you "yes" (the authority to give permission to integrate data from this archive into the portal).

Even in some of the most "chaotic" systems encountered by CENDARI and EHRI, an eventual solution was devised. And still, having an institution's data represented in digital research infrastructures by this sort of bespoke approach to reworking its data for a one-time import into the portals is useful for the infrastructures, but only of limited benefit for your institution. It becomes much more interesting and helpful when the institution engages in the dialogue and raises its self-knowledge and capacity level. So, therefore, we invite you to read along to see what possible help is out there and what benefits standardization and documented workflows can bring to your organization – each within its own possibilities and capacities, without needing to start from scratch or invest huge amounts of money or time in new systems. Mike Priddy and Linda Reijnhoudt and their colleagues at the Data Archiving and Networked Services (DANS, The Hague) introduced this CMM to our group and are working on developing this model and sharing a written check-list which introduces external reference points, illustrated with examples to involve you and your institution in improving your own capabilities to describe, publish and share your data.

Stating the obvious? Providing your institution with its own unique ID

A first point of attention concerns perhaps the most obvious, yet most frequently forgotten, aspects for putting your institution on the map, namely, describing your own organization. This is actually your institution's "highest access point": information about what type of archive you are, where you are situated and which collections you hold. But, if there was one task for which CENDARI and EHRI had to carry out much manual work, it was in describing the institutions according to a uniform and maximally complete format. In many cases, it involved a search in various places on an institution's website (if such was available) to collect official name(s), addresses and contact information, history, mission statements and information on holdings (such as references to online and analogue finding aids). The result of this is that others will make their own description of your institution and that a national archive, which is for example both described in CENDARI and EHRI, will be described in different ways. "Oh well, variations to a theme", one could say, but what is especially unfortunate is that computers cannot necessarily identify these descriptions as all being from the same institution. An institute may refer to itself as "State Archives of Belgium", but EHRI might speak of "State Archives and Archives in the Provinces" and CENDARI might not mention an English name, but stick to "Archives générales du Royaume et Archives de l'État dans les Provinces". A computer cannot recognize that these are actually one and the same entity.

OK, but what to do about this? There is only one full-proof solution for the issue, and that is using a unique standardized identification code for your institution. There exists such a standard (called the ISIL code, the "International Standard Identifier for Libraries and Related Organisations"), and when it is applied correctly allows you to communicate it so that anyone who wishes to refer to your institution can apply it as well, such that all descriptions of your institution can be connected in a non-ambiguous way. If you are looking for a standardized way to organize a description of your institution, there is a standard out there which can help you organize your information according to a fixed schema which will bring all this information together in one description, both for your own internal use as for use outside of the institution. So, stating the obvious is not such a bad idea: it provides your institution with its own personal ID, which allows for it to "travel" safely and unambiguously in the digital world.

The next step would be to include other basic ID information, apart from your institution's ISIL or "passport code for archives". As in our own passports we not only have an identification number – national number ("rijksregisternummer") or social security number – our passports also provide standard fields such as name, first name(s), date of birth, place of birth, nationality, and the like. There exists a standard which provides such basic fields to describe your institution as well. It is called the "International Standard for Describing Institutions with Archival Holdings" (ISDIAH). Archivists mostly refer to it under its abbreviation ISDIAH, and it is very often this type of abbreviation-speak that intimidates non-archivists (although we would think the abbreviations mostly came in to shorten the name). Another similar and frequently used standard is "Encoded Archival Guide" (EAG). This standard is for instance applied in the Archives Portal Europe. This project provides access to information on

archival material from different European countries as well as information on archival institutions throughout the continent.

Once one reads through the guidelines for these standards (freely available online, even In multiple languages, see links), one realizes that they are really helpful tools to organize the information about an institution, from the very basics such as its official name, address, opening times and contact information to its history, records management and collecting policies, a general description of its archival and other holdings as well as its finding aids, guides and publications. It is well worth your while to have a look at the different fields in these standards and to work with them, if not for the research infrastructures who wish to collect this information from you, then for publication on your own website. Stating the obvious tends to be often forgotten: very many institutions still need to provide a clear and concise overview of their institution's history and mandates or its collecting policies on their website, not to mention clear and easily found contact information and the like. So, make this one of the easily achievable, yet key first steps to improve the visibility and information of your institution to the outside world!

From the filing cabinets and printed catalogues to digital metadata

More obvious than making a standardized description for your institution is that you wish to describe the holdings of your institution, the precious archival materials which you are conserving on paper or in digital format along with other possible holdings (photo, audio, film, artefacts, etc.). It is not so long ago when the humanities and other departments at universities explained in their methodological courses the use of filing cabinets and stencilled finding aids. This was actually less than twenty years ago: even in the second half of the 1990s, instead of making a digital bibliography, students learned to compile a filing cabinet in their classes at the history department. If we are now talking about databases and different digital data organising tools, we have experienced a fast evolution, often with very steep learning curves and continuous changes in the system. The classic answer when someone asks about the (meta-)data management system used by an archive, is that "the system is currently in an adjustment process" or that "we are switching to a new program". Quite often, the enthusiasm to pick up on new trends has caused some chaos in the sense that new systems keep on being added to already existing ones and that in the end it becomes quite unclear what is described (and how and where) and where there are overlaps and the like.

Institutions which can look back to a considerable number of years of existence before the "digital age" face especially dramatic changes. Having been used to cardboard indexes and published catalogues, and finding aids written almost like books, providing detailed texts information about the creators and the history of the collection, the transition to the digital world is not so easy or straightforward. The amount of analogue information is often so vast that transferring all of it to the digital world is not done quickly

– indeed, very much on the contrary. It results in multiple systems existing one next to the other, with varying – and often unclear – degrees of overlap. Very often, a first step is to make the "classic finding aids" available online. For many institutions, this was a total culture shock and not an easy transition: keep in mind that before the digital age, every visitor was "known" to the archive, as he or she would have visited the institution, or would at least have contacted the archivist to inquire about its holdings. With the opening up of catalogues and finding aids online, the archives no longer controlled who was looking at their materials. Quite a few institutions publish their finding aids online as pdf documents. However, while this an important step forward to opening up to the outside world, this format does not allow much for systematic searches, especially not when going beyond the document one is looking at. Opening a large number pdf's and systematically searching them becomes a quite challenging endeavour. Therefore, it is interesting to learn about possible ways in which standardized descriptions, made available in equally standardized and shareable databases, can offer search facilities that throughway beyond the possibilities of a print publication or its digital counter-form online.

To complicate matters even further a new type of intermediary arose: aggregators, digital libraries and archives, and of course the digital research infrastructures and research portals. Some of them, such as Europeana (an online platform providing access to digitized collections from across Europe), concern gathering pre-existing digitized content into one common search portal. Others, such as CENDARI and EHRI, focus on bringing in descriptions of the holdings, not scans or digital images of the holdings themselves, as a way of bootstrapping the work of researchers. They aim to be discovery tools, some kind of virtual route planner, for researchers to arrive at their destination, namely, the sources they are seeking. However, as mentioned, bringing all the pieces together to build this route planner is not an easy task and has much to do with all the different formats and ways of describing the holdings at the various archival institutions. How to best deal with this? Let's look at a couple of examples of best practice, starting with how to describe your holdings.

How to find the needle in the haystack: Describing your holdings

Describing your holdings effectively is probably almost as important as preserving the actual sources. Without accurate and effective descriptions, finding the right materials and making this a repeatable process would prove to be impossible. In the filing cabinets, all file cards followed the same basic structure; in the finding aid books, the structure became clear from the table of contents and all descriptions within the book were written according to one standard template. Variations between institutions could occur, but the file cards, and the published finding aids, were never going to be partially or fully integrated into one large database – as physical objects, they remained within specific physical spaces, and were only "aggregated" in the minds and through the experiences of human users of multiple facilities. An alternative that does away with this human interpreter is very much present in the current digital world, however. And this is the main challenge we are facing at the moment: how to make sure that all these descriptions of sources and all these metadata can be accessed and searched

simultaneously, hence enlarging into an unprecedented degree the availability and visibility of information on sources spread over many different locations, both in the analogue and the digital world.

Following archival standards

Well before the digital age, archivists and archival theory worked on providing archives with a standard framework for the description of their holdings. What these standards actually do is provide a standardized template to compose collection descriptions. This does not mean that these standards are asking difficult questions for archivists to answer; rather, they organize the information the archivists are collecting into a fixed set of fields. For archival descriptions (whether you call them Collection, Record Group or Fonds), the standards "General International Standard Archival Description" (ISAD(G)) or "Encoded Archival Description" (EAD) can help. Once reading the standards, they already sound less daunting, the fields described in the standard are as logical as could be. As such there will be, for example, a field "title" to provide the title of the collection, or a field "scope and content" to describe the content of the collection. Even though these standards foresee a wide range of possible fields to fill out for descriptions of archival holdings, only a few are mandatory fields, so it depends on the policy within the archival institution or the needs of the collection being described which non-mandatory fields will be filled out as well. A very handy aspect of the standards is that the descriptions of each of the fields are freely available online, and, for the International Council on Archives ("ICA") standards, even in multiple languages. And even better is that most of the open access databases such as ICA-AtoM (a web-based archival description software following the ICA standards; "AtoM" stands for "Access to Memory") provide pop-up windows to help people encoding the descriptions with this same information on a field by field basis. Simple, right? Well, yes and no.

Indeed, it helps a great deal if everyone uses the same standardized fields, but as mentioned, humans are not half as consistent as computers and hence not everyone will interpret the rules in the same way. And this can be an individual interpretive difference, but it can equally be an interpretive difference at the institutional level. During the four years that CENDARI and EHRI have been working with the metadata from numerous archival institutions in Europe and beyond, we did not yet encounter two institutions that work in the exact same way, nor have we seen an institution which follows the ICA-standards for the full 100%. There is always a peculiarity for each and every one institution. Does this mean that it is not worth your while to invest time and energy in this matter? To the contrary! Even though there remain variations on a theme, at least everyone sings in the same key and the refrain is the same all over the place. Using the archival standards is a first and critically important step to allow efficient and effective communication of your holdings either via your own website, but especially in sharing them in a common platform together with the information from other archives, hence boosting their visibility and making them part of a larger community.

The most frequently used standards, apart from those for the description of archival holdings, are the standards to describe corporate bodies, persons and families (authority files in the jargon, see "International Standard Archival Authority Record for Corporate Bodies, Persons and Families", or ISAAR

(CPF)) and the already mentioned standard to describe archival institutions (ISDIAH). Describing corporate bodies, persons and families? What do you want us to do with that? Well, all archives come either from an institution (like a ministry, an organization or the like), from a person (be it a public figure or a less known private person) or family, or the archival holdings are describing such actors; hence we refer to them in the descriptions of our holdings as the "creator(s)" of the archive or as a keyword (access point) to the description to explain what it is about. Organizing all the information on these "authorities" can be challenging. Organisations from the former Soviet Union, for example, are referred to by their official name (in Cyrillic characters), under their abbreviation and in an English (or other translation); see for example "Чрезвычайная Государственная Комиссия", "ChGK", or "Soviet Extraordinary State Commission for the Investigation of the German-Fascist Crimes". Likewise, a person can be known under a nickname (such as a pen name or stage name), and you wish to register both the official name and the nickname. Or the person you are describing has changed his or her name. For example, you may wish to register for David Ben-Gurion his birth name, David Grün. And then you may often wish to include dates of birth/foundation, biographic information or the history of the institution. Here again, the standard foresees these fields and basically provides you with a template by which to organize your data in a systematic way. All of this has the extra benefit that other organisations using the standard will do so in the same way, hence allowing for easier data exchange.

Uniquely identifying your holdings

Now that we have the standards outlined, we still have to organize our archive and attribute identifiers to our holdings, both for digital and analogue materials. Obvious, you will say, but wait a minute. Quite often identifiers are not as straightforward as one may think. As such, we have encountered many institutions which change the identifiers for their collections over time. This is unfortunate, as researchers will refer to your holdings with different identifiers over time. Likewise, there are archival institutions that attribute multiple identifiers to one description, or multiple different descriptions to one identifier. While seemingly logical to the person describing the holdings, complex relationships that do not result in a one-on-one connection between the archival holdings and their description pose problems for the reader.

Another issue is the often too complex ways of attributing an identifier. A French archive we worked with began using Roman numerals as identifiers for its collections. While some people can still read Roman numerals without hesitation, most others have some problems as the numbers go up… Moreover, sorting the numbers also poses problems. Hence, using them is not such a great idea. This is equally valid for complicated numbers with many different characters. The key to good identifiers is a solid and simple structure and, most importantly, that they remain persistent and stable, and not change over time. Combined with the ISIL code, the unique identifier for archival holdings will be your key to opening up your sources in a sustainable way to the outside world.

But are unique and persistent identifiers the all-in-one solutions for sharing your data? Unfortunately not: sharing data, and especially seeing where the data of your institution fit with data from another

institution, be it in a common research platform or in the framework of data exchange, will always require a degree of mapping work, of seeing which of yours match with which of someone else's. Granted, doing this mapping with changing identifiers can make the mapping process a very time-consuming and non-sustainable activity, if not downright impossible. So, investing in unique and persistent identifiers is worth your while (really!).

Consider which levels of descriptions you want and need

Unlike filing cabinets and published books, databases have a tricky factor in that they require you to decide, before you begin, how the hierarchies in your descriptions need to be organized. Museums tend to have a high interest in object-level descriptions, as this is needed to build exhibits, catalogues and the like. The traditional archives tend to group their holdings in fonds, record groups or collections, and then further describe sub-levels and in some cases (but certainly not all) object-level descriptions. When deciding which levels of description your institution needs, keep in mind that you will want to share at least basic information about your holdings online (at least that is what we hope for!). Also, even if you are ready to share all your information, you may have to take into account privacy and/or copyright regulations which require certain limitations in data sharing (both on your own website and with other projects).

Another thing to carefully consider is that not all database and data management systems are flexible in changing hierarchies. Of course, there always seems to be a possibility to create "child-level" descriptions, so going top-down may not cause problems, but the opposite way, of creating parents for children, seems often to follow only biological reasoning, and hence may not be possible. Especially for those institutions which have a strong item-level focus, this can cause problems. Needless to say, creating layers in between, hence making the parent the grandparent and similar endeavours, is often not possible in the systems you are using. This is all the more reason to consider carefully before starting with your database and to first try to think of your holdings in as flat a structure as possible.

For those archives that hold copies from other repositories, we advise foreseeing a description at the same level as in the original archive. In practice, we often see that all copies from one repository are grouped into one copy-collection. In the best case, the different fonds and files from which the copies were made are mentioned in the finding aids, but often this is done without attributing a specific call number to this level. For example, RG12.001 may group copies from a departmental archive in France, but the fact that fonds A.36 was copied in there does not have an identifier of its own. In order to allow for linking between original and copy collections, this would be a quite neat asset. So keep this on your radar when bringing in those copies!

Standardizing the input

As mentioned in the introduction, the more we can standardize the information in our descriptions, the better. The best news in this paragraph is that there are actually some international standards we can

start applying in our institutions. The International Organization for Standardization or ISO has developed many such codes, which are then referred to as ISO codes (this is another abbreviation that is not half as scary as it first sounded). It is worth mentioning here that the ISO codes must have standard input for countries (ISO 3166) and for languages (ISO 639) – there are also ISO codes for (physical) containers, but we won't need those here. The computer-readable information from the codes can be applied in the language required and requested by its user. As such, "BE", "deu", "nld" and "fra" can be read as "Belgium" where one speaks "German", "Dutch" and "French" or as "België" with "Duits", "Nederlands" and "Frans".

Also for the indication of dates, where we see in a free-text environment all types of formats appearing, the ISO code 8601 – which standardizes this input like [YYYY]-[MM]-[DD] – has proven extremely useful. Oh well, you may say, we already use a standardised format and we always put it like so "12/11/2015". Great, but a European will read this date as 12 November 2015 and an American will read it as 11 December 2015. If we say we adhere to a standard, then at least we do not have to wonder: it would read 2015-11-12, because we were indeed referring to 12 November 2015, rather than 11 December of the same year. Sure, but if we just write all dates in full, then we would not have a problem either, would we? Well, actually, we would. Not only will your date-indication be limited to one language, it also becomes less machine-readable. The computer cannot easily make the link between "12 Novembre 2015" and "12 november 2015", whereas when using the standard, both would give "2015-11-12". Convincing, isn't it? The extra nice part is that the ICA conveniently refers to existing ISO codes in its description of the guidelines. So, really, it is worth your while looking into.

Now, of course the ISO does not offer a solution for all input we would like to see standardized. Many institutions, especially larger ones with more years of experience, work with a thesaurus. In the old days, these were large books lying on the desks of the employees, describing the holdings. Since publishing an updated version was not done frequently the thesauri were fairly stable tools. Nevertheless, minor to major mistakes happened along the road: from typos (like antisemitims) and variations on a word (think of anti-semitism versus antisemitism) to wayward employees creating terms out of the book and the like (such as Jew-hatred). Humans, again, can interpret these differences, but for computers the four examples above are four completely different terms. The consequence is that the four terms will not be linked, and hence will not all appear under your search results when you are using one of the four given possibilities. Using standard vocabularies, especially with a drop-down menu (meaning that you can choose the word you want from a list) are a wonderful tool in avoiding typos, creating consistency and making your work shareable, both within and beyond your institution.

Keeping the pieces together

With all the possibilities available in the digital world, many analogue sources have been digitized and are stored both in their analogue and digital format in the archives. More and more sources are in the meantime born as digital sources without an analogue original (they may have a print-out analogue representation of the digital original, though). Digital preservation is a challenge in itself, as it requires

clever storage and identification of the sources, especially when they exist in different formats. You can scan a document, then OCR it, then perform manual clean-up, then mark it with Named Entities (OCR, or "optical character recognition", converts images of text into editable and searchable machine-encoded text; Named Entity Recognition identifies elements in text such as names of persons, organizations, locations, etc.). This results in four different files, all from the same physical source, during four different stages of the digitization process. How do we keep provenance on this? It is important to create a connection, a link, between the digital data and their descriptions. Linking data with metadata (i.e., the source with the description) can be done in different ways. Here again it is important to carefully consider how your data management tools will (or will not) allow you to indicate which parts of your (meta-)data are for internal use only (staff only or consultation in the reading room), which (meta-)data can be shared on your own website and which of these can also be further communicated, such as to research infrastructures, online catalogues and the like.

Publishing about your holdings on your website

The most logical place to start opening up about your holdings online is on your own website. Most institutions offer basic to quite detailed access to their finding aids. These can be a real-time reflection of the current status of the database (live updated) or can be an output from your internal system at a given date which then requires updating as things progress. Institutions can also decide to restrict this information to classic internet pages (or "html pages"), but it could also be a decision to publish the descriptions of your holdings in such a format that they are ready to be picked up by others. Examples include .xml (Extensible Markup Language, a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable) and EAD (Encoded Archival Description, a standard for the encoding of finding aids for use in a networked (online) environment). The abbreviations may not say much, but the ideas behind them are what counts here.

"In our family we share" - export of information from your institution to the outside world

When my husband was a child – he is the second of five children – his family once went out for dinner with their grandparents. The oldest child, super excited about the event, quickly finished his whole plate and then looked at his grandmother's plate, still untouched – she was about to start eating; he contemplated for a moment, and then very matter-of-factly announced: "In *our* family we *share*..." Needless to say, the quote made it into the family collective history and is still brought up at family get-togethers. When starting to lead an international group of very diverse people with different language and cultural backgrounds, I wanted to create a sense of community and told this anecdote, upon which I added that I thought that it was actually also a very good quote for our purposes. Ever since, everyone in the group used the "in our family we share" quote. It made for very comfortable communications and a sharing positive atmosphere, not only during mealtimes at our meetings, but also in the virtual work environment we were working in. In the digital world, being known and seen is of utmost importance.

What institution is not counting the number of visitors to its website, the number of thumbs-up "likes" to its Facebook page or the amount of Twitter mentions and retweets it receives? Yet will it help us if we focus only on our own individual online presence? I don't think so. I am a strong believer that sharing will bring more benefits (not to mention a happier feeling) to all involved. At the end of the day, the effort of sharing and the diminishing of institution-exclusive information that is available only on your own website, will lead to more exposure of your institution, its holdings and the hard work you are investing in preserving and opening up the important cultural heritage you are responsible for. And let's also not forget how inclusion into a community can benefit each and every one of us.

Now, all the steps discussed above will greatly facilitate the export of information. An important factor not yet discussed is the readiness of the systems you use to export data. For this we have some golden advice: whenever you need outside assistance from an IT company to start up or improve your data management system, please do not only focus on the input and import of data, but also inquire, from the very start, as soon as the first description is in there, how to get it back out! The usability of your data management system stands or falls with its capacity to export and the formats it supports. Time and again, archives are not only set up with a system the institution cannot adjust themselves later on, they moreover also need to pay an outside company to export their data out of their own system. You do not want to be such an institution (and if you are at the moment, it's time to do something about it, or in any case be much more assertive about these matters – it's never too late to change). Having an OAI-PMH endpoint (Open Archives Initiative Protocol for Metadata Harvesting) may be a tool you wish to consider having.

Apart from being able to export your input, you will also wish to be able to make a selection in your exports, not only on topic level but also on the access level. You may wish to export only the descriptions of your photo collection on the topic of the First World War, or only the top-level descriptions of your paper archival collections on the post-Second World War trials, because the lower levels have to remain at internal level for privacy reasons. Equally, you will probably have some maintenance and other fields in the description which you wish to keep for internal use only, as well as work in-progress which is not ready to share. Ideally, all such specific requirements are on your radar when discussing which digital data management system to use. The more you can test things in the period you are installing a new program, the better.

When sharing with the outside world, the chance that there will be parallel information from different systems is real. There may, for example, be a research guide which covers all sources on the First World War in a certain country, compiled by a central institution participating in the same project as you are. Your institution's description of the sources would now be accompanied by a parallel description from this research guide. Or, you may have researchers who have indicated sources available at your institution and have described which parts of certain collections were relevant to their specific research topic. There are many other plausible scenarios. This should make us all the more aware that accountability – mostly created by a clear source reference – becomes much more important. So, before sharing your (meta-)data, make sure their provenance is clearly indicated, and also includes information on versioning, including at least the date of the last change in the description! Tracing back on your

tracks will otherwise prove to be a time-consuming and difficult process filled with uncertainties and unclarity.

Also, sharing with the outside digital world may also result in input for your institution. Typically, the digital research environments invite their visitors to be active on their portals and welcome annotations, enrichments – in short: feedback on the information presented on the website. It can well be that a researcher's notes can be highly useful for your institution and that you will wish to incorporate them into your own system. Suffice to say, without going into details on the technical requirements this entails, and the fact that here again provenance needs to be clearly documented and stated, this is a great new opportunity, making computers our medium to create a two-way exchange of information, not only output to the researchers but also input, and hence data enrichment, from the researchers.

Bad news for the anarchists: the importance of documented and well-implemented guidelines and policies

Along with standardization and sharing comes the need for clear policies, consistency in the processes and accountability or traceability of input. As long as you were working by yourself in a bubble, explaining to your visitors how you organized and described the sources you are storing, this need did not surface. In larger institutions and in a sharing environment however, one cannot go without thoroughly documented guidelines and policies and a form of quality control on the input that is being generated.

Let's illustrate this with an archival institution which started out as a small initiative in the 1990s, with three people putting together an exhibition and collecting materials for a museum which opened its doors in 1995. In the beginning, only one person of the team was describing all the identified sources for the creation of the exhibit (and the necessary background information to build the exhibit). As the material started to accumulate, she felt the need to organize the information in a structured way and so started working in an Access database, describing piece by piece the various items collected. The small museum opened and became a notable success, exceeding all expectations in the number of visitors. The museum expanded its activities with the opening of a documentation centre. Here as well, the holdings grew considerably over a relatively short time. Aware of the benefits of digitization, and with the goal of collecting all the museum's materials on the topic, even if it was in copy-format, the small museum and documentation invested greatly in digitization of materials. The data – be it in original or copy – were well-preserved. However, the adequate providing of metadata, tools by which to find the right materials and open them up for researchers, were too long neglected; this resulted in a flat text on the institution's website, which not only reflected only a fragment of the rich holdings but also changed the reference numbers for what was described as the website texts were updated. Only in the last few years, when international research infrastructures wished to open up the information on the

institution's holdings, did there grew awareness to work in a standardized digital environment so as to organize both data and metadata. The benefits of moving to standardized, structured digital metadata is not only a blessing for the visibility of the institution; it is also a key to documenting information within the organisation itself, which otherwise remains implicit, in the heads of people (who, as we know, do not always remain in the same job or position and even if they do, do not remain around forever). Having an organised system forces your organisation to establish structured internal information processes, helps you to share your information via the publication of (meta-)data on your own website, allows you to share the information of your choice with projects (research infrastructures) and creates opportunities for new methodologies, and hence helps you move to sustainable digital publishing of your archival catalogues.

Further reading?

If this text has convinced you of some or all of the benefits of sharing – and standardization as a prerequisite for this – we gladly refer provide you with some suggestions (not an exhaustive list) for further reading. Enjoy!

General International Standard Archival Description (ISAD(G))
Cendari White Book of Archives
Collaborative EuropeaN Digital Archival Research Infrastructure (CENDARI)
Data Management Planning
Digital Research Infrastructure for the Arts and Humanities (DARIAH)
Encoded Archival Description (EAD)
Encoded Archival Guide (EAG)
European Holocaust Research Infrastructure (EHRI)
Europeana
Free your metadata
Hope Wiki
ICA-AtoM (International Council on Archives - Access to Memory database)
International Council on Archives (ICA)
International Organization for Standardization (ISO)
International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR (CPF))
International Standard for Describing Institutions with Archival Holdings (ISDIAH)
International Standard Identifier for Libraries and Related Organisations (ISIL code)
Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
Standards in APEx