Building Bridges: Preparing the Automated Transfer of Metadata for an Upcoming Data Archive in Belgium by Mapping the Metadata Standards of Archives and Social Sciences

Benjamin Peuch

BUILDING BRIDGES: PREPARING THE AUTOMATED TRANSFER OF METADATA FOR AN UPCOMING DATA ARCHIVE IN BELGIUM BY MAPPING THE METADATA STANDARDS OF ARCHIVES AND SOCIAL SCIENCES

Benjamin PEUCH

Introduction

Since the 1970s European countries have set up scientific infrastructures in the form of data archives. Those particular kinds of archives are meant to preserve and provide access to research data produced by scientific institutions. Like regular archives, materials created by scientists are subject to deterioration overtime and they bear tremendous historic value. That is why countries have established specialized repositories since then. Some have a wide scope of interest; others rather focus on particular domains or disciplines.

Over time certain data archives enjoyed more development than others for a number of reasons. These include the economic situation of their home countries, as well as how open to collaboration data producers (researchers, research centers, and universities) are. Just like the lack of an archive endangers the documents which ought to help write the history of a country, the lack of a data archive implies that the history and even the integrity of science is under threat. The consequences can be especially drastic for data-intensive fields such as the social sciences. Demographers and sociologists collect large amounts of data, usually in predetermined/numerical form, from which they derive statistics and trends (i.e. quantitative data), but also oral or written answers from respondents to questionnaires or during interviews in more open-ended formats, which can be less easily aggregated (i.e. qualitative data).

Archives preserve historic documents so that historians may re-use them in the future; the same logic applies to data archives. Re-use of research data, also known as 'secondary data analysis', is critical for data-driven scientific fields such as the social sciences because it enables the replication of studies. It also allows researchers to skip the stage of data collection and make use of previously collected data for their own research purposes so that they can apply their own

methodologies to and test their own hypotheses with those data. In addition, the possibility of re-use provides invaluable help to young researchers[1].

Because social science data are often collected thanks to public funds, society at large should benefit from them; such is the open science philosophy promoted by the European Union[2]. But setting up a data archive infrastructure poses several challenges, not least of all the processing of scientific data, which are often complex, support- or software-dependent digital objects.

That is why the SODA project (Social Sciences Data Archive), a pilot study for a Belgian data archive for the social sciences, brings together representatives of the social science community as well as professional archivists, who pool their respective expertise in order to create Belgium's very own social science data archive[3]. The ensuing entity will integrate the Consortium of European Social Science Data Archives (CESSDA) and serve as a central platform for archiving and disseminating research data, following in the footsteps of such predecessors as GESIS (Germany)[4], DANS (Netherlands)[5], or UKDS (United Kingdom)[6].

Many legal and organizational questions remain unanswered at this point: What will be the legal form of this new entity? Can it seamlessly integrate representatives of Belgium's various political bodies, notably the linguistic Communities, the Regions, and the Federal State? Where and how exactly will the data be stored? While the project team investigates such issues, this paper will present an early technical realization which relies on the assumption that the future entity will somehow integrate both the State Archives and the Belgian universities, the former as managers and preservers of data, the latter as providers and consumers of data. In this scenario, part of the State Archives' infrastructure for the preservation and dissemination of data will be re-used for the needs of the future Belgian data archive[7].

Preserving and disseminating archival objects is made possible by documenting them following well-established metadata standards. We archivists rely on the

---

[1] K. B. Rasmussen, "Social Science Metadata and the Foundations of the DDI", in *IASSIST Quarterly*, 37. 1, 2014, p. 29-30.

[2] SPARC Europe, "PSI Directive Compromise Agrees to Make Publicly-Funded Research Data Open by Default", in *SPARC Europe.org: Setting the Default to Open*, 27 February 2019, https://sparceurope.org/psi_researchdata_openbydefault/, accessed 12 March 2019.

[3] The SODA project members are the State Archives of Belgium, the *Université catholique de Louvain* and the *Vrije Universiteit Brussel*.

[4] GESIS is part of the Leibniz Institute for the Social Sciences: https://www.gesis.org.

[5] DANS are the Data Archiving and Networked Services: https://dans.knaw.nl/en.

[6] UKDS is the United Kingdom Data Service: http://ukdataservice.ac.uk.

[7] An increasingly popular method for creating research institutions consists in partially relying on already existing infrastructures. See: B. Habert and C. Huc, "Building Together Digital Archives for Research in Social Sciences and Humanities", in *Social Science Information*, 49, 2010, HAL pagination: p. 9, https://doi.org/10.1177%2F0539018410371570.

Encoded Archival Description (EAD) to produce machine-readable versions of finding aids while social scientists work with the Data Documentation Initiative (DDI) standard to document their datasets. Like most large archive repositories, the State Archives' infrastructure for metadata management revolves around EAD. Utilizing this infrastructure for the dissemination of social science data would save a fair amount of resources both in terms of time and finances, but part of the DDI must be transferred over to EAD to this end. Since both metadata standards are based on the eXtensible Markup Language (XML), a crosswalk between them was successfully developed. The subsequent mapping of DDI and EAD tags has been presented earlier to the social science community with a strong technical focus[8]. This paper presents other aspects of the mapping but also delves longer on its rationale by highlighting the partnership between social scientists and archivists that lies at the heart of the SODA project.

I will now present the context of the SODA project and explain why this project benefits both the State Archives and Belgian researchers in social sciences. A few statements will follow with regard to the particularities of social science research data and how they differ with traditional archive records. Next will come a brief presentation DDI and a comparison with EAD. Finally, the crosswalk will be presented with special focus on the choices that determined the method for mapping both metadata standards and by underscoring how certain challenges could be resolved.

A Good Deal for Everyone: What Archives and Social Scientists Stand to Gain Together

One might not see at first what exactly archivists and social scientists have in common in terms of interest. The former usually have a historical background and resort to the analytical methods of the humanities, most prominently close reading; the latter, for the most part, are focused on quantitative analysis of numerical data from which sweeping conclusions can be drawn. Some historians have picked up on the digital humanities trend and use data mining software, but it seems that most still prefer the good old pen and paper, while most scholars trained in the social sciences have had to learn how to use tools for statistical computing such as SPSS or R.

In actual fact, the two disciplines are evolving in ways that make them criss-cross more and more. For one thing, digital humanities are growing rapidly, supported as they are by international networks like the Digital Research Infrastructure for the

[8] This work was first presented in B. Peuch, "Elaborating a Crosswalk Between Data Documentation Initiative (DDI) and Encoded Archival Description (EAD) for an Emerging Data Archive Service Provider", in *IASSIST Quarterly*, 42. 2, 2018, p. 1-24, https://doi.org/10.29173/iq924. The present paper focusses on the implications of social science research data for a traditional archive and expands on the initial findings.

Arts and Humanities (DARIAH)[9]. Historians and archivists draw more and more on aggregation techniques to make sense out of the large volumes of records that they curate. This likely stems in part from the newfound popularity of the 'More Product, Less Process' (MPLP) philosophy[10], which arose from the realization that archivists could no longer dedicate as much time as they used to process the ever-growing backlog of archives. More and more historians dare to challenge the methodological assumptions of their curriculum and experiment with tools and data.

The Specificities of Social Sciences

The term 'social sciences' designates a wide spectrum of disciplines and sub-disciplines. The newfound popularity of interdisciplinarity further blurs the lines between one method of doing social science research and another. Studies and methods can also be distinguished with the help of the quantitative/ qualitative paradigm, but here too it is more of a spectrum than of a clear-cut divide. That is to say that, because I will focus on a particular facet of social sciences, namely quantitative, data-intensive varieties, I am well aware that I do not encompass all of the social sciences.

Quantitative social sciences involve collecting large quantities of data that easily translates in numerical, aggregatable figures. If, for example, you want to conduct a study on the wellbeing of a certain population, students in an academic library for instance, and you ask them to sit for interviews and ask them open-ended questions, or if you request that they fill diaries in which they record their emotions and experiences, you are not doing the same thing as when you ask someone: 'What is your opinion on Donald Trump? Very favorable, favorable, no opinion, unfavorable, or very unfavorable?' or 'How many times have you moved house in the past 10 years?' In the second case, you can draw statistics from the collection of answers that you recorded.

In the first situation, the investigator will have to go through a lot of material, written or audio, become very familiar with it, and infer from the whole broad themes that could correspond to actual social tendencies. Depending on their method, the investigator might actually want to identify peculiarities and isolated elements that, while individual, are nonetheless significant with regard to certain theories. In other words, with this kind of approach, it is a human who performs the analysis, for the early stages of ingesting the materials to the later ones when conclusions can be drawn. In the second situation however, since the data have a much more consistent shape — proposed answers, numerical data… — they can

[9] See also for a recent publication on the subject an introduction to the digital humanities for students and researchers: S. Van Hooland, F. Gillet, S. Hengchen and M. De Wilde, *Introductions aux humanités numériques : Méthodes et pratiques*, Louvain-la-Neuve, 2016 (Méthodes en sciences humaines).

[10] M. A. Greene and D. Meissner, "More Product, Less Process: Revamping Traditional Archival Processing", in *The American Archivist*, 68, 2005, p. 208-263.

be combined together instantly with modern software. In other words, with that kind of approach, the first phases of the analysis are supported by computer programs.

Naturally, historians resort to numerical data of various kinds. But compare the toil of the scholar who scrutinizes large volumes of archival materials, who rifles through old documents that computers often cannot process, who dives into archive fonds to form a general impression of this collection in order to write a synthesis about a certain archive producer — compare this kind of work and the two methods previously described. It does seem that, altogether, the work of the historian bears more similarities with that of the qualitative methods in social sciences than with more quantitative methods.

This distinction is worth stressing out because quantitative data nowadays constitute the bulk of social sciences[11]. Moreover, while computer software can accelerate the process of analysis by amalgamating consistent data, quantitative data have their own complexity. For one thing, researchers often try to get as many data from their respondents as they can so as to be able to perform large-scale cross-reference analyses. Imagine that you want to learn more about the people who live in your city: you will want to look at features such as age, family composition, professional situation, health record, date of arrival… and one of these can come to the fore during the secondary analysis and yield results that will cast each time a different light on your working hypotheses.

The point of this explanation is to show that, because the methods of historians and archivists on the one hand, and that of researchers in quantitative social sciences on the other hand differ so vastly, the data that they produce likewise differ formally and fundamentally, and so do the metadata that they are expected to provide to document those data. Studies and research projects whose methodologies involve collecting a lot of numerical data are most of the time documented with 'books of codes', or codebooks, which serve as 'dictionaries of codes'[12]. Researchers handle the aforementioned demographic features — age, family composition… — by storing them in spreadsheet/statistical analysis software programs, such as Microsoft Excel or SPSS Statistics. In this context, the content of each cell is called a 'variable'[13] and the names of the columns are the 'categories'. The distinction matters because, when you are not a specialist in social sciences and you do not handle data on a daily basis, you are usually presented with 'data' like the following:

---

11 M. Vardigan and C. Whiteman, "ICPSR Meets OAIS: Applying the OAIS Reference Model to the Social Science Archive Context", in *Archival Science*, 7, 2007, p. 76, https://doi.org/10.1007/s10502-006-9037-z.

12 K. B. Rasmussen and G. Blank, "The Data Documentation Initiative: A Preservation Standard for Research", in *Archival Science*, 7, 2007, p. 309. https://doi.org/10.1007/s10502-006-9036-0.

13 Rasmussen and Blank, "The Data Documentation Initiative", p. 56-57.

Ages of Smokers	16-21	22-30	30-40	40-50	50-60	60-70
Number of individuals	167	357	835	1,037	569	324

Table 1. An example of data organization in a chart from a fictional dataset.

While such a chart certainly presents data, it is a case of aggregated data. This implies that something sits behind those neat numerical syntheses. Media often show the trees, but there is always a forest in the background. In our fictional example shown in Table 1, we have surveyed at least 3,289 individuals (and those are only the smokers; we are not taking the non-smokers into account). This means that, theoretically, there exist record cards like this one…

Record number	00357
Age	18
Gender	F
Monthly income (gross)	n/a
District	Farnham
Family situation	Student housing, 2 flatmates
Medical history	Asthmatic

Table 2. An example of an individual record from a fictional dataset[14].

… at least 3,289 times! — even if some of the data are missing (in case a respondent was not able or willing to provide a certain piece of information for example). That is why it is vital to distinguish the encompassing 'categories' from the very individual, and often very numerous, data 'variables'.

Another difficulty is that categories are seldom put in such common, human-readable terms as 'Age', 'Gender', etc. That is because most researchers in quantitative social sciences use powerful software programs which were specifically designed to handle, compile and manipulate large volumes of complex data. And because there are so many subdivisions within the data, and because researchers

[14] Notice how the name of the respondent is not recorded. It likely was at one point during the study, mainly for practical reasons (especially for follow-up procedures, in case it becomes necessary to contact the person again) but it is always imperative at the end of a research in social studies to anonymize respondents. Even so, it often remains possible to cross-reference data and thus identify individuals, which is why the curation of social science data raises endless issues of privacy protection.

need to be able to aggregate the data in various ways while keeping track of everything — say for instance that you want to know exactly how many individuals between the ages of 30 and 50 who have lung-related medical conditions and who live in two particular districts are smokers — these programs need to record all this information and categorize it in a systematic and unambiguous manner. That is why they assign generic identifiers to these categories such as 'V12' or 'Q0A'.

These are the 'codes' which the codebooks help elucidate. Datasets in the social sciences usually feature one or several spreadsheet files with long, cryptic lists of numeric values and no-less-cryptic names of columns such as these. Without the 'dictionary of codes' that is the codebook, it is usually impossible for users to work out their meaning. Without proper context it just looks like a nonsensical jumble of numbers and letters, as shown on the next page.

In this example, the first column seems intelligible enough: as indicated by the title, 'LANGUAGE', it is a list of varieties of American Indian languages spoken on American soil. But there is no telling what 'POP', 'VAPOP', 'MVAPOP', 'VACIT', and all the others mean.

Obviously, a lot of documentation and study hours are required to grasp the general context in which those data were collected as well as the particular meanings of those rows and columns. That is why the social sciences have adopted strict and complex metadata standards such as the Data Documentation Initiative (DDI).

The Data Documentation Initiative: A Standard for Quantitative Social Science Datasets

As noted by Karsten Boye Rasmussen and Grant Blank:

'Analysing undocumented data is impossible […]. But even with documentation the process of analysis is often difficult (e.g., the user must be able to understand the jargon of the documentation), error prone (e.g., the documentation might be imperfect, and/or the user might misunderstand the documentation), and time-consuming (e.g., users have to familiarize themselves with the documentation and the software). Providing a standard format for machine-readable metadata can reduce errors and simplify analysis. [For] these reasons, the DDI is intended to become the cornerstone of many scientific infrastructure projects.'[15]

[15] Rasmussen and Blank, "The Data Documentation Initiative", p. 58.

J	K	L	M	N	O	P	Q	R	S	T	U
LANGUAGE	POP	MPOP	VAPOP	MVAPOP	VACIT	MVACIT	VACLEP	MVACLEP	ILLIT	MILLIT	LEPPCT
Total Persons	308745540	0	234564070	0	21677005	57338	9826340	29312	951505	8953	
Hispanic	50477595	0	33346705	0	20879435	42438	4959140	20291	611665	6901	23209
American Indian alone or in combination	5220580	0	3569355	0	3269895	3618	181140	3785	25445	1736	848
American Indian (Apache)	111810	0	75945	0	66615	136	4505	245	505	69	21
American Indian (Arapaho)	10860	0	6680	0	6465	33	180	76	35	45	1
American Indian (Blackfeet)	105305	0	73850	0	71550	115	1565	198	110	30	7
American Indian (Canadian and French American Indian)	14820	0	11025	0	9570	212	350	110	35	40	2
American Indian (Central American Indian)	27845	0	20615	0	10935	1009	2630	651	315	216	12
American Indian (Cherokee)	819105	0	591555	0	580505	204	8890	421	655	70	42
American Indian (Cheyenne)	19050	0	12200	0	11865	44	270	81	15	24	1
American Indian (Chickasaw)	52280	0	34580	0	34170	29	335	79	40	37	2
American Indian (Chippewa)	170740	0	116095	0	114430	135	1455	235	70	47	7
American Indian (Choctaw)	195765	0	133745	0	131880	65	2535	375	265	96	12
American Indian (Colville)	10550	0	7035	0	6905	24	45	37	4	15	0
American Indian (Comanche)	23330	0	16020	0	15015	56	455	122	60	36	2
American Indian (Cree)	7985	0	5850	0	5485	131	95	54	4	24	0
American Indian (Creek)	88330	0	59755	0	59040	41	1010	179	55	34	5
American Indian (Crow)	15205	0	9630	0	9465	31	175	122	10	23	1
American Indian (Delaware)	18265	0	13555	0	13335	67	255	108	10	31	1
American Indian (Hopi)	18325	0	12075	0	11400	39	1365	252	95	114	6
American Indian (Houma)	10770	0	7020	0	6965	36	280	143	75	71	1
American Indian (Iroquois)	81000	0	58270	0	56845	160	955	151	60	21	4
American Indian (Kiowa)	13785	0	9340	0	9080	46	235	104	15	31	1
American Indian (Lumbee)	73690	0	50435	0	50210	112	305	138	10	39	1
American Indian (Menominee)	11135	0	7350	0	7260	23	190	132	4	35	1
American Indian (Mexican American Indian)	175495	0	117905	0	62040	2022	14315	1109	2415	410	67
American Indian (Navajo)	332130	0	215515	0	210715	121	36760	1536	9545	767	172
American Indian (Osage)	18575	0	12885	0	12715	32	145	69	4	21	1

Table 3. An excerpt from the United States' 2010 Decennial Census[16].

[16] United States Census Bureau, "Section 203 Determination Dataset — Census 2010", in *Decennial Census of Population and Housing*, 2010, available from: https://www.census.gov/programs-surveys/decennial-census/data/datasets.2010.html [Accessed 31st July 2018].

DDI has two branches: DDI-Codebook, from version 1 to version 2.5, and DDI-Lifecycle, from version 3 to version 3.2. DDI version 1 was published in 2000, and the distinction between Codebook and Lifecycle was introduced when DDI-Lifecycle 3.0 was published in 2008. Codebook and Lifecycle greatly differ because they do not share the same conceptual foundation: Codebook imitates the structure of a dictionary of codes while Lifecycle envisions the data in their whole life cycle, even before they were first collected[17].

Because DDI-Codebook and DDI-Lifecycle differ so much, a choice had to be made before mapping one of the two towards EAD. Back before I made my decision, I try to see how the data and metadata objects of social sciences on the one hand and archives on the other could connect. I arrived at the following preliminary mapping:

		Field of study	
		Social sciences	Archives (Archival science)
Data	Print	Datasets	Archive fonds
	Digital	Born-digital documents + scans	
Meta-data	Print	Codebooks	Finding aids
	Digital	Digitized codebooks + DDI	Digitized finding aids + EAD

Table 4. Parallels between the kinds of data and metadata used in the social sciences on the one hand and in archives on the other hand.

If the crosswalk was to make some sense, it would have to bridge two objects that are somewhat similar in shape and purpose. This seemed to be the case when laying things out in this manner: just like finding aids, codebooks are discrete, originally book-like documents that help readers appropriate source materials. A 'codebook' certainly felt closer to a finding aid, intuitively, than the general 'life cycle of research data', since the latter concept felt so abstract and, in itself, devoid of a stereotypical form. All in all, both DDI-Codebook and EAD are 'document-centric'[18].

[17] See the website of the DDI Alliance for more information about the different versions, their history, and their field level documentation: http://www.ddialliance.org/.

[18] The same term was used to characterize both metadata standards in two different papers: M. Vardigan, P. Heus and W. Thomas, "Data Documentation Initiative: Toward a Standard for the Social Sciences", *The International Journal of Digital Curation*, 1, 2008, p. 109; J. Riley and K. Shepherd, "A Brave New World: Archivists and Shareable Metadata", in *The American Archivist*, 72, 2009, p. 98, https://doi.org/10.17723/aarc.72.1.kl70j01223654874.

That being said, when you think about it, finding aids themselves do document the 'lifecycle of archives': they trace their origins — who produced them, where, when, why, and what do they consist of, what do they look like, what are they all about — up to the point where they were delivered to a repository — and what happened there and then too, how were they arranged, where are they stored now, whether we can consult them, etc. So then, it would not have been entirely inappropriate to draw parallels between EAD and DDI-Lifecycle in this respect.

But because a choice had to be made, and because, in the way both variants of DDI were structured, Codebook felt much simpler and easier to handle to a layperson than Lifecycle did, Codebook was chosen[19]. Furthermore, the latest version of DDI-Codebook, No. 2.5, was selected because it builds atop all of the other DDI-Codebook versions, so that a software program that can process DDI 2.5-compliant files can also process files that follow the rules of DDI version 1.0. Another argument is that Codebook is the version of DDI that can be automatically produced by Dataverse, a software program for depositing and accessing research data[20]. The SODA project team is currently running tests with Dataverse, which is a free, open-source program that could be used for the needs of the future Belgian data archive.

DDI-Codebook and EAD

DDI-Codebook is structured in five great sections. <codebook> – 'Document Description' contains information about the DDI file itself, including who created it, when, with what program, based on which codebook, and so on. <stdyDscr> – 'Study Description' contains information about how the study took place, which kind of study it was, which methods were used, who the people interviewed or the data sources consulted were, what data processing tools and functions were used, what was the timeframe, and so on. <fileDscr> – 'File Description' contains information about each individual file that constitute the dataset. <dataDscr> – 'Variable Description' contains information about the categories mentioned earlier and allows users to understand their meaning. Finally, <otherMat> – 'Other Study-Related Materials' contains extra information such as special sources that were consulted during the study, or materials which the authors advise to consult.

---

[19] Incidentally, it is commonly admitted in DDI-savvy circles that the vast majority of currently existing DDI files is still DDI-Codebook-compliant — quite simply because DDI-Codebook has been around for longer than DDI-Lifecycle. That being said, the Consortium of European Social Science Data Archives (CESSDA) pushes a lot for DDI-Lifecycle and various elements seem to indicate that the day of DDI-Codebook are numbered, at least on the long term. The author gathered this anecdotal evidence by attending such events as the 9th EDDI Conference in Lausanne, 2017, and the work sessions of the working group CESSDA Metadata Management (CMM) Phase 2.

[20] Dataverse was developed by the Institute for Quantitative Social Science (IQSS) at Harvard University, the Harvard University Library, and the Harvard University Information Technology organization.

At first sight, DDI-Codebook and EAD do not seem to have much in common. As we know, EAD is structured in three great sections: <eadheader>, which contains information about the EAD file itself; <archdesc>, which contains information about the archival materials of the fonds that the finding aid describes; and, within <archdesc>, <dsc>, the 'Description of Subordinate Components' or, in other words, the body of the finding aid, with its succession of titles and subtitles that help grasp the structure of a collection of archives. Three sections versus five, what can be done?

The question is, what exactly are we trying to map? What information elements do we need to transfer from one metadata file onto another?

Defining the Scope of the Mapping

It would not make much sense to want to transfer everything from one file to another since DDI, on the one hand, serves a purpose in the context of social sciences, and EAD, on the other hand, does so in the context of archives. As said earlier, there is a high chance that the future Belgian data archive will work within, or in close collaboration with, the State Archives of Belgium. If the State Archives could display information about the new data they are now responsible for in their online catalogue, this would represent a tremendous value-add in terms of collection development. A persistent challenge in the profession of archivist is to keep in mind that we do not — that we *cannot* — know what will stimulate the interest of the future historians[21], which is why selecting documents for preservation when we cannot afford to store everything is such a dilemma. That is why enriching the State Archives' collections with social science research data constitutes an investment for possible future cross-discipline investigations.

On the other hand, social scientists will have their own data access and discovery platform, which will directly process DDI. Therefore, only a certain core of elements from DDI needs to be fed into EAD so that we may utilize the State Archives' online catalogue and display social science metadata in it.

That being said, finding which elements of DDI would be interesting to map over to EAD was another challenge on its own. DDI-Codebook contains almost 2,000 elements. While the large sections can help when looking for specific information items, another problem was my lack of knowledge in social sciences. What is a cohort study? What distinguishes it from a longitudinal study? What does 'universe' mean in this context? How do you create an n-Cube?

---

[21] F. X. Blouin Jr and W. G. Rosenberg, *Processing the Past: Contesting Authorities in History and the Archives*, Oxford (UK) – New York (NY), 2011, p. 92; A. Flinn, "Chapter 6: Other Ways of Thinking, Other Ways of Being. Documenting the Margins and the Transitory: What to Preserve, How to Collect", in L. Craven (ed.), *What Are Archives? Cultural and Theoretical Perspectives: A Reader*, Abingdon (Oxon) – New York (NY), 2016 (1st ed. 2008), p. 110-111.

Those were the reasons why, although I knew only a subset of elements would have to be mapped over to EAD, I went through all of DDI-Codebook's elements and sought possible correspondences in EAD. The benefit from this approach was twofold: 1) Although this endeavor was time-consuming, I developed a great familiarity with DDI, and I can now use this knowledge for the SODA project by combining it with my knowledge of archives and historiography; 2) I made sure that no potentially interesting information element was left behind, and the ensuing crosswalk will be open for re-use once we publish it.

Results

It is difficult to discuss a more than 2,000-line-long list of items. However certain aspects of the mapping can be highlighted. A point of note for example is how, on several occasions, the structure of DDI enabled the mapping of whole section. A recurring group of elements in DDI-Codebook is <citation> – 'Bibliographic Citation', which contains information about a large variety of people who potentially contributed to the study and to the making of the dataset. <citation> can occur at five different locations in a DDI file, as there might be different contributors for different aspects of the study, respectively for the creation of the DDI file (<docDscr>), for the execution of the study and for other materials in direct relation to it (<stdyDscr>), for other materials that are not so directly related to the present study (<otherMat>), and for the authors of the original codebook (<docSrc> – 'Documentation Source' in <docDscr>).

The <citation> element contains many different wrappers and identifies a large amount of possible contributors, such as the producer, the distributor, the depositor, the 'authoring entity', the funding agency, the people responsible for each particular version of the dataset, the contact persons… But because this type of information always comes in a standardized manner, it could be easily mapped and distributed in specific EAD elements, namely various tags in the <eadheader> portion of EAD when the information concerns the creators of the DDI file, and in <archdesc> when it bears over the study investigators.

It is worth noting how the preoccupations of a community of users show in the information tools that they use. Not unlike EAD, DDI puts a lot of emphasis on the deposit process, with elements meant to document who deposited a dataset (<depositr>), what is that person or that institution's affiliation ('affiliation' attribute of <depositr>), when exactly the dataset was deposited (<depDate>)… Archivists base much of their work on the principle of provenance, so these elements naturally mapped well with the EAD tags <custodhist> – 'Custodial History' and <acqinfo> – 'Acquisition Information'.

Likewise, DDI distinguishes the 'Authoring Entity' (<AuthEnty>), the 'Producer' (<producer>), and the 'Distributor' (<distribtr>) among other entities who may hold responsibility over a dataset. This wide spectrum of potential contributors can

help in situations where responsibility cannot be uniformly allocated. As mentioned in the EAD 2002 tag library:

Although the repository providing intellectual access usually also has physical custody over the materials, this is not always the case. For example, an archives may assume responsibility for long-term intellectual access to electronic records, but the actual electronic data files or systems may continue to reside in the office where they were created and maintained, or they may be held for long-term storage by a unit such as a data library that is able to provide the appropriate technical facilities for storage and remounting. When it is clear that the physical custodian does not provide intellectual access, use <physloc> to identify the custodian and <repository> to designate the intellectual caretaker. When a distinction cannot be made, assume that the custodian of the physical objects also provides intellectual access to them and should be recognized as the <repository>[22].

And as noted for the <publisher> – 'Publisher' tag: 'Often this party is the same corporate body identified in the <repository> element in the finding aid'[23]. The keyword is 'often', as there may well be cases where those two entities differ and this has to be recorded in the documentation.

A phenomenon that frequently occurred was that, for some element, it is worthwhile to transfer the information from DDI over to EAD, but because the metadata element in DDI is very specific and its EAD counterpart is more general, it will be necessary to add an extra layer of information to avoid any ambiguity. Most of the time this can be done with a title. For example, in DDI's <dataColl> – 'Data Collection Methodology' element, there is a 'Custodian' element, which refers to the authority responsible for updating the sample frame which was used for the study. The sample frame designates the instrument which was used to select the population of the study, e.g. a phonebook. It would make sense to transfer this information in the <processinfo> – 'Processing Information' element in EAD, since the latter is meant to contain '[i]nformation about accessioning, arranging, describing, preserving, storing, or otherwise preparing the described materials for research use'[24]. However, the name of the entity that maintains the sample frame cannot just sit there in <processinfo> without any mention of its purpose. It will therefore be necessary to add a title such as 'Custodian of the Sample Frame'.

Finally, it was interesting to note that certain elements of DDI rarely ever seem to be used. This observation was made in the light of a corpus of DDI files[25]. Certain

---

[22] Encoded Archival Description Working Group of the Society of American Archivists and Network Development and MARC Standards Office of the Library of Congress, *Encoded Archival Description Tag Library: Version 2002*, Chicago, 2002, p. 221.
[23] Encoded Archival Description Working Group, *Encoded Archival Description Tag Library*, p. 214.
[24] Encoded Archival Description Working Group, *Encoded Archival Description Tag Library*, p. 206.
[25] See B. Peuch, "Elaborating a Crosswalk", p. 7-8.

key elements such as <controlledVocabUsed> – 'Controlled Vocabulary Used', for the controlled vocabularies used in the dataset, <guide> – 'Guide to Codebook', which can contain a specific glossary for a study, the various attributes of the <notes> element or the 'format' attribute of the <biblCit> – 'Bibliographic Citation' element, which can refer such manuals of style as the APA's or the Chicago MS, could never be found in actual DDI files. The likely reason for this is that those information items, when they are relevant for a particular dataset, are recorded elsewhere in the information infrastructure of the data archives that produce DDI files.

Conclusion and Next Steps

This paper introduced an early technical realization in the course of an ongoing project. Because the project is still running and involves different actors and disciplines, much contextualization was required before the mapping proper could be discussed. The technicality of the results might be an obstacle to their later integration in the global information system of the State Archives but, following the principles of knowledge management, several presentations of the mapping were given and the file itself was very much documented, with a description of each DDI element and the mention of problems or possibilities where needed.

The next steps are the following:

- After mapping DDI towards EAD, mapping the must-have and nice-to-have information elements of the State Archives' version of EAD towards DDI;
- Formatting the larger and the smaller crosswalks to make them publication-friendly, hopefully in open access;
- Integrating the new mapping in the technical and technological architecture of the upcoming Belgian data archive for the social sciences, which will largely be part of the State Archives'.

The idea of such a mapping endeavor accounts for the complexity of the SODA project, but also for its rich potential. Two worlds — archives and social sciences — and two types of institutions — a large federal entity on the one hand and several universities who are related to separate linguistic communities on the other — are coming together to build a new kind of administrative and scientific entity from which both types of actors will benefit. This entails devising practical means in order to rely on the strengths of both worlds so that both stand to gain something from this venture.