Elaborating a Crosswalk Between Data Documentation Initiative (DDI) and Encoded Archival Description (EAD) for an Emerging Data Archive Service Provider

Benjamin Peuch[1]

Abstract

Belgium has recently decided to integrate the Consortium of European Social Science Data Archives (CESSDA). The Social Sciences Data Archive (SODA) project aims at tackling the different challenges entailed by the setting up of a new research infrastructure in the form of a data archive. The SODA project involves an archival institution, the State Archives of Belgium, which, like most other large archival repositories around the world, work with Encoded Archival Description (EAD) for managing their metadata. There exists at the State Archives a large pipeline of programs and procedures that processes EAD documents and channels their content through different applications, such as the online catalog of the institution. Because there is a chance that the future Belgian data archive will be part of the State Archives and because DDI is the most widespread metadata standard in the social sciences as well as a requirement for joining CESSDA, the State Archives have developed a DDI-to-EAD crosswalk in order to re-use the State Archives' infrastructure for the needs of the future Belgian service provider. Technical illustrations highlight the conceptual differences between DDI and EAD and how these can be reconciled or escaped for the purpose of a data archive for the social sciences.

Keywords

crosswalk, mapping, Data Documentation Initiative (DDI), Encoded Archival Description (EAD), data archives, Consortium of European Social Science Data Archives (CESSDA ERIC)

Introduction

The Consortium of European Social Science Data Archives (CESSDA) was created in 1976 (CESSDA ERIC, 2017). Because it aims at bringing together researchers in social sciences in Europe, it now represents one of the key international institutions in this field on the continent. As such, it also constitutes one of the main networks for DDI users since its first objective is to make exchange and re-use of social science research data possible (Marker, 2013).

Fifteen European States ushered in this new organization in 1976, among which Belgium (Marker, 2013). Even so, various factors of social and political nature led this country to leave the consortium prematurely and to fall behind in terms of research infrastructure development for the social sciences (Schoups et al, 2008). Today however, Belgium has renewed its motivation to join CESSDA and become a full-fledged member in accordance with the organizational and technical requirements set out by the consortium. The setting up of a CESSDA-compliant data archive in Belgium is the objective of the Social Sciences Data Archive (SODA) project[2].

In this paper I would like to introduce a technical realization that was devised in the course of the ongoing SODA project: a mapping between two XML-based metadata standards, DDI on the one hand

and Encoded Archival Description (EAD) on the other hand. The idea of such a crosswalk stems from an institutional partnership between actors from the social science community in Belgium and archivists working for the State. After I recount the context of the project, I will explain the rationale of the crosswalk and describe the materials and methods that were used to develop the mapping. I will then provide some technical illustrations for the obstacles that arose and the solutions that were proposed and conclude by outlining the next steps in the process.

Social scientists and archivists

Following in the footsteps of the other CESSDA Service Providers, the SODA project seeks to bring together the Belgian researchers in social sciences in order to promote sharing and re-use of research data. As such, this project involves representatives from the social science community, but also archivists. The latter often have academic backgrounds in history and historiography. As such, their discipline is more akin to the humanities than to the social sciences. Still they are involved in the SODA project for several reasons which will be explained hereafter.

The SODA project brings together three partners: 1) the *Vrije Universiteit Brussel*[3], a Dutch-speaking university based in Brussels; 2) the Catholic University of Louvain, the French-speaking university that used to manage the now-defunct Belgian Archives for Social Sciences (BASS); and 3) the State Archives of Belgium, the institution responsible for preserving and giving access to administrative and historical records produced by Belgian public offices, courts, notaries, as well as private organizations and influential families.

The role of archives is to store documents that no longer serve a purpose for their original creators so that future historians may access them and use their contents to advance historical research. Such documents come in many shapes: reports, minutes, letters, accounts, charters, drafts… and they now also come in both analog (paper) and digital form. The key organizing principle in archives is provenance: materials should be grouped according to their original source (e.g. a government department or a particular court of justice) in order to preserve, along with the materials themselves, the context of their creation (Morris and Rose, 2010). Archivists document this context with finding aids, i.e. 'description[s] of records that giv[e] the repository physical and intellectual control over the materials and that assis[t] users to gain access to and understand the materials' (Pearce-Moses, 2005: 168). In other words, finding aids constitute the metadata of historiography. And just like the social science community procured an XML language for encoding its metadata in the form of DDI, historians devised the Encoded Archival Description (EAD) from that same 'meta-markup language' (van Hooland and Verborgh, 2014: 16, 28-43).

The purpose of EAD is to convert finding aids in digital form. Like DDI, it was designed in the mid-1990s (Pitti, 1997; 1999; Dryden, 2010; Rasmussen and Blank, 2007). Also like DDI, EAD evolved over time and was further developed by its research community of origin so that it now comprises several versions (DDI Alliance, 2017a; 2017b; Stevenson, 2016). Years ago, historians, genealogists and journalists had to go to the archives, sometimes at a heavy personal cost, before they could even be sure of the relevance of a repository's collections to their work. Nowadays, thanks to EAD, archival institutions around the world can share their finding aids online, thus simplifying the first steps of historical research significantly.

A crosswalk between two metadata standards

There now exist, in large repositories like the State Archives of Belgium, whole infrastructures that were devised around EAD so as to smoothen the process of writing finding aids, then converting them in machine-operable formats and finally making them accessible to users through open public access catalogs (OPACs). Because the current economic context deters the Belgian stakeholders from investing much into scientific projects that do not involve the hard sciences, the actors of the SODA project are working towards re-using the existing infrastructure and its underlying workflow at the State Archives for the needs of the future Belgian service provider. This workflow, or 'pipeline', comprises procedures, programs, tested and tried methods, and well-defined roles and tasks. If all or parts of it can be re-used for a social science data archive, then implementation will be quicker and costs will be reduced to a large extent.

The gains from this joint venture would be twofold: first, it would support the setting up of the infrastructure for a brand new institution, thus advancing social science research in Belgium; secondly, it would also benefit the field of historical research, as the State Archives themselves could re-use social science data as well as make it accessible to its very own audience, thus promoting interdisciplinarity[4]. Furthermore, the pilot study that is the SODA project currently involves only two of the twelve Belgian universities, but, as the future service provider will hopefully include all twelve in a cooperative endeavor along with the State Archives, it will thus bring together the country's two largest linguistic communities—the Dutch- and the French-speaking, upon whose governments the universities rely for funding—and the Federal State, to which the State Archives belong[5]. In this way, although the business plan of the future data archive is still in the making, the State Archives will take on the tasks of archiving and disseminating the research data, while researchers and their universities will be the depositors and the end users of the subsequent new service.

However, this entails bridging the gap between the aforementioned respective metadata standards: DDI, which is a CESSDA requirement (CMM Group, 2016) on the one hand, and EAD on the other. If DDI files can be converted into EAD files in a fully or partially automated manner, then the tools and the skills of the State Archives can be re-used for the needs of the future Belgian service provider just like researchers will be able to re-use the data made available. The idea is not fully to replace DDI files with EAD files, but to funnel the content of DDI files into the 'pipeline' of the State Archives by means of EAD even while retaining and making the DDI files accessible. Precisely what actions and services will be performed with the DDI files on the one hand and the EAD files on the other hand remain to be determined; yet the benefit of converting the former into the latter both for the future service provider and for the State Archives is a given. Hopefully this whole mapping enterprise comes off as heuristic, although it might also appear somewhat misguided. One of the reasons is the difficulty in envisioning the architecture of the future service provider, as much on the legal and institutional as on the technical and technological planes. Tackling other questions such as the choice of a data management program or the crafting of a data transfer policy for future depositors will likely affect the crosswalk as well. However, in order to procure concrete deliverables that can be placed before the stakeholders, the decision was made to undertake the mapping.

The conversion from DDI to EAD can be done by means of a crosswalk, or 'mapping', from one element or 'tag' to another. A crosswalk can be defined as follows: 'A *crosswalk* defines the semantic mappings of the fields of a source metadata schema to the fields of a target metadata schema, so as to semantically translate the description of sources encoded in different schemas. A crosswalk is expressed through a table that shows the equivalent metadata fields of the metadata schemas involved' (Gaitanou, Bountouri and Gergatsoulis, 2012: 264, emphasis in the original). The goal of crosswalks is often to allow 'metadata created by one community to be used by another group that employs a different metadata standard' (UNESCO, 2015: 53) as is one of the motivations of the SODA project.

Preliminary searches aimed at uncovering existing crosswalks between DDI and EAD proved unfruitful. The tag library of EAD 2002 itself features crosswalks between EAD and the Dublin Core as well as between EAD and ISAD(G)[6], the international standard for the description of archival materials (SAA, 2002). However no crosswalks involving EAD and DDI could be found, apart from a modest mapping between select elements from different schemas put out by the Emory University Libraries & Information Technology department (Emory U, 2017). While thorough and well thought out, this crosswalk seems to proceed from the spirit of metadata standards such as the Dublin Core (DC), i.e. a more minimalist approach, one focused on the very essentials of metadata across fields and disciplines (Doorn and Tjalsma, 2007; Méndez and van Hooland, 2014). A metadata standard such as DC is meant to record the purely descriptive and administrative metadata, i.e. the information related to the identification, use, and management of the described objects (Caplan, 2003). But archival description also heavily relies on technical, structural, and preservation metadata, which respectively specify the use requirements, the internal organization, and the storage conditions of the documents (Oliver and Harvey, 2016). The 2002 version of EAD offers 146 elements for use (vs. 15 for DC) and though only a subset is mandatory, this number hints at the large and diverse range of information that may be encoded for the proper description of a group of archives.

DDI and EAD

Because DDI and EAD are two complex, high-level markup languages applied to two different scientific disciplines, we at the State Archives assumed at the outset of the project that a perfect one-to-one mapping between the two could not be achieved. However, as previously stated, the rationale of the crosswalk was not to do away with DDI, but to make the DDI-encoded metadata intelligible for the EAD-minded workflow of the State Archives. The future data archive will preserve the DDI files and use them whenever the already existing procedures cannot address the needs of the users[7].

Little ought to be said about DDI here. However, it should be said that, for reasons that will be made clear further along, the version of DDI that was selected for the purpose of the mapping was DDI-Codebook because it was considered—with only a basic knowledge of the two DDI branches, Codebook and Lifecycle—the one most likely to be similar to EAD.

As for EAD, as was previously stated, it now comes into three versions: EAD 1.0, which was published in 1998 (Pitti, 1999); EAD 2002; and EAD3, which was published in 2015 (Stevenson, 2016). To this date the States Archives of Belgium as well as the Archives Portal Europe (APE), which disseminates the metadata of archival repositories in Europe, still use EAD 2002. Moreover, accounts of full-fledged implementations of EAD3 or of transitions from EAD 2002 to EAD3 are still awaited. That is why the mapping was made only towards EAD 2002 for now.

It has been more than 15 years since the publication of EAD 2002. Contrary to EAD3, the 2002 version came out at a time when yet few finding aids had been converted by archival repositories around the world: back then, archivists were still struggling to embrace the digital transition (Allison-Bunnell, 2016). Today EAD has truly become a standard, for it was adopted by most large archival institutions (Francisco-Revilla et al, 2014). That being said, it has drawn much criticism, even shortly after its publication. First, the complexity of the metadata language came under fire: as it turned out, the learning curve was very steep for such an expansive library of tags (Yakel and Kim, 2005). The fact that EAD inherited the peculiar, much-criticized syntax of XML did not help[8], nor did the fact that, while computer programmers themselves scoffed XML-based languages (Pilgrim, 2009-2011), archivists and historians with varying, uneven levels of proficiency in computer science were bound to struggle with this new standard. It took some time for universities to include EAD or at least XML training in their curricula destined to future archivists and librarians (Fox, 2005), yet one can see now that EAD is more and more frequently required by employers in the field (Riggs, 2005)[9].

Other critiques targeted the excessive flexibility of EAD 2002: several authors pointed out that the lack of mandatory tags and of guidelines resulted in different archivists encoding the same kind of information in different places in EAD files (Shaw, 2001; Francisco-Revilla et al, 2014). For instance, Luis Francisco-Revilla and his co-authors noted that such an essential element as 'Repository' (<repository>), whose purpose is to record the name of the archival institution responsible for providing access to the described materials, can be left out of an EAD 2002 document (Francisco-Revilla et al, 2014).

However it is unfair, as many authors did, to heap on EAD criticism that actually pertains to the deficiencies of the archival milieu at large. For instance, the subjective difficulty in appropriating EAD was and still is largely due to the lack of training (or interest) from archivists into computer matters (Shaw, 2001; Yaco, 2008; Dow, 2009). Lack of funding for the purchase of computers, software, training sessions, and other such spendings for the implementation of EAD cannot be seriously attributed to EAD itself either: technology always comes at a cost. Furthermore, as Dennis Meissner and his colleagues were brave enough to admit, attempts to convert finding aids into EAD might incidentally reveal to some archivists the flaws in the finding aids themselves. Meissner and his colleagues acted upon this by rewriting their finding aids, a task that proved strenuous and time-consuming yet with great benefits (Meissner, 1997). Technological transitions, even the most successful ones, are always painful at any rate: as Elizabeth Dow reminds her readers: '[T]he archival community could take a lesson from the agonies librarians suffered as they converted card catalogs to online public access catalogs. By all accounts, the cleanliness of the data on the cards made a huge difference in the cost, efficiency, and sense of success of the project. The lesson: create clean data to start with' (Dow, 2009: 114).

It is worth noting that much of the criticism leveled at EAD echoes with the critiques of DDI: the latter was also said to be very complex and sometimes out of touch with the true needs of the social science researchers in terms of data documentation. Similarly, problems that affect the community at large like the lack of proper tools or training, the need for organizational change, and a slow adoption rate hindered the initial widespread acceptance of DDI (Wackerow and Vardigan, 2013). Nevertheless, just like EAD within the archives, DDI has now made its way into the world of social sciences and is actively promoted through such vast networks as CESSDA or the standard's dedicated conferences, EDDI and NADDI.

Materials and methods

DDI versions

The choice of one of the two branches of DDI, Codebook and Lifecycle, was not an easy one, as it depended upon several factors of differing nature. First, there was the fact that the author does not have an extensive background in social sciences, so knowledge of DDI and of the current issues in metadata management in that same field had to be acquired through on-the-job training[10]. Next, there was the fact that the SODA partners knew little about the specific Belgian context: how widespread are data documentation practices in the Belgian social science academia? Do researchers know about Nesstar or Dataverse? How much inclined would they be to share their datasets? A survey was designed to gather information on this topic, but the modest resources of the project and the tight schedule that they entailed forced the project members to move on and produce deliverables, sometimes in a counter-intuitive order.

Assuming, as we were eventually able to, that the data documentation practices are lacking in the Belgian social science community, we saw in this situation an opportunity to instill good practices where there were few or none by offering to our target audience a consistent framework that had been thought out for their needs. However, such a framework had to be coherent with the State Archives' technical resources as well as its own interests in the matter. For one thing, like all archival repositories, the State Archives have devised a particular EAD template for their finding aids, selecting some of the tags from the library and leaving others out. Knowing this, the question that arose was: which, of DDI-Codebook and DDI-Lifecycle, is more likely to parallel EAD in general and the State Archives' own version more specifically?

The idea arose that a codebook is more akin to a finding aid than a conceptual construct of the whole lifecycle of a dataset. As a discrete intellectual work that seeks to convey the meaning of codes who, without proper context, are unreadable, a codebook fulfills missions that are similar to that of finding aids, which provide researchers with the historical circumstances that determined the creation of records. Just like a codebook will help researchers understand the role of variable 'V29', whose sole name is of no help in grasping its purpose, a finding aid will guide historians as they long to learn about, for example, the enigmatic USA Board of Tea Appeals[11].

A valid critique of this stance is conveyed by the fact that, as finding aids serve to record the context of creation of documents, by documenting the custodial history and the conditions of acquisition of said documents by a repository before they were processed and arranged in a certain way by one or more archivists, one might say that finding aids in fact document the lifecycle of those records. This consideration regrettably occurred to the author only after the choice of DDI-Codebook was made, and likely so for two reasons: first, to an outsider, DDI-Lifecycle, both in its concept and in its peculiar XML rendition, can be very perplexing[12], as it entails a good understanding of the actual lifecycle of social science research data nowadays as well as a very good command of XML; second, the idea of a codebook—both conceptually and, in more practical terms, as a literal 'book of codes' that one might hold in one's hands—may prove more intuitive and relatable to other, more common forms of media.

At first, only a high-level, almost superficial comparison of the general structure of EAD 2002 on the one hand and of DDI-Codebook and of -Lifecycle on other hand could be made, working with peer-reviewed publications as general guides and with the token DDI files and the DDI tag library ('online field level documentation') put out by the DDI Alliance on its website (DDI Alliance, 2017c). As will be shown when discussing the method behind the process, it seemed at this level of analysis that broad, general correspondences could be drawn between the key sections of DDI-Codebook and EAD, and while it arose along the mapping process that this claim ought to be mitigated, this further supported the choice of DDI-Codebook.

DDI 2 only builds upon the very first version of DDI (DDI Alliance, 2017d), and because DDI 2.5, the latest Codebook version, is backward compatible with DDI 2.1 (DDI Alliance, 2017e), DDI 2.5 was chosen as the 'source language' for the crosswalk[13]. Thus, a mapping that could process files following the rules of DDI 2.5 would also be able to handle files that draw on the DDI 1 specifications.

A DDI corpus

Thanks to the hard work of the CESSDA Service Providers, many full-fledged data archives now exist and disseminate, especially via Nesstar servers, datasets and metadata. The latter files are most often accessible to all, which is why actual instances of XML-DDI files could be gathered online and assembled into a corpus that enabled systematic search for specific occurrences of elements or attributes. In this way, it was possible to see them in context as well as to observe variation among the different uses and practices of metadata managers. For example, in cases where only one date was encoded in the <stdyDscr> wrapper, suggesting that this was the date when the study as a whole had reached its conclusion, this date could be found either in prodDate, in prodDate ATT date, in distDate, in distDate ATT date, in depDate, in version, or in version ATT date.

Files produced in conformity with the syntax of DDI 2 were gathered from the online, open-access repositories of several data archives. The corpus was not put together in a thoroughly systematic fashion, by seeking files from all the European data archives for instance; rather, the governing criterion was the ease of retrieval of the metadata files. The DDI files were the following:

Title	**Identifier**	**Data Archive**
Labour Force Survey, January 2017 [Canada]	LFS-71M0001-E-2017-January	ABACUS (Canada)
Class and Social Structure of the Population of Czechoslovakia in 1984 - module for individuals	CSDA00111en	CSDA (Czechia)
Politbarometer Short Inquiry 2002 (Accumulated Dataset)	ZA3851	GESIS (Germany)
American National Election Study, 2004: Pre- and Post-Election Survey	4245 10.3886/ICPSR04245.v2	ICPSR (Unites States of America)
National Travel Survey, 2002-2015	5340	ICPSR (Unites States of America)
Euro-Barometer 10 -- October - November, 1978	7728	ICPSR (Unites States of America)

Migrations between Africa and Europe – MAFE Senegal (2008)	IE0216A	Ined (France)
Selected Teagasc National Farm Survey Data 2007	Amended!Teagasc!NFS!2007!Portal!Data	ISSDA (Ireland)
ISSP 2015: Darbo orientacijos IV, 2015 m. spalis - gruodis, 2 leidimas	LiDA_ISSP_0296_STUDY_02	LiDA (Lithuania)
A Smile Is Not Enough, Part I: Developing an intervention for continuous positive feedback, 2013	NSD2045	NSD (Norway)
Baromètre Politique Français 2006-2007 Vague 1	fr.cdsp.ddi.BPF2007-R1	Réseau Quetelet (France)
ISSP Slovensko 2009-2010	SASD 2009001	SASD (Slovaquia)
Institutional Trust 2013	SND0963-001	SND (Sweden)
Socioeconomic inequalities in Thrace: Living conditions, education and employment	A1_en	So.Da.Net (Greece)
General Election Study Belgium 2003	Electionsbelges2003b	SOHDA (Belgium)[14]
OECD Main Economic Indicators Databank, 1960-2017	4744 10.5257/oecd/mei/2013-04	UKDA (United Kingdom)
Multiscopo ISTAT – Time use - 2002-2003	it.adpss-sociodata.ddi.SN054	UniData (Italy)

Table 1. List of DDI files that constitute the test corpus.

Several of these files only provide information for the description of the study and of the DDI document (<docDscr> and <stdyDscr>); others feature variable-level description (<varDscr>). Some tags or attributes could not be found, such as the sdatrefs, methrefs, and pubrefs attributes of such elements as <fileDscr>, or the <controlledVocabUsed> element within <docDscr> and all of its sub-elements. Yet the assumption was not made that those elements were rarely if ever used since the present corpus was created only as a tool for the purpose of the mapping and not as a representative sample of DDI files.

Software

Because a mapping consists in a correlation table, it seemed only natural to resort to Microsoft Excel so as to flesh out the correspondences in an orderly and systematic fashion. While the need for structure was high, the complexity of the ensuing model was low, which is why more versatile software like LaTeX did not come under consideration. Cross-files searches into the corpus of DDI files were performed with Notepad++.

Method

In order to present the method with which the mapping process was carried out, a word must be said about its two guiding principles. First, there was the 'orientation' of the mapping, i.e. determining which of DDI and EAD would be the starting point of the mapping and which would be the receiving end. Both configurations can be advocated for a number of reasons: on the one hand, DDI is the source material since this is all about the management of social science metadata; on the other hand, if the State Archives are to fulfill several of the tasks of the future Belgian service provider and if they are to do so by re-using their EAD-based software infrastructure for the needs of the data archive, then EAD must be given conceptual precedence. Which standard ought to be adapted in order to transfer easily into the other proved to be a tricky question. Eventually, the decision was made to begin with DDI because, since the author did not have an extensive knowledge of social sciences, it seemed more logical to embrace the whole of DDI and thus get better acquainted with this scientific field through its standard before looking for its conceptual equivalents in EAD. Right at the outset, the idea arose that, after the DDI-to-EAD crosswalk would be finished, a 'cross-mapping' might have to be produced in turn, although more specifically from the State Archives' specific take on EAD towards DDI.

The second guiding principle consisted in the awareness of the opposition between syntax and semantics in computer science. Depending on the context, precedence goes to one of these two concepts, which both originate from linguistics and which designate respectively 'the way in which linguistic elements (such as words) are put together to form constituents (such as phrases or clauses)' and 'the historical and psychological study and the classification of changes in the signification of words or forms viewed as factors in linguistic development' (Merriam-Webster, 2018).

At the beginning of the project, the author had to acquire knowledge about the social sciences in general and about DDI in particular by himself, on the job. To this end, the insight of the university project partners, two Belgian social scientists, proved most valuable. Yet limitations came into light when the focus shifted to DDI: like most traditional historians and archivists do not know about EAD, leaving such technical topics to computer scientists, it turned out that social scientists have often never heard about DDI, as they turn to academic librarians, the ICT department in their institutions or graduate students for questions of documenting and archiving research materials. At best, researchers know about such programs as Nesstar or Dataverse, but the underlying technicalities such as the metadata standard in use remain in the shadows for them.

And yet a metadata standard is certainly not the most 'technical' aspect of data management—i.e. neither the most complex, nor the most vital one—compared, for example, with the maintenance of the whole hardware infrastructure of a data archive. In most cases, the data encoded in EAD or DDI are subject to only few checks and procedures, as opposed to other computer operations where the input of invalid data can lead to highly detrimental consequences such as the need to restart a lengthy and costly procedure or the corruption of data. Let us imagine that a person in charge of documenting a dataset, instead of encoding the description of the study in stdyInfo/abstract, inserts the descriptive text in method/dataColl/timeMeth. Since <timeMeth> allows for character data, this would not constitute a syntax error, although it would be a semantic one. While problematic, this mishap would probably not cripple the readability or usability of the dataset; at worst, the problem would be easily noticeable on the interface that displays the metadata of the dataset and could be quickly solved. Compare this situation with that of a computer programmer who inadvertently removes a single comma from an algorithm upon which several automated processes rely. Because of the unforgiving syntax rules of most programming languages, this would amount to a spanner thrown in the works, and it might be long before the cause of the ensuing problems is finally identified. Apparently, the notion that the failure of the launch of the Mariner 1 space shuttle was due to a misplaced comma in lieu of a period somewhere in the FORTRAN computer code is inaccurate; still, part of the true cause of the problem was the (infinitesimal yet essential) lack of a 'superscript bar' in the transcript of a command for a smooth function (De Florio, 2009: 32).

That is to say, in effect, that contrary to operational computer functions where proper syntax is key, many elements in XML languages such as EAD or DDI are but 'meaningful titles': they are primarily containers meant to store text, which will be parsed on a superficial level by various programs yet truly 'processed' only later on by human beings. Like parsing programs verify the syntax of computer code, humans perform a similar operation when they process information, although on the semantic plane, where rules are far less strict and there is much more room for interpretation. One might say that the meaning attached to this or that element of an XML language is the 'human factor' of the standard: it is subject to interpretation, therefore to variation, and one must strive to apprehend its rationale so as not to adulterate its function. This was an important concern in the making of the mapping: while there was no doubt that many cases would prove to be problematic and require hard choices, much work was dedicated to grasping the original purpose of the DDI-Codebook elements and attributes so as to reduce as much as possible the potential gap between them and EAD's own elements and attributes.

Results and discussion

At the end of the mapping procedure, the starting hypothesis was successfully verified: because, *in abstracto*, EAD is meant to document objects that require archiving, it could be bent for the needs of a data archive for the social sciences with appreciable pliability. Like special characters that require particular encoding procedures in certain computer environments, some elements of DDI had to be 'escaped' in various ways in order to conform to the syntax of EAD, yet the flexibility of EAD allowed for such remedial reallocations.

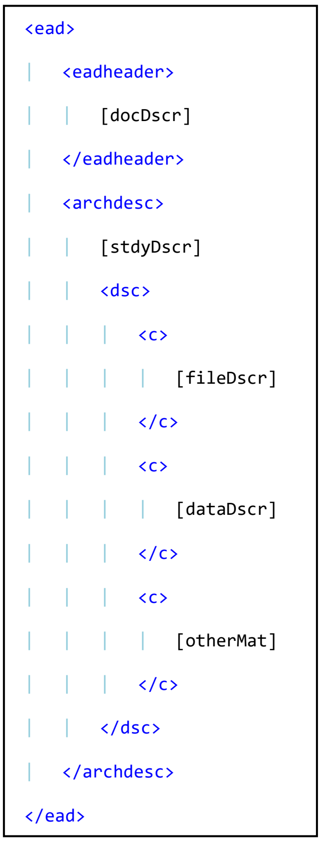
To illustrate, the following schemas show high-level templates for both metadata standards and the third figure presents a possible structural mapping:

```
<codebook>
|   <docDscr>
|   |   Information about the DDI file and the source codebook
|   </docDscr>
|   <stdyDscr>
|   |   Information about the study and the dataset
|   </stdyDscr>
|   <fileDscr>
|   |   Information about individual files
|   </fileDscr>
|   <dataDscr>
|   |   Information about variables
|   </dataDscr>
|   <otherMat>
|   |   Other study-related materials
|   </otherMat>
</codebook>
```

Figure 1. High-level DDI-Codebook template.

```
<ead>
|   <eadheader>
|   |   Bibliographic and descriptive information about the finding aid
|   </eadheader>
|   <archdesc>
|   |   Information about the content, context, and extent of the body of archival materials
|   |   <dsc>
|   |   |   Hierarchical groupings of the archival materials
|   |   </dsc>
|   </archdesc>
</ead>
```

Figure 2. High-level EAD template.



```
<ead>
|   <eadheader>
|   |   [docDscr]
|   </eadheader>
|   <archdesc>
|   |   [stdyDscr]
|   |   <dsc>
|   |   |   <c>
|   |   |   |   [fileDscr]
|   |   |   </c>
|   |   |   <c>
|   |   |   |   [dataDscr]
|   |   |   </c>
|   |   |   <c>
|   |   |   |   [otherMat]
|   |   |   </c>
|   |   </dsc>
|   </archdesc>
</ead>
```

Figure 3. Possible structural mapping of DDI-Codebook towards EAD.

Figure 3 shows what could have been and what, in part, is in the crosswalk. Mostly, the contents of the instances of <fileDscr>, of <dataDscr>, and the <otherMat> section are conveniently transferred into 'Component' (<c>) sections in EAD with proper labeling, both to facilitate their identification should a technician need to look at the 'source code' in the archival institution (the EAD file in this context) and to prepare the subsequent transfer of the formerly DDI-encoded information from the EAD document to a more dynamic, user-friendly format—typically on a bibliographic record displayed through an OPAC.

However the seemingly one-to-one relocation of the key <docDscr> and <stdyDscr> sections into the 'EAD Header' (<eadheader>) and the 'Archival Description' (<archdesc>) sections respectively does not quite correspond to the eventual crosswalk. This is in spite of the fact that one could set down the following hypothetical analogy:

EAD	**DDI**
'EAD Header' section <eadheader>	'Document Description' section <docDscr>
Concerns: the finding aid(s) covering the archival materials	Concerns: the DDI file and the source codebook
'Archival Description' section <archdesc>	'Study Description' section <stdyDscr>

Concerns: the archival materials proper	Concerns: the study and the dataset

Table 2. A seemingly accurate parallel between the objects concerned by EAD and DDI.

At this level of analysis, these sections seem to mirror each other. The archival materials and the study results, on the one hand, constitute the 'data' of history and social sciences respectively, and, on the other hand, the finding aid and the DDI file (which explicitly corresponds to the codebook in this instance) are the 'metadata' proper. However, in depth-analysis soon reveals key structural differences between their respective encoding constraints.

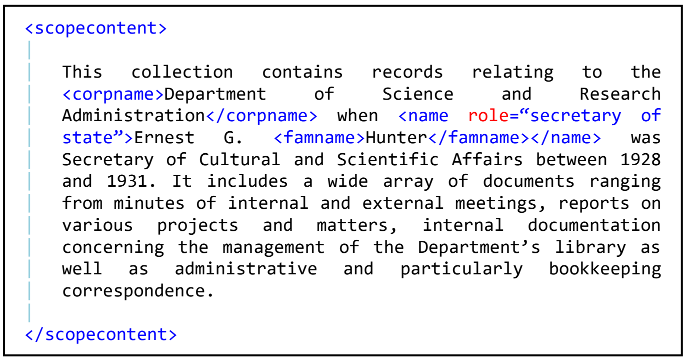
'Document Description' (<docDscr>) and 'Study Description' (<stdyDscr>) contain the most essential items of information concerning the dataset and its origins, specifically the descriptive and administrative metadata, such as the names of the people who partook in the study that resulted in the very dataset. Such information could not simply be split as initially envisaged. A contributor who helped collect the data during the field work procedures should appear in the descriptive and administrative metadata along another contributor who, instead, encoded the information in the study-related codebook into a DDI file, even though their names and affiliations are not 'stored' in the same section in DDI. This reflects the choice of the developers of DDI, who decided to distribute the information into a potentially vast array of elements and subelements, e.g. the many possible instances of 'Bibliographic Citation'. However, this was one of the instances where EAD proved to work rather differently. An illustrative example to contrast these sections is the part played by, on the one hand, the large group of sub-elements contained in 'Bibliographic Citation' in DDI, and, on the other hand, the isolated 'Author' (<author>) element in EAD.

Authors and contributors

DDI's 'Bibliographic Citation' section allows for the use of a large group of elements that describe the various contributors and participants whether for the production of the dataset, the fulfillment of the study, the production of a work or the fulfillment of another study referenced in the metadata, etc. Between the 'Producer' (<producer>), the individuals or corporate bodies identified in the 'Version Responsibility Statement' (<verResp>), those responsible for the various 'Notes and Comments' (<notes>, ATT resp), the 'Distributor' (<distrbtr>), the 'Depositor' (<depositr>), the 'Contact Persons' (<contact>), the 'Authoring Entity' or 'Primary Investigator' (<AuthEnty>), and the 'Other Identifications / Acknowledgments' (<othId>), the encoder is almost spoilt for choice when it comes to recording the names and roles of the various people who contributed in one way or another to the realization of the study and of its subsequent dataset.

On the other hand, in the case of EAD, there are quite fewer tags for encoding person- or institution-related metadata, and they are fairly tightly confined to certain parts of the metadata document. Essentially, the 'Author' element records the '[n]ame(s) of institution(s) or individual(s) responsible for compiling the intellectual content of the finding aid' (SAA, 2002: 48) and, for other uses and requirements, the 'Name' (<name>) element and its more specific variants, 'Family Name' (<famname>) and 'Corporate Name' (<corpname>), can be used. The latter three elements were designed for tagging significant names in the sections of the EAD document that describe the archival materials *per se*, as opposed to 'Author', whose purpose is to encode the name of the finding aid's creator. For example, the names of certain public servants or historical figures may be tagged with

<name> or <famname> so as to signal their presence in or relation to the archival materials and thus guide the researchers looking for archives that mention them. The following example shows a fictional instance of the 'Scope and Content' (<scopecontent>) section of an EAD document, which is meant to 'summarizing the range and topical coverage of the described materials' (SAA, 2002: 229), with illustrations of the use of the name-related tags:



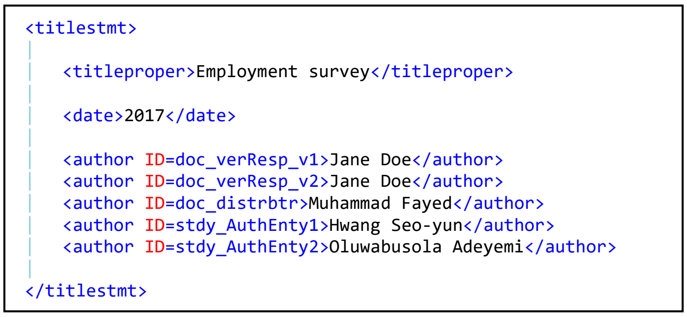
```
<scopecontent>

    This collection contains records relating to the
    <corpname>Department of Science and Research
    Administration</corpname> when <name role="secretary of
    state">Ernest G. <famname>Hunter</famname></name> was
    Secretary of Cultural and Scientific Affairs between 1928
    and 1931. It includes a wide array of documents ranging
    from minutes of internal and external meetings, reports on
    various projects and matters, internal documentation
    concerning the management of the Department's library as
    well as administrative and particularly bookkeeping
    correspondence.

</scopecontent>
```

Figure 4. A fictional example of a 'Scope and Content' section in a typical EAD document.

As illustrated, the various <name> elements in EAD are tags meant to be used for labeling data and not metadata. This is why referencing the names of those who participated in fulfilling the study and producing the dataset—the people referenced in the 'Study Description' section—into the 'Archival Description' (<archdesc>) section in EAD would inevitably break with the spirit of the archival standard. The problem might be circumvented by specifying the nature of the contribution of each participant through the ATT role of the 'Name' (<name>) element. Yet not only would an ATT type, which <name> is not endowed with, be preferable in this case; it would still be a hassle to separate, on the one hand, the names of those who partook in the creation of the source codebook and of the ensuing DDI file, and, on the other hand, the names of those who directly contributed to the realization of the study and the production of the related dataset[15]. While the question of authorship is an endless debate, especially in the case of complex creations such as motion pictures or social sciences datasets, the methods and precepts of archivists, which are fairly well translated by the structure of EAD, advise for the concentration of all contributors in one large super-section, i.e. the 'Title Statement' (<titlestmt>) within 'EAD Header' (<eadheader>)[16].

The 'Author' element in EAD has few attributes that could help distinguish the various types of contributors. However, this can be easily remedied by automatically assigning meaningful identifiers to the instances of that element (all elements have a unique ID attribute in EAD).

The resulting 'Title Statement' section in EAD could then look as follows:



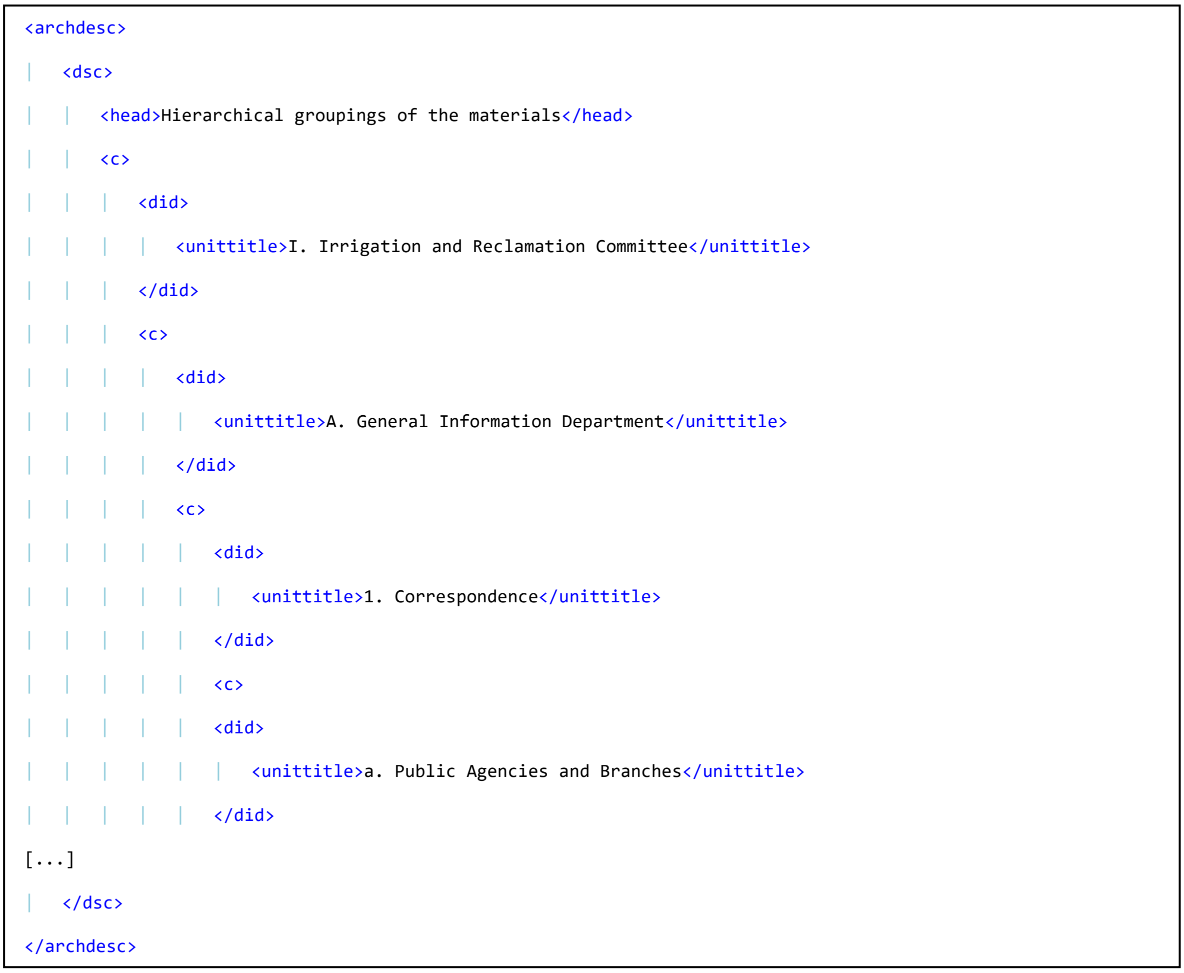
```
<titlestmt>

    <titleproper>Employment survey</titleproper>

    <date>2017</date>

    <author ID=doc_verResp_v1>Jane Doe</author>
    <author ID=doc_verResp_v2>Jane Doe</author>
    <author ID=doc_distrbtr>Muhammad Fayed</author>
    <author ID=stdy_AuthEnty1>Hwang Seo-yun</author>
    <author ID=stdy_AuthEnty2>Oluwabusola Adeyemi</author>

</titlestmt>
```

Figure 5. A fictional example of a ‘Title Statement’ section in a DDI-based EAD document.

To make sure that identifiers are, in conformity with the rules of EAD, unique for each element in the file, an algorithm will have to generate identifiers with strict criteria: the identification of the provenance section; the retrieval of the role played by the contributor, whether as-is or according to a local controlled vocabulary; the inclusion of implicit information, i.e. information that is not originally contained within the mapped element itself, such as the version number as seen in Figure 5; and, finally, when there are multiple individuals for one category, a disambiguation number, as in Figure 5 in the case of the ‘Authoring Entities / Primary Investigators’.

Identifier overload

Another potentially problematic case is that of identifiers. Here, EAD shows practical flexibility, as each and every element can receive an ID attribute, which must be unique. This allows for much granularity, which can be especially useful for complex archive collections. For example, a finding aid that describes a small group of documents might require only a few identifiers, such as a dedicated permanent identifier and a call number. On the other hand, more identifiers might be required for a finding aid that covers a very large and complex collection of archival materials. Some of the most problematic cases include miscellaneous records from different archive producers, existing in various formats, some of which require specific preservation measures, which possibly entails scattering the materials across one or several repositories in the worst scenarios. In order to give both material and intellectual structure to such collections, archivists sometimes must compose very complex and hierarchical descriptions with many divisions and subdivisions, resulting in deep-reaching ‘trees’ of embedded ‘Component’ (<c>) elements, as illustrated with Figure 6:



```
<archdesc>
|   <dsc>
|   |   <head>Hierarchical groupings of the materials</head>
|   |   <c>
|   |   |   <did>
|   |   |   |   <unittitle>I. Irrigation and Reclamation Committee</unittitle>
|   |   |   </did>
|   |   |   <c>
|   |   |   |   <did>
|   |   |   |   |   <unittitle>A. General Information Department</unittitle>
|   |   |   |   </did>
|   |   |   |   <c>
|   |   |   |   |   <did>
|   |   |   |   |   |   <unittitle>1. Correspondence</unittitle>
|   |   |   |   |   </did>
|   |   |   |   |   <c>
|   |   |   |   |   <did>
|   |   |   |   |   |   <unittitle>a. Public Agencies and Branches</unittitle>
|   |   |   |   |   </did>
[...]
|   </dsc>
</archdesc>
```

Figure 6. A fictional example of a deep-reaching EAD ‘tree’ of embedded ‘Component’ (<c>) elements.

In such cases, it can be especially helpful to allocate identifiers either to the deepest components, whether with the ‘ID of the Unit’ (<unitid>) element or with the ID attribute of the ‘Component’ elements, or to all echelons of the hierarchy, although the latter situation is rare since it requires much work of generation and management of identifiers.

The problem that might arise when it comes to mapping DDI to EAD would be the following: what is to be done with the identifiers already present in DDI? Even if the future EAD files are not meant fully to replace the DDI files, it would still be interesting to transfer the original DDI-encoded identifiers over to EAD. Yet the future data archive is likely to devise a policy for generating and managing identifiers of its own. Would it be possible to retain the original identifiers on the one hand while enriching the EAD documents with a set of new identifiers? It could be interesting to present the ‘structure’ of the dataset by listing the different files that make it up into EAD’s ‘Description of Subordinate Components’ (<dsc>) section. But where exactly should the identifying information transfer? Into <unitid>, and/or in the ID attribute of that same element, and/or in that of ‘Title of the Unit’ (<unittitle>), and/or in that of ‘Descriptive Identification’ (<did>), or perhaps in the ID attribute of the highest element, <c>? Further, there could be at least three identifiers to juggle with: the name of the file, the identifier of the file (allocated to it in the DDI-Codebook file), and the data archive’s own identifier for each file.

This is one of those situations where ruling out technical questions shows that such questions bear larger organizational implications. The threat of ‘identifier overload’[17] forces information systems engineers and decision-makers to determine what roles exactly the EAD files will fulfill on the one hand and what other purposes the DDI files will serve on the other. In a way, it also broaches upon the question of the scale of the future institution: will it be only a branch of the State Archives, which will simply draw upon the existing practices for identifier allocation, or will it be a larger entity with needs such as that of a specific policy for identifier generation and management? Perhaps an exploratory study of the technical implications of the SODA project was required before this question could come to light, but it must now be answered before a definitive technical solution can be implemented.

Distributing the information

The fact that DDI-Codebook and EAD have their own inner logic transpires in how certain elements mapped well yet required that the various items of information sometimes take very different paths. For instance, DDI’s <holdings> (‘Holdings Information’), with its four specific attributes, ‘location’, ‘callno’, ‘URI’ and ‘media’, contains valuable information from the standpoint of EAD. However, it is distributed as follows:

DDI	EAD
<holdings> ‘Holdings Information’	archdesc/custodhist or archdesc/repository/corpname (+ <address>)
<holdings> ATT location	
<holdings> ATT callno	archdesc/dsc/c/did/container or archdesc/dsc/c/did/unitid (ATT ID)
<holdings> ATT URI	eadheader/eadid ATT URN or URL
<holdings> ATT media	archdesc/phystech/genreform

Table 3. Distribution of the information contained in <holdings> into different EAD endpoints.

As illustrated, the information that can be encoded within one single element along with its attributes in DDI has to be reallocated in possibly five different elements and/or attributes within all three main wrappers of EAD, <eadheader>, <archdesc> and <dsc>. The element <holdings> itself or its location attribute seem to be where the name of the institution responsible for giving access to the dataset is most often encoded. If it gives the name of the ‘home institution’, i.e. the university from which the researchers behind the dataset hail, then its content should be transferred into the EAD section ‘Custodial History’ (<custodhist>); if, on the other hand, it already records the name of the data archive to which it is destined, then its content should be transferred to the key EAD element <repository>, possibly followed by an <address> wrapper composed of <addressline> elements in case a postal address was also encoded. The call number (ATT callno) is transferred either to the ‘Container’ (<container>) element in EAD, or to the <unitid> element, or to that element’s ID attribute. The EAD 2002 Tag Library advises one of the latter two choices, acknowledging however that different practices coexist (SAA, 2002: 80). Next, the content of the URI attribute of <holdings> is copied over to either ATT URN or ATT URL of <eadid>. Finally, EAD's 'Genre / Physical Characteristics' (<genreform>) element, which is meant to record 'the types of material being described, by naming the style or technique of their intellectual content (genre); order of information or object function (form); and physical characteristics' (SAA, 2002: 150) is the perfect equivalent of <holdings>' media attribute.

We ought to point out that the abovementioned problem of 'identifier overload' might occur with the URI attribute of <holdings>. If the attribute contains a URI that is not a URL hyperlink, then the question of identifier redundancy arises once more. But if, as was often observed in the DDI-Codebook corpus, the attribute simply contains a URL hyperlink that leads to the Web page of the home institution of the researchers who authored the study, then we need only transfer said hyperlink to 'Custodial History' and make it dynamic with the href attribute of the 'External Reference' (<extref>) element in EAD.

While the various paths that each item of information must follow during the transfer might appear sprawling overall, the information nevertheless finds conveniently close equivalents in terms of XML elements according to the logic both of the source language, DDI, and the language of destination, EAD. However, such scattering of the source information shows that the rationale of the mapping is a one-way transfer of information. It is unlikely that DDI files can be reconstructed from the EAD files simply by reverting the 'orientation' of the mapping, hence the need to preserve the DDI files in spite of their conversion towards EAD.

Conclusion and future work

The potential inclusion of the State Archives of Belgium into the future data archive for the social sciences will require the harnessing of the existing infrastructures for the needs of said data archive. Because the State Archives are a cultural heritage institution, one of their missions is to preserve and communicate the context of the data stored in their collections, a task that they share with data archives. In order to perform this mission for the future Belgian CESSDA Service Provider, a crosswalk between two metadata standards was undertaken and completed.

At this stage the mapping of elements and attributes only exists, so to speak, *in vitro*: it has been laid out in a spreadsheet with much documentation as well as technical recommendations in a dedicated column of an XLSX file; but it has yet to be integrated into a program and then a series of processes—the State Archives' 'pipeline'—that will ultimately produce the desired EAD document.

The following tasks will be undertaken in order to make the most of the crosswalk:

- First, the mapping will need to be adapted to the template EAD document laid out by the State Archives, thus providing an example of adaptation to local needs and therefore a proper use case;
- Second, Encoded Archival Context (EAC) will have to be included if possible, so as to make the transfer of data about the scientific contributors even more effective;

- Thirdly, extensive testing will have to be performed with real instances of DDI files to make sure that the transfer is feasible and that the resulting EAD documents conform to the rules of the standard as well as those of the State Archives;
- Finally the mapping will have to fit into the pipeline of the State Archives, along with secondary algorithms such as the one meant to identify (i.e. automatically generate identifiers for) the various elements that designate contributors to the original study.

The spreadsheet that contains the mapping will likely be published, hopefully in open access, at some later point. Additional research is required to determine what are the best formats (CSV, TSV, PDF, XSD…) for making the crosswalk readable and readily re-usable by other users. This will benefit the publication of the subsequent adaptation to the State Archives' EAD template, which will probably also take the form of a mapping.

The rationale of such a crosswalk does not rest solely in the re-use of computer infrastructure, nor does it merely consist in reducing expenditures. It also demonstrates the validity of the archivist in the organization chart of a data archive. In several CESSDA service providers, the focus was visibly put on academic qualifications in social sciences for the constitution of the workforces. This is quite logical, considering the data archives federated by CESSDA permit the re-use of research data from the social sciences, by social scientists, and for social scientists. Still, archivists seldom appear on the 'who's who' webpages of CESSDA service providers. Perhaps a DDI-EAD crosswalk might inspire some curators and show the way towards fruitful collaborations. Archivists are, after all, professional data and metadata managers, and they are used to handling highly heterogeneous archival objects. By working in close collaboration with the scientists from whom they receive data and for whom they will preserve and disseminate datasets, archivists can contribute to scientific research beyond their 'organic' attributions.

Although the SODA project officially began several years ago, the involvement of the State Archives is fairly recent, which is why many key questions still remain unanswered at this point: the exact business plan, cost model, and the roles that the State Archives will take on in the new service provider; the legal form of the future entity; the types of contracts and/or agreements that will bind the data purveyors (the universities and their researchers) to the data archive. Moreover, the transfer of (meta)data can only be done with actual (meta)data in the first place. The next grand step in terms of technical and organizational developments will entail gaining the trust of data providers and supplying the adequate tools and guidelines for documenting research data to them.

Acknowledgements

The Social Sciences Data Archive (SODA) project is funded by the Belgian Federal Public Planning Service Science Policy Office (BELSPO) under contract #FR/00/SO3.

I would like to express my thanks to my manager, D^{r} Rolande Depoortere, whose careful and trustful supervision made it possible for me to move about and explore different possibilities before setting one course of action. I also wish to thank my colleague, Samira Hajji, whose hard work helped better envision the legal form of the future Belgian service provider. My thanks go to the reviewers, who took the time to read my contribution and to formulate tactful suggestions for improvement. I am also grateful to Mari Kleemola and the CESSDA Metadata Management (CMM) working group, who showed kind interest into my work and who provided some valuable feedback, especially Irena Vipavc Brvar and Wolfgang Zenk-Möltgen. Finally, I wish to thank everyone at the EDDI 17 conference who attended and gave feedback after my presentation, especially Mari Kleemola, Franck Cotton, and Joachim Wackerow, as well as the organizing committee of the Swiss Centre of Expertise in the Social Sciences (FORS) at large.

References

AKÇEŞME, Banu, BAKTIR, Hasan and STEELE, Eugene (eds.) (2016) *Interdisciplinarity, Multidisciplinarity and Transdisciplinarity in Humanities*, Newcastle upon Tyne, Cambridge Scholars.

ALLISON-BUNNELL, Jodi (2016) 'Review of Encoded Archival Description Tag Library – Version EAD3', *Journal of Western Archives*, vol. 7, no. 1, article 6.

BOYDENS, Isabelle (2011) 'Strategic Issues Relating to Data Quality for E-Government: Learning from an Approach Adopted in Belgium'. In: ASSAR, Saïd, BOUGHZALA, Imed and BOYDENS, Isabelle (eds.) *Practical Studies in E-Government: Best Practices from Around the World*, New York, Springer, pp. 113-130, https://doi.org/10.1007/978-1-4419-7533-1_7.

CAPLAN, Priscilla (2003) *Metadata Fundamentals for All Librarians*, Chicago, American Library Association.

CESSDA ERIC [Consortium of European Social Science Data Archives European Research Infrastructure Consortium] (2017) CESSDA ERIC – History, [Online], Available: https://www.cessda.eu/About/History [18 May 2018].

CMM Group [CESSDA Metadata Management Working Group] (2016) CESSDA Service Providers' Metadata Practices: Standards, Controlled Vocabularies and Requirements for the CESSDA Portfolio. CESSDA Metadata Management Project Combined Deliverable D1 & D2, [Online], Available: https://www.cessda.eu/content/download/834/7776/file/CMM_ServiceProvidersMetadataPractices_2016.pdf [18 May 2018].

DDI Alliance (2017a) History of the Standard, [Online], Available: https://www.ddialliance.org/what/history.html [18 May 2018].

DDI Alliance (2017b) DDI Specification, [Online], Available: http://www.ddialliance.org/Specification/ [18 May 2018].

DDI Alliance (2017c) Explore Documentation, [Online], Available: https://www.ddialliance.org/explore-documentation [18 May 2018].

DDI Alliance (2017d) DDI Codebook 2.1, [Online], Available: http://www.ddialliance.org/Specification/DDI-Codebook/2.1/ [18 May 2018].

DDI Alliance (2017e) DDI-Codebook, [Online], Available: http://www.ddialliance.org/Specification/DDI-Codebook/ [18 May 2018].

DE FLORIO, Vincenzo (2009) *Application-Layer Fault-Tolerance Protocols*, Hershey (PA), Information Science Reference.

DESCHOUWER, Kris (2005) 'Kingdom of Belgium'. In: KINCAID, John and TARR, G. Alan (eds.) *Constitutional Origins, Structure, and Change in Federal Countries: Volume I. A Global Dialogue on Federalism*, Montreal, McGill-Queen's UP, pp. 48-75.

Doorn, Peter and Tjalsma, Heiko (2007) ‘Introduction: Archiving Research Data’, *Archival Science*, vol. 7, no. 1, pp. 1-20, https://doi.org/10.1007/s10502-007-9054-6.

Dow, Elizabeth H. (2009) ‘Encoded Archival Description As a Halfway Technology’, *Journal of Archival Organization*, vol. 7, no. 3, pp. 108-115, https://doi.org/10.1080/15332740903117701.

Dryden, John. (2010) ‘A Structure Standard for Archival Context: EAC-CPF Is Here’, *Journal of Archival Organization*, vol. 8, no. 2, pp. 160-163, https://doi.org/10.1080/15332748.2010.513325.

Emory U [Emory University, Libraries & Information Technology] (2017) Crosswalk of Core Metadata: Emory Core Metadata Mapping to Selected Standards and Systems, [Online], Available: http://metadata.emory.edu/guidelines/descriptive/crosswalk.html [18 May 2018].

Fox, Michael (2005) ‘Professional Training for Encoded Archival Description in Europe’, *Journal of Archival Organization*, vol. 3, nos. 2-3, pp. 71-82, https://doi.org/10.1300/J201v03n02_06.

Francisco-Revilla, Luis, Trace, Ciaran B., Haoyang, Li and Buchanan, Sarah A. (2014) ‘Encoded Archival Description: Data Quality and Analysis’, *Proceedings of the American Society for Information Science and Technology*, vol. 51, no. 1, pp. 1-10, https://doi.org/10.1002/meet.2014.14505101043.

Gaitanou, Panorea, Bountouri, Lina and Gergatsoulis, Manolis (2012) ‘Automatic Generation of Crosswalks Through CIDOC CRM’, *Metadata and Semantics Research: Proceedings of the 6th Research Conference, MTSR 2012*, Berlin, Springer, pp. 264-276, https://doi.org/10.1007/978-3-642-35233-1_26.

Huddleston, Rob (2008) *XML: Your Visual Blueprint™ for Building Expert Web Sites With XML, CSS, XHTML, and XSLT*, Hoboken (NJ), Wiley.

ICA [International Council on Archives] (2000) Z695.2.I83 2000 *ISAD(G): General International Standard Archival Description*, 2nd edition, Ottawa, International Council on Archives.

Lawson, Gary (2004) *Federal Administrative Law*, 3rd edition, St. Paul (MN), West Academic.

Marker, Hans Jørgen (2013) ‘Data Documentation, Access, and Dissemination Systems’. In: Kleiner, Brian, Renschler, Isabelle, Wernli, Boris, Farago, Peter and Joye, Dominique (eds.) *Understanding Research Infrastructures in the Social Sciences*, Zurich, Seismo, pp. 39-46.

Meissner, Dennis (1997) ‘First Things First: Reengineering Finding Aids for Implementation of EAD’, *The American Archivist*, vol. 60, no. 4, pp. 372-387, https://doi.org/10.17723/aarc.60.4.6405275227647220.

Méndez, Eva and van Hooland, Seth (2014) ‘Metadata Typology and Metadata Uses’. In: Sicilia, Miguel-Angel (ed.) *Handbook of Metadata, Semantics and Ontologies*, Singapore, World Scientific, pp. 9-40, https://doi.org/10.1142/9789812836304_0002.

Merriam-Webster (2018) Dictionary, [Online], Availaible: https://www.merriam-webster.com/ [18 May 2018].

Morris, Sammie L. and Rose, Shirley K. (2010) ‘Invisible Hands: Recognizing Archivists’ Work to Make Records Accessible’. In: Ramsey, Alexis E., Sharer, Wendy B., L’Eplattenier, Barbara and Mastrangelo, Lisa S. (eds.) *Working in the Archives: Practical Research Methods for Rhetoric and Composition*, Carbondale, Southern Illinois UP, pp. 51-78.

Oliver, Gillian and Harvey, Ross (2016) *Digital Curation*, 2nd edition, [no place], American Library Association.

Pearce-Moses, Richard (2005) *A Glossary of Archival and Records Terminology*, Chicago (IL), The Society of American Archivists.

PILGRIM, Mark (2009-2011) Everything You Know About XHTML Is Wrong, [Online], Available: http://diveintohtml5.info/past.html [18 May 2018].

PITTI, Daniel V. (1997) 'Encoded Archival Description: The Development of an Encoding Standard for Archival Finding Aids', *The American Archivist*, vol. 60, no. 3, pp. 268-283, https://doi.org/10.17723/aarc.60.3.f5102tt644q123lx.

PITTI, Daniel V. (1999) 'Encoded Archival Description: An Introduction and Overview', *D-Lib*, vol. 5, no. 11, https://doi.org/10.1080/13614579909516936.

RASMUSSEN, Karsten Boye and BLANK, Grant (2007) 'The Data Documentation Initiative: A Preservation Standard for Research', *Archival Science*, vol. 7, no. 1, pp. 55-71, https://doi.org/10.1007/s10502-006-9036-0.

RAŢĂ, Georgeta, ARSLAN, Hasan, RUNCAN, Patricia-Luciana and AKDEMIR, Ali (eds.) (2014) *Interdisciplinary Perspectives on Social Sciences*, Newcastle upon Tyne, Cambridge Scholars.

RIGGS, Michelle (2005) 'The Correlation of Archival Education and Job Requirements Since the Advent of Encoded Archival Description', *Journal of Archival Organization*, vol. 3, no. 1, pp. 61-79, https://doi.org/10.1300/J201v03n01_06.

SAA [Society of American Archivists] (2002) *Encoded Archival Description Tag Library*, Chicago (IL), The Society of American Archivists.

SCHOUPS, Inge, LOBET-MARIS, Claire, LAURENT, Véronique, POULLET, Yves and LEFEVER, Nathalie (2008) *SODA Prospect Study: Feasibility Study of a Computerized Archive Service for the Social Sciences (SODA)* [*Studie SODA: Haalbaarheid van een data-archief voor de sociale wetenschappen / Étude prospect SODA : Faisabilité d'un service d'archivage de données pour les sciences sociales*], Expertisecentrum DAVID, Cellule interdisciplinaire de Technology Assessment FUNDP and Centre de recherche Informatique et Droit FUNDP, report without a number.

SHAW, Elizabeth J. (2001) 'Rethinking EAD: Balancing Flexibility and Interoperability', *New Review of Information Networking*, vol. 7, no. 1, pp. 117-131, https://doi.org/10.1080/13614570109516972.

STEVENSON, Jane (2016) 'Encoded Archival Description Tag Library, Version EAD3', *Archives and Records*, vol. 37, no. 2, pp. 257-260, https://doi.org/10.1080/23257962.2016.1220362.

UNESCO (2015) *Interoperability and Retrieval*, Paris, UNESCO, [Online], Available: http://unesdoc.unesco.org/images/0023/002321/232199E.pdf [18 May 2018].

VAN HOOLAND, Seth and VERBORGH, Ruben (2014) *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata*, London (UK), Facet.

WACKEROW, Joachim (2017) *Current Status of DDI 4*. [Presentation] 9th Annual European DDI User Conference (EDDI17), Swiss Centre of Expertise in the Social Sciences (FORS), 6th December.

WACKEROW, Joachim and VARDIGAN, Mary (2013) 'An Established International Metadata Standard: The Data Documentation Initiative (DDI)'. In: KLEINER, Brian, RENSCHLER, Isabelle, WERNLI, Boris, FARAGO, Peter and JOYE, Dominique (eds.) *Understanding Research Infrastructures in the Social Sciences*, Zurich, Seismo, pp. 158-167.

YACO, Sonia (2008) 'It's Complicated: Barriers to EAD Implementation', *The American Archivist*, vol. 71, no. 2, pp. 456-475, https://doi.org/10.17723/aarc.71.2.678t26623402p552.

YAKEL, Elizabeth and KIM, Jihyun (2005) 'Adoption and Diffusion of Encoded Archival Description', *Journal of the American Society for Information Science and Technology*, vol. 56, no. 13, pp. 1427-1437, https://doi.org/10.1002/asi.20236.

End-notes

[1] Benjamin Peuch is a contractual researcher at the State Archives of Belgium. A literature and information science graduate, he has been working on the Social Sciences Data Archive (SODA) project since April 2017 and can be reached by email at benjamin.peuch@arch.be or benjamin.peuch@gmail.com.

[2] The project was previously known as the 'Social Sciences *and Humanities* Data Archive (SOHDA) project' and has been recently renamed.

[3] The name of this university literally means 'Free University of Brussels'. However, it is preferable not to translate it because there is another university in Brussels whose name also literally translates like so: the French-speaking *Université libre de Bruxelles*.

[4] On interdisciplinarity, see for example Raţă et al, 2014 and Akçeşme, Baktır and Steele, 2016.

[5] For more information on the institutional landscape of Belgium, see Deschouwer, 2005.

[6] 'ISAD(G)' stands for 'General International Standard Archival Description'. The standard was conceived by the International Council on Archives and published in 1988 (ICA, 2000).

[7] This is likely to be determined to a large extent by the kind of data management software, such as Nesstar or Dataverse, that will be chosen for the Belgian data archive. A review has yet to be conducted before the choice is made.

[8] 'A common criticism of the W3C Schema language is that it is too complex and tries to do too many things' (Huddleston, 2008: 52).

[9] As a form of anecdotal evidence, in 2016, knowledge of EAD was required for the current job position of the author at the State Archives of Belgium.

[10] 'Metadata demands knowledge. Obviously knowledge about the data and the research being carried out is needed, but the further cost is that it demands knowledge of metadata description, the structure and content of metadata and some technical insight regarding the practical arrangement of the metadata' (Rasmussen and Blank, 2007: 60).

[11] The Board of Tea Appeals, abolished in 1996, was tasked with 'adjudicat[ing] the claims of tea importers whose products were denied entry into the United States by federal tea-tasters' (Lawson, 2004: 7).

[12] A very bad pun to make about the core structure of DDI-Lifecycle would be to say that it is exceptionally ‘fragmented’.

[13] Moreover, the DDI Alliance announced at the 9th EDDI conference that they would continue to support DDI 2 and DDI 3 even with the imminent release of DDI 4 (Wackerow, 2017).

[14] This file was produced as a small test case several years before the State Archives became actively involved in the SO(H)DA project.

[15] Furthermore, experience shows that, oftentimes, in the few cases when information about the various contributors is encoded into the ‘Document Description’ section, it turns out that the researchers themselves encoded their own codebooks.

[16] This is not entirely true when it comes to EAD 2002 in particular: in this version, the identity and the roles played by the people who contributed to the creation of the EAD file (as opposed to the usually paper-based finding aid) are to be encoded in the ‘Creation’ (<creation>) section under ‘EAD Header’ (<eadheader>), and <author> may not be used in <creation>. This however was solved in EAD 3, where <creation> was replaced by the ‘Maintenance Event’ (<maintenanceevent>) section, which allows for the use of a personal tag, ‘Agent’ (<agent>).

[17] On the problem of identifier overload, see Boydens, 2011: 122-123.