UNIVERSITY OF SZEGED
DOCTORAL SCHOOL OF EDUCATION

PHD PROGRAMME FOR
INFORMATION AND COMMUNICATION TECHNOLOGIES IN EDUCATION

Ingo Barkow

THE CHALLENGES OF METADATA MANAGEMENT IN COMPUTER-BASED SURVEYS AND ASSESSMENTS

Summary of the dissertation

Supervisor:
Prof. Dr. Benő Csapó
Institute of Education
University of Szeged

Szeged
2016

Introduction

Data management in the educational sciences and especially in the area of assessment of students consisted for many years out of handling different kinds of paper material. To give an example from the home organization of the author. The whole basement of the German Institute of International Educational Research (DIPF) consists of a test archive where standardized cognitive tests from several decades are available. Scientists can visit the archive and look at items, instruments, scales and results in paper form. Data Management in this sense consists only of a register in which box which test from which decade can be found (which should not mean keeping this information is a trivial task). The situation nevertheless gets much more complicated with the introduction of computer-based testing into the Educational Sciences. Computer-based testing has a lot of advantages in comparison to paper&pencil testing procedures like automatic scoring and thus the possibility to develop adaptive tests or the creating of log files which can be analysed to get a richer picture about the participants in a test than a simple „true/false" scenario, but the processes around data management are much more challenging. A modern researcher is basically able to re-use an item from paper&pencil test from the 1950s by copying the material. There might be issues about the layout like old fonts and pictures from this time period which look awkward to the modern student and there are certainly political, cultural or methodological dimensions to be considered, but there is no technological boundary which avoids the usage of the material. Furthermore, the information on the paper is sufficient to modernize the item. In computer-based testing we are facing other difficulties when we think of data management and archiving of instruments.

Literature Review

If one considers the project title of the dissertation „the challenges of metadata management in computer-based assessment and surveys" it already contains four different topics which are combined in one very specific subject matter. As the focus is rather small there are not many people who worked on this specific subject which also leads to a problem in finding fitting literature. Basically the focus gets smaller by the following questions, where the sort order has been changed to reflect the underlying analysis process:

- What can be considered metadata in general?
- Which metadata are relevant for educational sciences?
- Which of these metadata are relevant for computer-based assessment?
- Can these metadata be standardized or combined?

There is literature about metadata in general, but already the second limitation – metadata for the educational sciences – limits the amount of literature significantly. Furthermore, there is currently almost no scientific literature about standardized metadata in computer-based assessment. Therefore, this work can be considered a new research area or a blank spot, which might lead to the question if these processes or questions are relevant or necessary. Currently in computer-based assessment though there is a high amount of money spent for item development for research studies, high-stakes testing or marketing surveys there is only little money for standardizing metadata or archival of items. This also reflects in the research work about this topic. The focus of a literature review had therefore to be extended towards neighbouring disciplines like social sciences and data management in statistical archives.

The Generic Longitudinal Business Process Model (GLBPM)

The processes of documentation for statistical production lead to the development of business and information models for defining data management processes. This work also influenced other domains like the social sciences. Inspired by the work of the GSBPM a modified version called the Generic

Longitudinal Business Process Model (GLBPM) was developed by a group of scientist from different agencies at the DDI Alliance Dagstuhl Workshop in September 2011 (*Barkow, Block, Greenfield, Gregory, Hebing, Hoyle & Zenk-Möltgen, 2012*). The main difference between the GSBPM and the GLBPM is the latter model tries to specify the business processes in a longitudinal study in social sciences rather than showing the processes for storing statistical products. From the pure model the overlap is significant in the horizontal view.

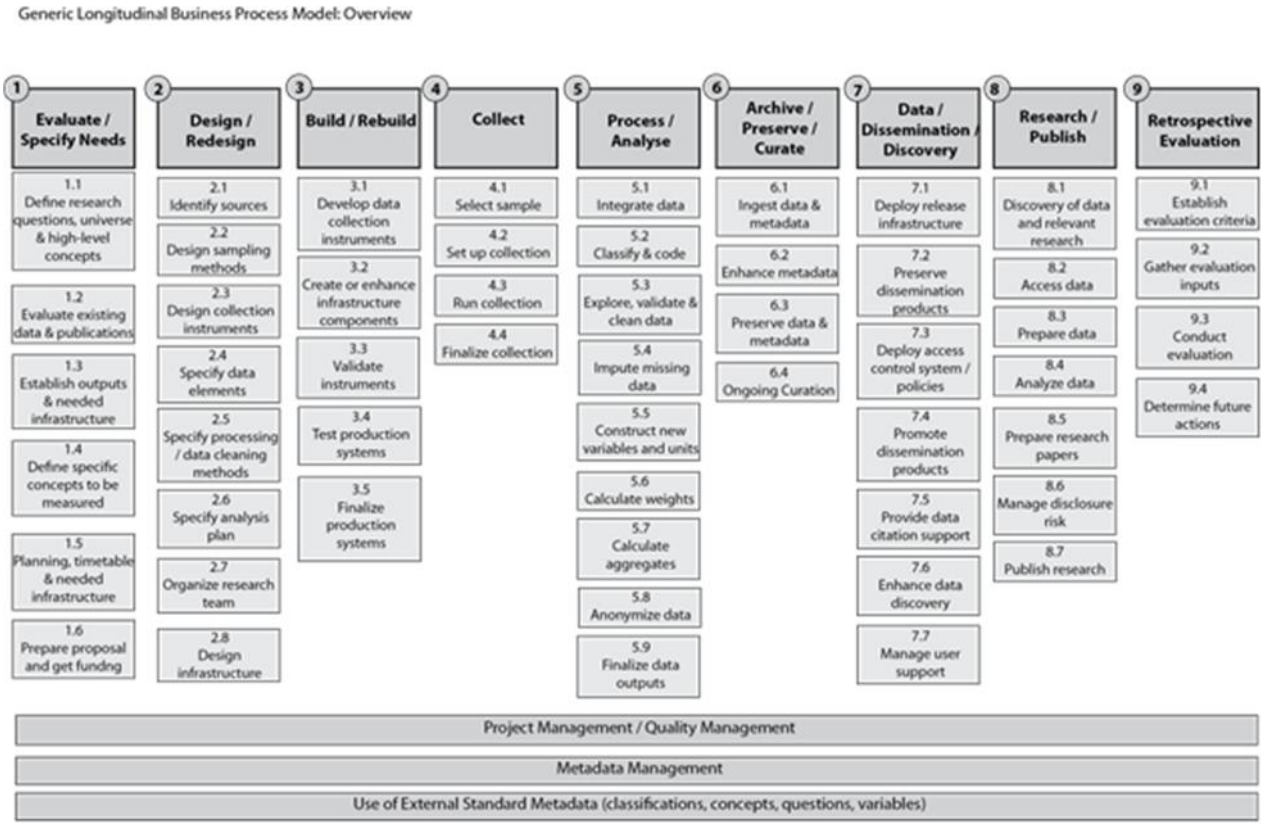


Figure 1. The Generic Longitudinal Business Process Model (GLBPM).
Source: Barkow, et.al. 2012 p. 7.

The model shows the processes how to setup the instruments and repurpose them in every wave of a longitudinal study. Essentially the proceedings in a longitudinal study are a repeating cycle where the instruments are created before the first wave (processes „Design/Redesign", „Build/Rebuild"), a data collection is performed (process „Collect"), the scientific dissemination processes are performed (processes „Process/Analyze", „Archive/Preserve/Curate", „Data / Dissemination / Discovery", „Research / Publish" and finally the next wave is prepared (process „Retrospective Evaluation"). The important factor is all the processes are built for re-use which is represented in the terms (e.g. „Redesign" or „Rebuild" imply the next waves) or the permanent accompanying processes „Metadata Management", "Project Management / Quality Management" and "Use of External Standard Metadata". GLBPM defines each of the sub-processes for its business logic. Here the differences to the GSBPM can be seen. GLBPM divides the processes more strictly and references to techniques used in the social sciences. It also connects the processes to external metadata standards like DDI. Nevertheless again also this model reflects the perspective of social sciences from the view of handling of longitudinal panel study using questionnaires. There are no processes regarding cognitive item construction or the specialities of handling computer-based assessment like it would be needed for a study like PISA or

PIAAC. Therefore this model again has to be modified if it should fit the needs of the educational sciences.

Different metadata standards and their capabilities

The dissertation introduces in the literature review different metadata standards and their capabilities. To get a clearer picture in this chapter here is an overview. For the purpose of identifying which metadata standards are able to deliver which part of the survey process the Generic Longitudinal Business Process Model will serve as a guideline to the processes where the metadata standards will be categorized against.

Table 1. Comparison between GLBPM processes and different metadata standards

Metadata Standard	Evaluate / Specify Needs	Design / Redesign	Build / Rebuild	Collect	Process / Analyse	Archive / Preserve / Curate	Data Dissemination / Discovery	Research / Publish	Retrospective Evaluation
LOM v1.0	Partly	No	No	No	No	Partly	Partly	No	No
LTI v1.0	Partly	No	Partly	No	No	No	Partly	No	No
QTI v2.1	Partly	Partly	Yes	Partly	No	No	No	No	No
APIP v1.0	Partly	Partly	Yes	Partly	No	No	No	No	No
SCORM 2004 (4th ed.)	Partly	No	Partly	No	No	No	Partly	No	No
DDI v3.2	Yes	Yes	Yes	Partly	Partly	Partly	Partly	Partly	Partly
SDMX v2.1	No	Partly	Partly	No	Yes	Yes	Yes	Partly	Partly
Dublin Core	No	No	No	No	No	Partly	Partly	Partly	No
DataCite Metadata Scheme	No	No	No	No	No	Partly	Partly	Partly	No

The first result which can be seen from the table with the high-level processes is none of the metadata standards is able to cover the demands of the business model in all details. The metadata standard filling most of the fields is Data Documentation Initiative v3.2, which fulfils at least all processes partly. Nevertheless, it is lacking sub-processes like “Setup Collection” or “Run Collection” in “Collect” as it does not focus on paradata or disposition codes. In DDI 3.2 the answer to the question on how to archive proceedings from field work would be to store this information in a file and document it with the “Archive / Preserve / Curate” processes. This is not sufficient if this information should be available for researchers in searchable databases. As the DDI Alliance is open for internal and external input the chances of filling these gaps are quite high in the long run. DDI Lifecycle developed from a metadata standard, which was primarily used in archiving into a standard which can also be used for survey design. An extension towards data collection seems to be a next logical step. Also seen from the comparison can be the complimentary use of DDI and SDMX, which is the basis for the DDI and SDMX dialogue

in 2010 and 2011. DDI focuses on the instruments while SDMX strengthens statistical production and fulfils most requirements there. A combination of these two standards would make sense as long as they do not cover each other's domains this well. Unfortunately as stated before there has not been much progress since 2011. The reason for this might lie in the development of GSIM as the combining information model and the work on DDI 4 which has similar goals. Furthermore as the model shows the educational metadata standards are lacking more processes as similar models from social sciences. The main focus is either to connect e-learning platforms (like in LOM or LTI) or pure item banking (like in QTI or APIP). There is currently no standard which even remotely tries to fulfil any requirements of data management or data archiving. Research work on this topic is therefore very much on the beginning. The Generic Longitudinal Business Process Model does not only cover singular processes but also offers continuous processes like project management or the ingest of external metadata.

Table 2. Comparison between GLBPM's ongoing processes and metadata standards

Metadata Standard	Project Management / Quality Management	Metadata Management	Use of External Metadata Standards
LOM v1.0	No	No	No
LTI v1.0	No	No	No
QTI v2.1	No	No	No
APIP v1.0	No	No	No
SCORM 2004 (4th ed.)	No	No	No
DDI v3.2	No	Partly	Partly
SDMX v2.1	Partly	Partly	Partly
Dublin Core	No	No	No
DataCite Metadata Scheme	No	No	Partly

The ongoing processes are the ones which are difficult to model in metadata standards and the question is if rather external sources should be used. It is therefore not surprising all metadata standards are lacking support here. For project management or quality management it makes sense to resort to known standards like PRINCE2 (project management) or ISO 9000 standards (quality management). From current research literature it can be seen metadata standards for educational sciences and especially computer-based assessment do not fully fulfil the demands of data management and data archiving for long-term preservation. Also, the standards from social sciences are further developed in those fields, but cannot be used "as-is" for our demands. Therefore a requirements analysis is necessary to figure out how either educational metadata standards can be improved or if standards from social sciences can be modified accordingly.

Research questions

Currently it is not known how much scientists know about data management processes for computer-based assessment and their willingness to participate in such tasks. Data management is time intensive and puts a burden on both scientists as well as the developers of assessment software as the software has to be extended in various respects from survey documentation to item development and finally to data analysis.

This leads to the following research questions:

- Which categories of different metadata standards do scientist know?
- How can those metadata standards be integrated into their daily work?
- What are the requirements of educational scientists towards metadata standards?
- What is missing to provide a high degree of re-usability (e.g. items, instruments)?

- What is missing to describe as much as possible for researchers who want to do a secondary analysis of the data?
- How could a model for storing a computer-based study including all metadata and paradata in the long term look like?
- Do researchers in the educational sciences need a work-around in the meantime?
- How can especially complex item types like simulations in complex problem solving (e.g. MicroFIN, see Greif & Funke, 2010) be modelled in a meaningful way?
- How can the results be stored for long time preservation or archiving?

Hypotheses

Before the start of the survey between the research questions and the field work also some hypotheses with underlying can be formulated on basis of the literature review.

- We expect more familiarity with data management and metadata standards from social scientist or researchers working in both fields than from pure educational scientist.

This is based on the higher qualitative level of the social sciences metadata standards and the higher number of institutions available for this purpose. Educational scientist haves less possibilities to use a proper standard or to find and consult an expert.

- We expect a general positive attitude towards metadata management as researchers understand the advantages, but a mediocre or even negative attitude towards current standards and their implementation.

Based on Punch (2009) secondary analysis of datasets generates interest of scientist and furthermore these techniques are used by large scale studies. Therefore the message of usefulness should have reached most scientists. Nevertheless the previous chapters show what metadata standards especially in the educational sciences are lacking. This should create a certain unhappiness only to be counterbalanced if the scientist do not know what they are missing.

- We expect most metadata standards will only be known by name or not at all by most researchers. There will not be many experts.

Normally data management though it is demanded from stakeholders is a process most scientists are not familiar with. They might only know those standards if the software they are using for their research supports it in the background or as a ‘sales argument’ of research data centres offering their services. It cannot be expected researchers are familiar with the standards as they normally do not fall into their domain.

- We expect significant interest of educational researchers in secondary analysis of datasets.

As surveys and fieldwork become more and more expensive researchers get more interested in re-using previous datasets or instruments.

- We expect younger researchers to have a more positive attitude towards sharing of their research data than more senior researchers.

Re-using of datasets needs an attitude of sharing data which has only be promoted in the recent years. Before a high linkage between data and primary researcher existed. Therefore, we expect some traditional views (“this data is mine”) from scientist, which started their careers with a different paradigm.

- We expect the average time for data management and documentation in studies and surveys not to be sufficient for generating high-level metadata.

This hypotheses bases on the professional experience of the other and the missing elements of most metadata standards. A normal ingest of research data into a data archive consists of a codebook describing the core variables and the dataset in the format of a statistical package (e.g. SPSS, Stata) in different levels of quality. There are normally no further information (e.g. instruments, reports, analytical scripts).

Methodological considerations in questionnaire design

The first question about a survey about metadata management for educational sciences is the mode in which it should be performed. Basically the following modes are available (Bethlehem & Biffignandi, 2012).

- Paper and Pencil Interview (PAPI) – paper questionnaire conducted in house by an interviewer
- Computer-Assisted Personal Interview (CAPI) – computer-based questionnaire conducted in house by an interviewer
- Computer-Assisted Web Interview (CAWI) – web survey filled out by the participants themselves
- Computer-Assisted Self Interview (CASI) – computer-based questionnaire filled out by participants in a facility, sometimes observed by audio or video
- Computer-Assisted Telephone Interview (CATI) – computer-based questionnaire conducted by an interviewer via phone

As this dissertation is not directly connected to a research project there is no budget whatsoever for data collection. It is therefore not possible to pay interviewers. This eliminates the modes PAPI, CAPI and CASI. Though the costs for CATI are considerably lower than sending out interviewers the costs for a phone survey are still too high. The only affordable solution is therefore conducting a web survey or CAWI mode though web surveys have clear advantages and disadvantages.

Sampling

As there is no world-wide central register of the groups described in the last chapter and the participation is completely voluntary the sampling has to be quite random. It is only possible to identify the candidates by demographic information and their job biography they give at the very end of the questionnaire. As discussed at the beginning of the chapter this information is not reliable. On the other hand as the target group are researchers the expectation is they will be more interested in the results themselves. To get a sufficient number of respondents an invitation for participation was sent out to big research organizations, mailing lists for researchers and individual contacts. A typical problem of web surveys a real sampling with sampling plan is not possible as there is no world-wide register of social scientists or educational scientist. In addition, web surveys usually have very low response rates. The identification of researchers is therefore only possible by questions about highest educational degree, scientific background, academic titles and years spent in research. As the data analysis will separate between different groups (e.g. social scientists vs. educational scientists) the results have to be checked if the number of collected interviews is sufficient for data analytics. To enable these analysis the Cohen table (Cohen, 1988) on effect sizes will be used. A very prominent tool for this purpose is G*Power (Faul et. al., 2009). This tool will be used in the later data analytics chapters to determine if the group sizes are sufficient for the effects to be considered significant as the response rate can be expected to be small.

Final data analytics on the combined dataset

The first data collection in December 2014 did not have enough participants to allow for more sophisticated data analytics. For this reason a second data collection took place in spring and summer 2015 with a shorter questionnaire and more advertisement to generate a higher response rate. This

chapter describes the results of the combined dataset from the first and second data collection and includes some more advanced data analysis procedures.

Description of the combined dataset

The combined dataset consists from the 120 completed interviews from the first data collection plus the 198 completed interviews from the second data collection. Both datasets were attached only for question items appearing in both questionnaires. In the second questionnaire there was a question to identify persons who have also participated in the first data collection. These individuals have been cleaned out of the dataset resulting in n=272 completed interviews for the combined dataset. Though this amount is still far from being impressive according to G*Power the field strength was good enough to allow for separation into academic status and research area. For academic status it was possible to separate the participants into four groups (PhD students, researchers without PhD, researchers with PhD and Professors). There were not enough bachelor or master students to form a students' group. In research area three groups could be formed (researchers from educational sciences, social sciences or both). There were not enough participants from other disciplines (e.g. computer science as third largest group) to form further groups.

Descriptive statistics on the combined dataset

As to be expected from the first and second data collection the combined results also show the same results as the individual datasets which differed only in minor details. This can be proven already in Table 3 about knowledge on metadata standards.

Table 3. Frequency (%) of Metadata Standards – combined dataset

Standards	Usage "often"	Usage "rarely"	Usage "never"	Usage "Do not know"
Data Documentation Initiative (DDI)	15.0	10.9	27.1	47.0
Dublin Core	7.6	8.7	19.3	64.4
Learning Object Model (LOM)	1.2	3.9	19.4	75.6
Learning and Test Interoperability (LTI)	1.2	1.6	17.8	79.5
Machine Readable Cataloguing (MARC)	2.3	6.2	16.0	75.5
Metadata Encoding and Transmission Standard (METS)	1.6	3.5	17.9	77.0
QueDex	0.4	0.8	15.4	83.4
Questionnaire and Text Interoperability (QTI)	3.5	2.7	16.9	76.9
Shareable Content Object Reference Model (SCORM)	0.8	5.4	17.4	76.4
Statistical Data and Metadata Exchange (SDMX)	2.7	2.3	23.5	71.5
Text Encoding Initiative (TEI)	1.5	3.1	17.8	77.6

The results from the second data collection allowed for a little more variety than the first data collection. As both datasets are now combined this effect is moderated by the lower spread in the first dataset. Nevertheless the combination of datasets does not change the results from the first and second data collection which states a high level of uncertainty and lack of details in regards to metadata standards. Metadata standards are largely unknown to researchers and the ranking stays stable in all three datasets.

Table 4. Distribution of the statements regarding metadata standards – combined dataset

Statement	1	2	3	4	5	6
Identification of metadata standards (MS0601)	3.5	15.7	15.4	14.6	9.4	41.3
Metadata standards and domains (MS0602)	4.7	11.0	10.2	18.9	25.2	29.9
Importance of metadata standards (MS0603)	21.7	38.3	9.9	1.2	1.2	27.7
Complexity of studies in regards to metadata standards (MS0604)	2.0	4.3	12.6	30.3	19.7	31.3

1 = Strongly agree; 2 = Agree; 3 = Neither agree nor disagree; 4 = Disagree; 5 = Strongly disagree; 6 = Do not know

Metadata standards are recognized to be important, but researchers feel their knowledge and handling not to be adequate. This again leads to the overall question if researchers need to be the persons to be in charge or the driver of these processes. To identify which kinds of data sources the researchers use they were asked to specify their most common research design.

Table 5.Frequencies of data sources (%) – combined dataset

Data Source	Always	Often	Sometimes	Rarely	Never
Structure cognitive test (e.g. computer-based assessment)	10.7	22.2	17.9	15.9	33.3
Structured questionnaire (e.g. multiple choice)	35.2	42.7	10.7	3.6	7.9
Qualitative text data (e.g. narrative interviews, oral history)	6.3	19.0	21.8	26.2	26.6
Experimental designs	5.6	16.7	24.7	23.9	29.1
Observational data (e.g. recordings, transcripts)	5.6	18.7	29.0	21.8	25.0

Not surprisingly Table 5 shows a heavy preference towards questionnaires and structured cognitive tests with the first data collection tipping the scale even more towards quantitative research methods. This result can also explain why metadata standards and data description are underdeveloped in these areas. As stated before these kinds of research design should not be ignored in data management, but it leads to the question if an archive can provide a service structure which can accommodate the needs for managing quantitative and qualitative datasets. Nevertheless, this question is out-of-bounds for this dissertation, as it cannot be discussed on basis of these datasets.

Table 6 - Frequencies of the different metadata provided (%) – combined dataset

Category	Always	Often	Sometimes	Rarely	Never	Do not know
Analytic scripts	16.8	18.8	18.3	17.3	22.1	6.7
Answer categories	44.4	20.8	16.9	6.8	5.8	5.3
Concepts to be measured	23.9	27.3	23.4	9.3	10.2	5.9
Flow logic	15.5	18.4	17.5	9.7	24.3	14.6
Fragments of software	3.9	13.1	15.0	18.9	38.3	10.7
Interviewer instructions	26.1	25.1	17.7	10.3	12.3	8.4
Layout definitions	9.9	18.2	18.2	18.2	24.1	11.3

References to research literature	26.3	18.0	23.9	9.8	17.1	4.9
Scoring rules	20.1	18.6	17.6	8.3	20.6	14.7
Statistical values from previous studies	6.8	22.0	22.0	16.1	26.3	6.8
Stimuli in reusable formats	5.4	12.7	14.2	19.6	34.8	13.2
Variable definitions	28.4	22.1	16.7	7.8	12.7	12.3

Answer categories, concepts, variable definitions, interviewer instruction and references to research literature are more likely to be preserved while the technically more advanced metadata are not. The more technically oriented metadata seem to be out-of-scope for researchers.

Table 7. Frequency of the access to datasets (%) – combined dataset

Category	Always	Often	Sometimes	Rarely	Never	Do not know
Public Use Files	8.8	22.8	25.7	12.3	29.3	1.2
Scientific Use Files	9.3	25.6	23.8	18.6	21.5	1.2
Raw datasets	8.8	31.6	18.1	15.2	25.1	1.2
Data enclaves, secure data services or Virtual Research Environments	1.7	8.1	8.7	18.0	54.7	8.7
Relational databases	2.9	11.6	15.7	12.8	51.8	5.8
Data Warehouses or analytical databases	0.6	4.1	11.6	16.9	54.1	12.8
Remote calculation or job submission	0.6	1.8	6.4	8.2	65.5	17.5
Personal extracts created by portal websites	0.6	2.9	9.9	14.5	59.3	12.8

Table 7 shows a tendency towards file-based approaches (Public Use Files, Scientific Use Files and Raw Datasets) which is again no deviation from the results of the individual data collections. Remote calculation, Virtual Research Environments or personal extracts by variable shopping basket systems are virtually unknown. The result is especially interesting as in international studies the datasets for dissemination are provided by data collection or survey organizations and undergo cleaning processes before release. Either the researchers only analyse small internal datasets or they are not aware their "raw data" is actually not directly from the field as they assumed. The next question asked if the researchers shared metadata, paradata or data from their research with others. Like in the individual data collections almost half of the participants did not want to participate in such processes. Here the combined result:

- 59.2% Yes
- 40.8% No

The participants in the second data collection were a little more open with almost two-thirds sharing their data, but here the result from the first data collection (55%) mediated the result. From a data archive perspective, all the results are actually disappointing. The next question therefore asked with whom they usually share or do not share their data.

Table 8 - Recipients of data (%) – combined dataset

Statement	Share data	Don't share
I deliver to data to research data center or archive.	20.2	79.8
I hand over the data to a designated person at my organization.	14.0	86.0
I share my data only with researchers I know.	14.3	85.7
I share my data with other researchers from the same field.	15.4	84.6
I share my data with other researchers from other fields.	9.6	90.4
I share my data only with researchers from my organization.	10.7	89.3
I share my data with other persons who are interested in my research.	16.5	83.5

As can be seen sharing the data (59.2%) does not necessarily mean delivering it to a research data facility or data archive (20.2%). It can be expected the values might even go down further if there was a separation between institutional archive and external archive. This of course raises the question why researchers do not share their data. The following question only was asked by filter to the n=79 researchers (40.8%) who answered not to share their data.

Table 9 - Reasons to not share data (%) – combined dataset

Opinion	1	2	3	4	5
I do not know where to hand in my data for archiving	12.7	19.0	36.7	17.7	13.9
I have no resources or support from my employer to hand in the data	19.2	29.5	29.5	14.1	7.7
Preparation of data needs to much effort	17.9	38.5	29.5	6.4	7.7
My organization or I have security concern	13.9	32.9	24.1	11.4	17.7
Somebody could misinterpret my results	7.7	15.4	30.8	25.6	20.5
Somebody could use my data before me	12.8	28.2	19.2	20.5	19.2

1 = Strongly agree; 2 = Agree; 3 = Neither agree nor disagree; 4 = Disagree; 5 = Strongly disagree

Obviously preparing the data for an external usage needs too much effort and there are also trust issues with other organizations. Research data centres and archives might use these results to improve their processes. The hurdles for researchers to hand in their data might simply be too big.

Differences between groups

Though the overall results provide some interesting insights there were researchers from different backgrounds participating as well as persons who have been in academia for different amounts of time. The question is if scientific field or status in academia have any impact on the results. To answer those questions has be performed mainly by the Mann-Whitney U test (Mann & Whitney, 1947) or Kruskal-Wallis H Test (Kruskal & Wallis, 1952) as they are considered to be the best fit for questionnaires like they were used in both data collections.

Academic status

As described before the combined response rate allowed for a separation into four groups – PhD students, researchers without PhD, researchers with PhD (post-docs) and Professors. In the usage the

different type of metadata standards there were no significant differences between the groups. In the opinion regarding metadata standards were significant differences between the groups in regards to importance of metadata standards ($\chi^2(3)=7.88$ $p<.05$). In detail there were significant differences between students and the researchers without PhD ($U=1030$ $p<.05$) plus PhD Students and researchers with PhD ($U=712$ $p<.05$). The PhD students evaluated the importance of the metadata standards lower than the other two groups. In other cases there were no significant differences between the groups. In the usage of data sources were no significant differences between the groups. In different kinds of metadata provided by researchers there were significant differences between the groups in analytic scripts ($\chi^2=8.63$ $p<.05$), concepts to be measured ($\chi^2=8.82$ $p<.05$), fragment of software ($\chi^2=8.73$ $p<.05$) and scoring rules ($\chi^2=8.95$ $p<.05$). In case of analytic scripts the researchers without PhD, researchers with PhD and Professors provide more often the analytic scripts than the PhD students ($U_{nophd}=545.5$ $p<.05$; $U_{phd}=797$ $p<.05$, $U_{prof}=592.5$ $p<.05$). In the other cases there were no significant differences. In the concepts to be measured the researchers with PhD and Professors provided more often concepts to be measured than PhD students ($U_{phd}=912$ $p<.05$, $U_{prof}=597$ $p<.05$). In the other cases there were no significant differences. In fragments of software the researchers without PhD and Professors provide more often the fragments of software than PhD students ($U_{nophd}=504$ $p<.05$; $U_{prof}=610.5$ $p<.05$). In the other cases there were no significant differences. In scoring rules researchers with PhD and Professors provide more often the scoring rules than PhD students ($U_{phd}=765.5$ $p<.05$, $U_{prof}=583.5$ $p<.05$). In the other cases there were no significant differences. In summary the Professors delivered more often additional metadata while the PhD students at least in these data collections do not seem very prone to the practices of data documentation. In frequency of the access to datasets were significant differences between the groups in the usage of relational databases ($\chi^2=9.28$ $p<.05$). The researchers without PhD, researchers with PhD and Professors used more often relational databases than the PhD students ($U_1=430$ $p<.05$; $U_2=618$ $p<.05$; $U_3=442.5$ $p<.05$). In the other cases there were no significant differences between the groups.

Research area

Similarly to the academic status the researchers were also asked about their field. From the amount of participants it was possible to create three groups – researchers who come from the field of educational sciences, social sciences or both. Here their answers on the different items.

Table 10 – Usage of different kinds of metadata – Kruskal-Wallis test

Category	$\chi 2$	P
Data Documentation Initiative (DDI)	38.20	<.01
Dublin Core	19.73	<.01
Machine Readable Cataloguing (MARC)	12.93	<.05
Metadata Encoding and Transmission Standard (METS)	7.68	<.05
QueDex	9.28	<.05
Statistical Data and Metadata Exchange (SDMX)	12.18	<.05
Text Encoding Initiative (TEI)	12.15	<.05

As Table 10 shows there were significant differences between the groups in these categories. The social scientists used more often metadata standards than educational scientists. There were no significant differences between the educational scientists and the both group, except in the usage of the metadata standard DDI ($U=1082.5$ $p<.05$). Both group used more often the rest of the metadata standards than social scientists with two exceptions: MARC and METS. In these last two categories there were no significant differences between them.

Table 11 - Usage of different kinds of metadata – Mann-Whitney test

Category	Educational – social sciences	Social sciences – both group
Data Documentation Initiative (DDI)	946.0(p<.01)	1052.5(p<.05)
Dublin Core	1385.0(p<.01)	1125.0(p<.05)
Machine Readable Cataloguing (MARC)	1485.0(p<.01)	1153.5(p>.05)
Metadata Encoding and Transmission Standard (METS)	1673.5(p<.05)	1249.0(p>.05)
QueDex	1698.5(p<.05)	1185.0(p<.05)
Statistical Data and Metadata Exchange (SDMX)	1578.0(p<.05)	1112.5(p<.05)
Text Encoding Initiative (TEI)	1567.0(p<.05)	1189.0(p<.05)

As table 12 shows there were significant differences between the groups in identification (χ^2=16.02 p<.01) and importance (χ^2=19.49 p<.01) of metadata standards. The social scientist and both group agreed more these statements than educational scientist. There were no significant differences between the social scientist and the both group.

Table 12. Opinion regarding metadata standards – Mann-Whitney test

Category	Educational – social sciences	Educational sciences – both group
Identification of metadata standards	1342.5(p<.01)	1152.5(p<.05)
Importance of metadata standards	1260.0(p<.01)	1143.5(p<.05)

In the usage of data sources there were significant differences between the groups in usage of structure cognitive tests (χ^2=37.54 p<.01), qualitative text data (χ^2=8.19 p<.05) and observational data (χ^2=11.81 p<.05). Not surprisingly the educational scientist and the both group used significantly more times the cognitive test than social science group (U_1=1034.0 p<.01; U_2=682.0 p<.01). There was no significant difference between the educational sciences and the both group. In qualitative text data both group used more often the qualitative text data than the educational scientists (U=1052.0 p<.05). In other cases there were no significant differences. In observational data educational scientists used more often observational data than social scientist (U=1599.0 p<.05) and the both group used it more than the educational scientists (U=1039.5 p<.05). There was no significant difference between both group and social scientist.

Table 13 – Metadata provided by researchers – Kruskal-Wallis test

Category	χ2	P
Answer categories	8.14	<.05
Flow logic	12.11	<.05
Scoring rules	8.83	<.05
Statistical values from previous studies	7.85	<.05
Stimuli in reusable formats	7.82	<.05

There were significant differences between the groups in these categories (Table 13). As the Mann-Whitney test shows there were significant differences between educational and social scientist (Table 14), except in statistical values from previous studies. The social scientists provided more often answer categories and flow logic than educational scientists, but in the remaining categories the situation is reversed. There were no significant differences between the both group and the educational scientists, except in flow logic. The both group provided more often flow logic than educational scientists (U=809.5 p<.05). The both group provided more often these metadata types than the social scientists (Table 14), except answer categories and flow logic, because in these cases there were no significant differences.

Table 13 – Metadata provided by researchers – Mann-Whitney test

Category	Educational – social sciences	Social sciences – both group
Answer categories	1276.5(p<.05)	1110.5(p>.05)
Flow logic	1274.0(p<.05)	1202.5(p>.05)
Scoring rules	1286.5(p<.05)	886.5(p<.05)
Statistical values from previous studies	1457.0(p>.05)	860.0(p<.05)
Stimuli in reusable formats	1289.5(p<.05)	925.5(p<.05)

In access to datasets there were significant differences between the groups in access of scientific use files ($\chi 2=37.54$ $p<.01$) and data enclaves ($\chi 2=8.19$ $p<.05$). Social scientists and the both group access more often these data products than educational scientists, and there was no significant difference between both group and social scientist (Table 15).

Table 15 – Access to datasets – Mann-Whitney test

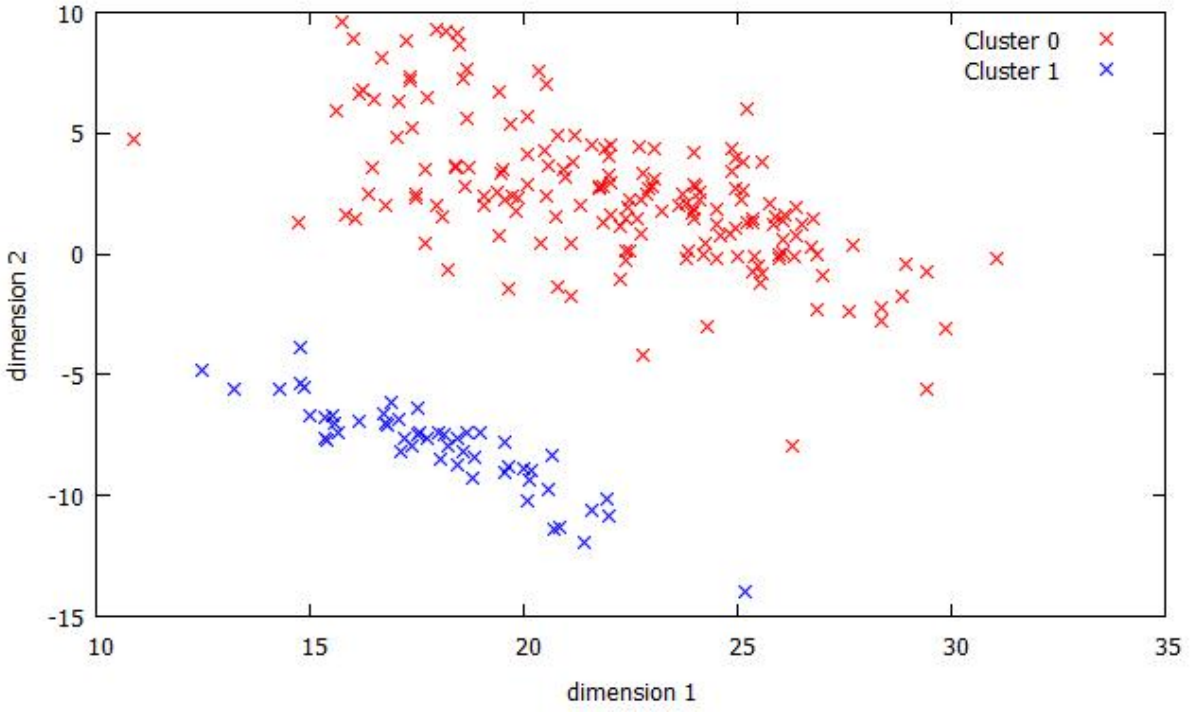
Category	Educational – social sciences	Educational sciences – both group
Scientific use file	880.0(p<.01)	509.0(p<.05)
Data enclaves	1134.5(p<.05)	486.0(p<.01)

This result is remarkable as scientific use files should be very common also for educational scientists. The question is what they use instead as raw files or primary research data are less common especially in international large scale assessment.

Data Mining

To better understand the results shown by the cluster analysis the idea was to perform a deeper assessment e.g. with the help of advanced data mining techniques. In late 2015 after the analysis of the second data collection and the combination of the datasets the author experimented with the dataset from both data collections e.g. by applying Structural Equation Modelling (SEM) (e.g. Ullman & Bentler, 2003) to it which did not provide interesting results as the data derived from the questionnaire did not show any latent variables except for measuring opinion. Another candidate for a more explorative approach could have been Educational Data Mining, but for applying techniques from this field a proper audit trail is necessary. Audit trail means all interactions like clicking buttons or mouse movements are recorded as well as changing of answers including a timestamp in milliseconds. Limesurvey as a questionnaire system does not offer this high level of granularity as in surveys the timing is not of the essence in an interview. It only records the total interview time and answer time for individual items. This is also not uncommon for educational studies. In PIAAC the interview system only records the individual time of items to detect cheating by the interviewer (i.e. faking interviews) not for data analysis purposes (see Zabal et. al. 2012). Real exact time measurement only starts in the cognitive assessment. This is also the reason why the technical standards and guidelines of this study allow a break or coming back on another day during the Background Questionnaire, but not during the Computer-Based Assessment. Timing of interactions simply does not matter during the interview. This is also the reason why most web survey systems like Limesurvey do not possess this functionality. Nevertheless it is possible to apply data mining directly to the dataset in an exploratory approach to detect more groups or clusters than the ones based on categories. The basic technique is to combine all participants with all possible variables as dimensions and start reduction processes so clusters might show up. In this case the dimensions of the data were set by manually selecting 46 variables as the required memory and time for processing the data increase exponentially with the growth of the number of dimensions. As

described before the database had instances with missing values and unfilled variables. Therefore, instances with more than 20 undefined features had to be removed from the dataset as the processing steps would be meaningless (too many zeros). After this filtering step the sample size was reduced to 224 participants in the database for further analysis. The remaining missing values in the dataset were as hinted at before replaced by zeros, since this value has no effect on the result of the matrix-vector multiplication which was is in the upcoming dimension reduction step. Instead of clustering the 46-dimensional data and analyzing the result of the method (e.g. measuring the distances between the cluster centroids or the compactness of the clusters) the dimension of the data was reduced to a two-dimensional plot as human understanding cannot grasp the complexity of a 46-dimensional hypercube (Bellman defined this phenomenon as the "curse of dimensionality" – Bellman, 1956). This allows to plot the instances to a plane and try to identify patterns. This plot can be seen in Figure 50 and it indicates there are two clusters in the dataset. The technique used to reduce the dimensions is called Singular Value Decomposition (SVD) (Golub & Reinsch, 1970) which is very commonly used in the area of machine learning. It searches for the bases of a subspace which maximizes the variance in the dataset. Another option to plane the dataset could have been the Principal Component Analysis (PCA) method (e.g. Jolliffe, 2002) which is one of the most commonly used dimensionality reduction technique. It has the same basics as Singular Value Decomposition). The difference is PCA uses the eigenvectors calculated on the covariance matrix of the data and SVD uses the singular values which are the left and right eigenvectors of the dataset. The advantages of the feature extraction methods against the selection ones are the following as it preserves the relative distances among the data elements, all of the resulting dimension is a linear combination of the original feature space and more information can be preserved by transforming the data into the two dimensional space instead of just select the two best features in the original space.



Note: Dimensions 1/2 (x/y axis) correspond to the direction of the first/second singular vector, which means these directions have the largest variance.

Figure 2 – Scatter plot of planed clusters

These two clusters become visible by slicing the hypercube into different two-dimensional planes. Nevertheless what do they mean and where do they differ? As described before they do not base on human defined categories like academic status or research discipline which were used in the previous chapters. When looking at the representations of the variables they are a bit surprising as they differ in relatively uncommon metadata standards mainly in the area of qualitative research (METS, QueDex, SDMX and TEI), use of experimental designs, use of fragments of software and virtually all of the different kinds of data access from Public Use Files to Data Enclaves. If human categories should be used the representation could be interpreted as clusters of more qualitative and experimental oriented scientists vs. empirically driven researchers in social sciences and educational sciences. The even more surprising result is the differences between Cluster 0 and Cluster 1 in regards to data access. None of them (with n=53) uses any kind of data access to secondary research data. As this represents the qualitative and experimental oriented persons which fall out of scope of this dissertation the topic will not be discussed further, but it certainly has a lot of potential to be researched in upcoming papers.

Hypotheses revisited

In this chapter the hypotheses will be revisited and compared with the data if they were true or not.

- *We expect more familiarity with data management and metadata standards from social scientist or researchers working in both fields than from pure educational scientist.*

This hypothesis proved to be true. Social scientist know more metadata standards than their counterparts from the educational sciences. Nevertheless, the distinction might not be that clear as the analysis show a big overlap in the usage of cognitive instruments by social scientists and questionnaires by educational sciences. The categories in general might not be that clear. Another result is data management is also less known to the group of computer scientists so hoping this task can easily by shifted into the software as there are experts by default in this field is also futile.

- *We expect a general positive attitude towards metadata management as researchers understand the advantages, but a mediocre or even negative attitude towards current standards and their implementation.*

This hypothesis proved to be true. The importance of metadata standards was evaluated highly by all groups regardless of discipline or academic status. On the other hand most metadata standards are unknown to the groups and there are a lot of missing elements which were identified by the participants. Especially in the handling of metadata for log file analysis there is confusion about standards and formats. More work in specification is therefore necessary.

- *We expect most metadata standards will only be known by name or not at all by most researchers. There will not be many experts.*

This hypothesis proved to be true as pointed out in the first bullet point. Metadata standards are for researchers more like an opportunity than a reality. They only know very few standards in detail and the software products they are using are exactly the ones, which do not support anything except their own propriety formats.

- *We expect significant interest of educational researchers in secondary analysis of datasets.*

This hypothesis proved to be true though there is no significant difference between the educational researchers and the other groups. All the researchers in the study were interested in the secondary analysis of datasets and a large portion of them are also actively doing this. It can be expected from the other values this trend will continue and get even stronger.

- *We expect younger researchers to have a more positive attitude towards sharing of their research data than more senior researchers.*

This hypothesis proved to be false. There was no significant difference between academic status or years in research and the willingness to share data. Though researchers are seeing the benefits of metadata standards and do secondary analysis the willingness to share the own data especially with persons outside the own organization is rather low.

- *We expect the average time for data management and documentation in studies and surveys not to be sufficient for generating high level metadata.*

This hypothesis proved to be true. The time specified for data management in an 18-month survey is about six weeks. This can be cross-referenced with the metadata elements data producers from the both disciplines offer. In end data users can consistently expect codebooks with variable information and data files. Every further information is optional and it is not clear if it is produced by the original research team or the staff at research data centres performing the process of enhancing metadata.

Answering the research questions

Here are the research questions and the comparison with the respective data.

- *Which categories of different metadata standards do scientist know?*

The survey shows scientists neither in the social sciences nor the educational sciences have deeper knowledge about the metadata standards and their internal structure themselves. DDI and QTI are known by name and scientists are also aware of the advantages of metadata standards, but cannot really use them without the help of data management personnel. Unfortunately, also the software basing on metadata standards is less used than software using propriety formats. Therefore, there is also not much indirect usage of metadata standards.

- *How can those metadata standards be integrated into their daily work?*

As the awareness about metadata standards is low especially in the educational sciences which is the main focus of this dissertation more impulses will have to come from research data centres towards researchers or IT personnel to change current procedures. This can be either by consulting researchers on data management procedures and helping them to embed them into their processes with the least impact on time. Furthermore also IT personnel can be trained from data management side to embed the creation of standardized metadata into software development so they become a by-product of the research process and therefore do not produce additional time constraints for researchers (ideally new software products should even save time for researchers).

- *What are the requirements of educational scientists towards metadata standards?*

The requirements do not differ from those of social scientists as long as both are using questionnaires and for this purpose DDI seems to fulfil most demands.

Scientists from both disciplines would like to store:

- Answer categories
- Concepts to be measured
- Scoring rules
- Variable definitions
- Flow logic
- Interviewer instructions

The situation changes when the focus is moved towards computer-based testing with cognitive instruments. The requirements on the survey lifecycle are very similar and have extended demands in regards to instruments (e.g. scoring, stimuli, complexity), but the metadata standards like QTI or APIP are much inferior. Currently there is no metadata standard for the educational sciences which fulfils those demands.

- *What is missing to provide a high degree of re-usability of items and instruments for computer-based assessment?*

Current metadata standards for computer-based testing in the educational sciences like QTI only focus on the design and delivery of simple item types. The processes of "Survey Design", "Data Processing", "Distribution", "Discovery", "Analysis" and "Re-Purposing" in comparison to DDI Lifecycle are missing. QTI can store the design of a simple item without layout information for re-use.

- *What is missing for educational researchers who want to do a secondary analysis of the data?*

Data producers offer a least a basic metadata set along with the data and the data users seem to be satisfied with this. For pure analysis this might be good enough. Nevertheless, this can only be accounted for internal data as most researchers do not hand in their data to third parties. This means in many cases the data is not available externally though the scientists would be willing to share the data.

- *How could a model for storing a computer-based assessment including all metadata and paradata in the long term look like?*

Based on the survey results it looked similar to the GLBPM or DDI Lifecycle model, but with an additional focus on scoring, stimuli and complexity in the area of instrument development and delivery.

- *Do researchers in the educational sciences need a work-around in the meantime?*

Yes, as the current metadata standards do not fulfil the demands completely a mix of metadata standards from different domains plus manual documentation is needed even for simple items and instruments.

- *How can especially complex item types like simulations in complex problem solving (e.g. MicroFIN, see Greif and Funke 2010) be modelled in a meaningful way?*

With the current metadata standards complex problem solving cannot be modelled as neither QTI nor APIP deliver answer modes, scoring mechanisms or support for graphical elements which can express this level of complexity. For those items a completely new development in form of an abstraction layer or even standardized modelling language is necessary.

- *How can the results be stored for long time preservation or archiving?*

The results from studies are the smallest problem encountered in the survey. The formats for statistical packages are quite stable over the years and known to the researchers. In a worst-case scenario the tabular files can be converted into CSV files – a standard which is unchanged since the 1970s and supported as the lowest demeanour also in the coming decades. Nevertheless there might be data loss involved e.g. if the original data was stored in a relational database or analytical database with additional information and in a more meaningful way. The answer therefore is implementing data management, data archiving or data stewardship processes which means dedicating personnel which preservation and understanding of these data.

Impact of the results on current metadata standards

As the research questions prove currently handling metadata for computer-based surveys and assessment is a problem as the metadata standards do not cover all the demands from a researcher's perspective. A

dissertation should provide mainly new insights but it is also not a mistake to provide answers towards the applicability of those results. Therefore this chapter will contain a discussion how the situation can be improved using current standards. At the end there will also be a topic about creating a completely new standard.

Extending Data Documentation Initiative (DDI) with educational content

In the survey DDI proved to be the most prominent of the metadata standards and is used by social scientists and educational scientists as well. DDI covers also most of the lifecycle tasks specified in the GLBPM, GSBPM and GSIM. Furthermore, DDI Lifecycle is currently undergoing a change from a XML scheme towards an ontology using multiple representations (version shift from DDI Lifecycle 3.2 to DDI Lifecycle 4.0 in the DDI Moving Forward project). The new DDI is supposed to be modular so it can be extended with other domains outside of the core set (e.g. healthcare metadata). With this modular structure it would be possible to add an extension especially for educational content. The module „educational survey“ could be a branch of „simple survey“ or „advanced survey“ depending how these packages will look once they have been fully specified by the DDI Alliance. The current development process regarding the core packages for DDI4 is running since 2012 and a first release candidate is scheduled for 2015, but the actual fully released and tested version 4.0 of DDI might take more time than that.

An involvement for educational researchers in DDI for creating an own module is a useful enterprise for several reasons:

- Educational researchers also heavily use questionnaires and can therefore get the advantages of handling both types of instruments within the same standard
- DDI 3.2 already contains some extensions derived from the educational sciences (e.g. stimuli, item batteries) as same institutions from this sector already were part of the development process in the past (e.g. DIPF)
- DDI contains a lot of support for statistical packages and publications which were also topics in high value by educational scientists

Nevertheless if educational scientists start their involvement now it will take some years until the results will be visible in the metadata standard itself and even more time until they will influence software products. Counting on DDI as an extended educational standard is therefore promising but only a long-term option. Furthermore it should be mentioned the German Institute for Educational Progress (IQB) started this discussion of embedding educational content into DDI already once in form of the EduDDI initiative (Mechtel, 2009), but this did not lead to a lasting success at this time. This can be explained by the very early version of DDI Lifecycle (version 3.0 at this time) and the lack of awareness in the educational sector. Maybe the process of developments around DDI4 prove a more successful ground.

Extending Questionnaire and Test Interoperability (QTI) with content from lifecycle models

The other option to create a better metadata standard for computer-based surveys and assessments for the educational sciences would be to extend the most prominent standard there. Similar to DDI also QTI is currently under further development due to the merging process with APIP. Nevertheless, this process might be more difficult than extending DDI as there are some challenges:

- QTI does not contain any objects outside of the instrument design domain, so processes regarding survey design, data collection, dissemination and publication have to be added. This is a huge endeavour as several hundreds of object classes and types have to be specified.
- The development cycles in QTI are long. The further development from QTI 2.0 to 2.1 took ten years from 2003 to 2013. This process might speed up due to the APIP involvement.

- QTI is not well-known even among educational researchers as the survey has shown.

In summary hoping for an improvement of QTI into a direction of more lifecycle-oriented workflows might take more time than the developments surrounding the DDI Moving Forward project. It might nevertheless be an idea for educational researchers interested in metadata management to join the community at IMS Global surrounding QTI and influence those developments.

Developing a fully new metadata standard for computer-based assessment

Another idea might be a fully new creation of a metadata standard especially for computer-based assessment specifying missing parts like complex item types, log file standards or scoring rules. Nevertheless, the organizational and political dimensions of such a project are huge. The development cycles for versions of Dublin Core, DDI or QTI take years with a multitude of organizations being present in these alliances. Building up structures like these with an appropriate number of members who then also commit to use the standard in their respective organizations takes very long. In the case of DDI it started beginning of the 1990s with the first version. Specifying a completely new standard which also has to compete with existing standards might therefore not be the ideal way to go.

Development of a model for software development based on the results

Unfortunately, the further development of standards will not provide a quick solution for the lack of metadata management in the educational sciences. Therefore, the question is how these problems can be handled in the meantime while the standards are progressing. The chapter will therefore discuss some scenarios to improve the current situation of a gap between demands from researchers, processes available in data management and software developed by a multitude of vendors.

Combining metadata standards for educational research as an interim solution

As the literature review has shown the current metadata standards are not sufficient to represent content from computer-based assessment. Furthermore, until DDI, QTI or other standards have filled the gaps and white spaces more time will be lost. Interim solutions are therefore necessary to bridge the development time for the metadata standards. To show a possible matrix of metadata standards the following two axis have been built. On the one hand processes from the Generic Longitudinal Business Process Model (GLBPM) have been taken as they represent the typical workflows which are also common to surveys from the educational sector. On the other axis the following areas according to the previous survey have been identified:

- Survey information (describing the research project from a scientist's perspective)
- Education-systematic information (metadata information necessary for educational institutions like schools and universities)
- Questionnaires (as they are also very common with educational researchers)
- Simple item types (e.g. multiple choice items)
- Complex item types (e.g. simulations)

The preferred standards to be used in this system are DDI Lifecycle 3.2 and QTI v2.1 as they were the most familiar to educational researchers in the survey though in general the knowledge about metadata standards themselves can be considered to be low. DDI Lifecycle 3.2 provides another advantage which can be used for combining it with other standards. Elements which do not exist within DDI but are necessary to represent content for an agency can be defined as a user-attribute pair (DDI, 2009) for a custom extension. It is therefore possible to embed another standard into DDI without violations. Nevertheless, these user-attribute pairs are only understood by the agency which originally created them or other agencies using the same layout. A completely unaware DDI instance will not be able to parse user-attribute pairs defined by another agency without implementing the same layout.

As can be seen the majority of processes can be handled within DDI 3.2 while some extension can be done by using other standards:

- Educational-systematic information can be handled by using the Learning Object Model though it has to be observed LOM is actually adapted to the educational system of different countries (e.g. UK-LOM, LOM-CH). This means the correct "flavour" of LOM has to be chosen. Furthermore it would be possible to embed LOM into a DDI 3.2 structure using user-attribute pairs.
- Simple cognitive items can be handled by using QTI v2.1 as a standard and similar to the procedure for LOM it can be represented by user-attribute pairs within DDI 3.2 enabling the integration from one standard into the other.
- There is no good solution for complex items as none of the standards is able to handle them. A workaround could be to store additional information about the item (e.g. screenshots, videos, graphical elements, documentation about the workflow) in archival formats and try to preserve the original software code by using Open Source repositories thus preserving as much as possible. Hopefully the challenge of complex cognitive items will be addressed by a metadata standard for education in the future.
- Unfortunately, data collection procedures and the paradata derived from it are also not handled very well within DDI 3.2. It is therefore necessary to document information like sampling, disposition codes, interviewer information or field reporting into archival formats which can easily be imported and exported between databases. A recommendation would be specifying a propriety XML-based scheme including a description of the layout as long as metadata standards do not offer any support for paradata.

Furthermore, there are additional challenges derived from the details of the survey.

- There is no format to handle scoring rules. QTI offers a limited set of rule definitions but they are not sufficient for complex workflows. If there are rules which cannot be expressed in QTI they have to be documented externally used procedures like described above.
- The same problem exists for log file metadata. There is no unified format for these information. It is therefore necessary to document the layout of the log files so they can be analysed in the future.

A metadata and data repository following these rules will be able to store computer-based instruments and the datasets derived from it in a meaningful way. Of course the information in this paragraph is not detailed to develop such a system from scratch. This would also not be possible as the content has to be adapted to the requirements of the researchers within the facility. It can nevertheless serve as a blue print for first design considerations which then have to be extended by an internal requirement analysis between the researchers, data management personnel and computer-scientists if the software is developed in-house.

Embedding data management personnel into the software development cycle

As the last paragraph of the previous chapters describes embedding data management into an organization involves three players. From the survey we know neither educational researchers nor computer-scientists on their own have sufficient knowledge about data management processes to properly embed them into their research or software development cycles. Thus, data management personnel has to be embedded into those processes.

This can be done by the following ways:

- Data Managers can act as advisors to researchers for choosing the proper software products, metadata standards and processes to support their survey work
- Data Managers can act as advisors to software developers who develop survey tools, assessment software, data collection software or repository software to support processes from the very beginning
- All three groups can be involved in the development of new processes and new standards at certification bodies or alliances with their expertise thus creating an data management supported research process

For this vision to come true it is needed researchers see the benefit of sharing their data and enriching them with metadata in the first place. For the researchers stating in the survey they are willing to support these processes this needs convincing arguments and a change of paradigm. In a world where financial resources for conducting surveys gets less and less there might be no other option then strengthening the usage of high quality datasets from research data centres with additional research questions and analysis.

Data management and computer-based assessment

A clear result of the dissertation is the necessity of data management for computer-based assessment. The question should therefore not be if data management has to be implemented but rather how this task can be done. Unfortunately, it can be stated educational scientists are in the normal day-to-day work not aware of the problems the lack of data management produces surrounding their items, instruments and datasets. Given the current trend towards computer-based assessment the problem is very likely to increase over time. This means in a worst case scenario in 15-20 years there will be a dramatic loss of metadata and data as nobody is able to reproduce the items and results from studies which were not big enough to have own data management procedures or were considered to be sufficiently important for a research data centre to actively inquire for it. It can only be hoped this dissertation or similar movements from data managers will result in software which supports data management procedures as a by-product of the item development and data collection process as it can be considered unlikely stakeholders will start extra funding for more advanced data management procedures. Nevertheless it can be expected stakeholders will pressure scientists more via funding regulations to embed data management procedures, e.g. formulating the necessity to have a data management plan for a survey. Currently more and more funding bodies are moving towards this direction. Unfortunately this does not necessarily lead to a certain quality as the data management plan might not be sufficient e.g. if it was designed without the help or data management personnel. The author has seen in his professional careers data management plans like "*at the end of the study we hand over the codebooks and the datasets to a research data centre*". Though it can be stated positively at least the survey is willing to share the data externally we have no idea which quality the data documentation will reach if the research data centre is not involved during the process. Actually, a valid description of the current situation is metadata and data is handed over to RDCs at the end of the study on stakeholder pressure. The datasets are very often not documented sufficiently to be handled by the repository systems and portal solutions a RDC uses and also not prepared for long-term preservation. Very often a process which could be described as "metadata forensics" starts where a scientist not involved in the original survey tries to reconstruct information about it. This process is time-consuming and might provide errors as the work is not done by an insider but an external scientist who has to resort to interpretation. Software and metadata standards have thus to become more user friendly and support the full survey process ideally in one software package which can handle the whole workflow or at least the interfaces have to be designed to support a seamless migration from tool to tool.

References

Barkow, I., Block, B., Greenfield, J., Gregory, A., Hebing, M., Hoyle, L., & Zenk-Möltgen W. (2012). Generic longitudinal business process model. *DDI Working Paper Series, Longitudinal Best Practices II, 5()*, 1−26. Retrieved from http://ddionrails.org/glbpm/GLBPM.pdf

Csapó, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues of computer-based assessment of 21st century skills. In Griffin, P., McGaw, B. & Care, E. (Eds.) *Assessment and teaching of 21st century skills*, New York: Springer, 143–230.

Csapó, B. (2011). *Developing an online assessment system: Results of the first phase and future plans*. Third Szeged Workshop on Educational Evaluation (SWEE). Szeged, 27 April 2011. Retrieved from: http://www.edu.u-szeged.hu/swee/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd Edition, Hillsdale.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Data Documentation Initiative (2009). *Technical specification, Part I: Overview, Version 3.1*. Retrieved from http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/

Data Documentation Initiative (2012). *Developing a model-driven DDI specification*. Retrieved from http://www.ddialliance.org/system/files/DevelopingaModel-DrivenDDISpecification2013_05_15.pdf

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160.

Generic Statistical Business Process Model (2013). *Introduction to GSBPM v5.0*. Retrieved from http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626.

Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische mathematik*, *14*(5), 403–420.

Gregory, A., & Heus, P. (2007). DDI and SDMX: Complementary, not competing, standards. *Open Data Foundation Paper*. Retrieved from

http://www.opendatafoundation.org/papers/DDI_and_SDMX.pdf.

Greiff, S., & Funke, J. (2010). Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme. *Zeitschrift für Pädagogische Psychologie*, *56* (Beiheft), 216–227.

Harman, K. & Koohang, A. (Eds.). (2007) *Learning Objects: Standards, Metadata, Repositories, and LCMS*. Santa Rosa, California: Informing Science Press.

Helic, D. (2006). *Template-based approach to development of interactive tests with IMS question & test interoperability*. In World Conference on Educational Multimedia, Hypermedia and Telecommunications, *1*, 2075–2081.

IMS Global Learning Consortium (2012). *IMS accessible portable item protocol (APIP) Overview - Version 1.0*. Retrieved from http://www.imsglobal.org/apip/apipv1p0/APIP_OVW_v1p0.html

IMS Global Learning Consortium (2006). *IMS question and test interoperability overview - Version 2.1 Public Draft (revision 2) Specification*. Retrieved from

http://www.imsglobal.org/question/qti_v2p0/imsqti_oviewv2p0.html

Jesukiewicz, Paul (2009). SCORM 2004 4th Edition. Content aggregation model. Retrieved from: http://www.adlnet.org/wp-content/uploads/2013/09/SCORM_2004_4ED_v1_1_CAM_20090814.pdf

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association, 47*(260), 583–621.

Learning Technology Standards Committee of the IEEE (2002). *IEEE 1484.12.1-2002, Learning object metadata standard*. Retrieved from http://www.imsglobal.org/metadata/mdv1p3pd/imsmd_bestv1p3pd.html

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.

Mechtel, M. (2009). *EduDDI - an application of DDI 3.0 for large-scale assessments in education*. Paper presented at the annual meeting of the International Association of Social Science Information Service and Technology (IASSIST), Tampere, Finland, May 2009. Retrieved from http://www.ddialliance.org/node/191

Solga, H., & Wagner, G. G. (2007). A Modern statistical infrastructure for excellent research and policy advice: Report on the German Council for Social and Economic Data during its first period in office (2004–2006). *RatSWD Working Paper No. 2*.

Retrieved from http://dx.doi.org/10.2139/ssrn.1417451

Smythe, C., & Roberts, P. (2000). An overview of the IMS question & test interoperability specification. In *Proceedings of the 4th CAA Conference*. Loughborough: Loughborough University. Retrieved August 19, 2016, from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1784/1/smythec00.pdf

Punch, K. F. (2009). *Introduction to research methods in education*. London: Sage.

Ullman, J. B., & Bentler, P. M. (2003). Structural equation modeling. In I. B. Weiner (Ed.), *Handbook of Psychology, Second Edition* (pp. 607-634). John Wiley & Sons, Inc.

Vale, S. (2010). Exploring the relationship between DDI, SDMX and the generic statistical business process model. *DDI Working Paper Series*. Retrieved from http://dx.doi.org/10.3886/DDIOtherTopics01

Vardigan, M., Pascal H., Wendy T. (2008). Data documentation initiative: Toward a standard for the social sciences. *The International Journal of Digital Curation*, *3*(1), 107–113.

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). Dublin core metadata for resource discovery (No. RFC 2413).

Wright, K. B. (2005). Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, *10*(3). Retrieved August 19, 2016, from http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00259.x/full

Zabal, A., Silke, M., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I. & Rammstedt, B. (2014). *PIAAC Germany 2012: Technical report*. Münster: Waxmann.

Publications in connection to the dissertation

Amin, A., Barkow, I., Kramer, S., Schiller, D., & Williams, J. (2012). Representing and utilizing DDI in relational databases.

Amin, A., Barkow, I., Kramer, S., Schiller, D., & Williams, J. (2015). Design Considerations for DDI-Based Data Systems. *IASSIST Quarterly, 39(3).*

Barkow, I., Leopold, D. S. T., Raab, D. S. M., Schiller, D., & Rittberger, M. (2011). 20 RemoteNEPS: data dissemination in a collaborative workspace. Zeitschrift für Erziehungswissenschaft, 14(2), 315-325.

Barkow, I. (2012, October). Modelling the Lifecycle into a combination of tools. *In EDDI12–4th Annual European DDI User Conference.*

Barkow, I., Block, B., Greenfield, J., Gregory, A., Hebing, M., Hoyle, L., & Zenk-Möltgen W. (2012). Generic longitudinal business process model. *DDI Working Paper Series, Longitudinal Best Practices II, 5()*, 1−26. Retrieved from http://ddionrails.org/glbpm/GLBPM.pdf

Barkow, I., & Schiller, D. (2013, October). An Update on the Rogatus Platform. *In EDDI13–5th Annual European DDI User Conference.*

Upsing, B., Goldhammer, F., Schnitzler, M., Baumann, R., Johannes, R., Barkow, I. & Jadoul, R. (2013). Development of the Cognitive Items.

Schiller, D., & Barkow, I. (2013). Administrative Data in the IAB Metadata Management System.

Schiller, D., & Barkow, I. (2013, October). Proposing a Metadata Solution over Multiple RDCs in the German Context. In *EDDI13–5th Annual European DDI User Conference.*

Wenzig, K., Matyas, C., Bela, D., Barkow, I., & Rittberger, M. (2016). Management of metadata: An integrated approach to structured documentation. In Methodological Issues of Longitudinal Surveys (pp. 627-647). Springer Fachmedien Wiesbaden.

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014). PIAAC Germany 2012: Technical Report. Waxmann Verlag Gmbh, Germ.

The research work for the dissertation was carried out between 2011 and 2015 from my full-time position as a senior researcher at the German Institute for International Educational Research (DIPF) which also financed all travels and publications. During this period I was also enrolled as an external PhD student at the University of Szeged. Thanks to both organizations for their support.