OAIS Compliance
GESIS
Indexing Research Data
ELSST
Research Data Management
EMORY

IN THIS ISSUE
3 Papers from IASSIST Conferences

ON PAGE 5
Editor Karsten Boye Rasmussen

ON PAGE 23
Membership Information
In this issue

5 Editor’s Notes
Karsten Boye Rasmussen

6 De-mystifying OAIS compliance:
Benefits and challenges of mapping the OAIS reference model to the GESIS Data Archive by Natascha Schumann and Astrid Recker

12 Thesaurus-Based Indexing of Research Data in the Social Sciences:
Opportunities and Difficulties of Internationalization Efforts by Katrin Baum and Andreas Oskar Kempf

16 Differences among Faculty Ranks in Views on Research Data Management by Katherine G. Akers and Jennifer Doty
Editor’s notes

Research data in demystified compliance, with “glocal” indexing, and archival development

Welcome to this volume 36-2 2012 of the IASSIST Quarterly (IQ). This editorial is written in October 2013, so you might rightfully wonder what time zone IQ is in. Yes, we are truly sorry that there have been hiccups in the process. However, the good news is that we have several issues in the pipeline being compiled by guest editors. With the many efforts of these busy people we hope to catch up on the production schedule with these coming special issues.

Now, however, we have three articles in this issue with investigations emanating from institutions that we term as “data archives” or “data libraries”.

Natascha Schumann and Astrid Recker from the Data Archive for the Social Sciences, at the GESIS - Leibniz Institute for the Social Sciences in Cologne, are demystifying some central issues on data archiving in their paper “De-mystifying OAIS compliance: Benefits and challenges of mapping the OAIS reference model to the GESIS Data Archive”. Papers for the IQ often evolve from presentations at conferences, as did this one, being updated from the IASSIST 2013 conference session “Beyond Bits and Bytes: The Organizational Dimension of Digital Preservation.” The authors are exploring the use of OAIS (Open Archival Information System) as an abstract reference model as it is being mapped to the GESIS data archive. “Welcome to the real world!” The authors focus on how the OAIS is considered to be a language and as such does not present a concrete solution. OAIS has since the late 1990s shaped and influenced the digital preservation discourse. The GESIS Data Archive followed in the footsteps of ICPSR and the UK Data Archive when it tested “OAIS compliance”, when investigating the functions of the archival information system by mapping processes to the OAIS. The investigation presented in the paper will be of benefit to others working at data archives and similar institutions. By the way, the authors conclude that “OAIS compliance” is not simply a Yes/No question. Compliance is complicated!

Also from GESIS – Leibniz Institute for the Social Sciences - comes the next paper “Thesaurus-Based Indexing of Research Data in the Social Sciences: Opportunities and Difficulties of Internationalization Efforts” by Katrin Baum and Andreas Oskar Kempf. Internationalization and standardization are areas supported and enhanced by IASSIST. The authors cite the Data Documentation Initiative (DDI) as an example of an international standard for describing data, facilitating international data exchange. Before data can be exchanged it has to be identified. The authors are investigating international indexing which enhances the precision of searches for data materials. The paper highlights some international organizations offering support for indexing, for example through descriptions in the “European Language Social Science Thesaurus” (ELSST) and the “Topic Classification” used by the member organizations of CESSSDA (European data archives). International indexing and local indexing both have their pros and cons. The paper is proposing a “glocal” solution that combines and integrates positive contributions from both types of indexing.

The last paper “Differences among Faculty Ranks in Views on Research Data Management” is research carried out by Katherine G. Akers and Jennifer Doty when they worked in e-Science and data management at Emory University Libraries. Katherine G. Akers now works as a postdoc at the University of Michigan Libraries. The authors investigated how faculty researchers manage their data during and after their research projects, and their views on data sharing and preservation. The pragmatic purpose is for the libraries to develop research data services tailored to their specific needs. Researchers’ age and amount of experience are often thought to be important factors, but other studies have not shown this conclusively. This research found that senior faculty stated more often that sharing their research data requires too much time and effort. I suggest that this gives data archives further incentive to continue to improve the ease of depositing research data.

Articles for the IASSIST Quarterly are always very welcome. They can be papers from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. Authors are permitted “deep links” where you link directly to your paper published in the IQ. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is also much appreciated as the information reaches many more people than the session participants, and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to take a look at the instructions and layout: http://iassistdata.org/iq/instructions-authors.

Authors can also contact me via e-mail: kbr@sam.sdu.dk. Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
October 2013
Editor
De-mystifying OAIS compliance:

Benefits and challenges of mapping the OAIS reference model to the GESIS Data Archive by Natascha Schumann¹
and Astrid Recker²

Abstract
Since its initial publication over a decade ago, the OAIS Reference Model, its concepts and terminology, have become essential to the digital preservation discourse. In this discourse, the topos – or myth – of “OAIS compliance” continues to play a central role as archives and repositories seek to demonstrate their fitness for the challenge of digital preservation. After briefly considering what OAIS is (and can be used for) and what it is not – namely, an abstract reference model, but not an architecture that can be implemented directly –, we will use the GESIS Data Archive for the Social Sciences as an example of mapping OAIS to an existing archive. We will then explore positive effects and benefits, as well as difficulties of completing this process. Thus, such a mapping can be taxing for an established archive: As most of the workflows have grown and proven their adequacy over a considerable period of time, taking a step back and viewing these processes from a new perspective is a challenge in itself.

Keywords
OAIS, standards, trusted digital repositories

Introduction
The importance of the Reference Model for an Open Archival Information System (OAIS), which since the releases of its first draft versions in 1997 and 1999 has shaped and influenced digital preservation discourse like hardly any other model, is undisputed (see, for example, Lee, 2012; Allinson, 2006; Oßwald, 2010). The OAIS standard has not only provided us with a common language – and thereby a common understanding of what it is that archives do when they preserve digital information objects; it has also given important impulses to move towards greater standardization in the field of digital preservation, including the development of criteria and procedures to analyze and assess archival preservation and dissemination practice (e.g. ISO 16363:2012 “Audit and certification of trustworthy digital repositories”). Despite – or possibly because – of the model’s influence, the ubiquity of its terminology and concepts, one frequently encounters misconceptions as to what OAIS is and what it is for. Often, these seem to be linked to a misunderstanding of what a reference model is. On a more concrete level, it is the notion of OAIS compliance and its – sometimes seemingly unreflected – use in archive self-portrayals or in the description of software packages which appears problematic.

OAIS is a reference model
Often, one will hear or read about “OAIS being implemented” in some organization or another. What is usually meant by this is that a system is being built or adapted which conforms to the OAIS model in some way. Such statements are misleading, however, because as a reference model, OAIS can...
by definition not be directly implemented: It is an abstract and highly generic conceptualization of a preservation and dissemination environment.

Thus, as defined by the Organization for the Advancement of Structured Information Standards, “[a] reference model is an abstract framework for understanding significant relationships among the entities of some environment...” (n.d.). As such, it “is not directly tied to any standards, technologies or other concrete implementation details, but it does seek to provide a common semantics that can be used unambiguously across and between different implementations” (ibid.). This is in accordance with the purpose of the OAIS model as given in the standard itself, which among other things states that the model:

- “provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future Archives”
- “provides a framework for describing and comparing different Long Term Preservation strategies and techniques” (CCSDS 2012, p. 1-1)

The OAIS reference model can be compared to a language operating on a meta-level, allowing us to speak about archives, their architectures and processes. Therefore, OAIS will not make a certain preservation strategy or technique a requirement. It will define the characteristics of a strategy it deems successful, but it will not prescribe a concrete, monolithic solution.

This means that OAIS cannot be used as a check list which can be ticked off as one builds an archival information system. Instead, to make this meta-language useful in building such a system, a translation process is required to create an architecture from the reference model which can then in turn be implemented. In this process, abstract OAIS concepts have to be translated into concrete system elements and processes tailored to work in a specific environment.

It is for this reason that to speak of an OAIS implementation is misleading. While this may seem quibbling over details, it is important to understand that the OAIS reference model will not translate into a real-world system seamlessly, and that this has an impact on the notion of OAIS compliance as put forward in the model, and as interpreted or translated by a given archive or preservation service provider.

**A mythical creature: OAIS compliance**

Because of the OAIS model's abstract character, the notion of OAIS compliance is, as has been pointed out repeatedly, “necessarily vague” (Lavoie, 2004, p. 17). To comply with the OAIS model means complying with a set of very abstract requirements which themselves need interpretation, translation, and concretization if they are to be useful.

Thus, the standard itself makes only two requirements for compliance:

1. “Support” (itself a rather vague notion) of the OAIS information model described in chapter 2.2, including among other things the concept of information packages composed of content information and accompanying metadata.

2. Fulfill the set of mandatory responsibilities described in chapter 3.1 of the standard (see CCSDS, 2012, p. 1-3). The latter (see box 1) are high-level requirements that, as Beedham et al. note, “it would be difficult for any functioning archive not to comply with” (2005, p. 10).

Regardless of this vagueness, “OAIS compliance” has almost become a topos in digital preservation discourse, a label that is applied to repositories and their hosting institutions “to underscore [their] trustworthiness” (CCSDS, 2011, pp. 1-1). Yet this label remains largely meaningless without context and specification. Thus, for any organization or repository labeling itself as OAIS-compliant, it has to be clear what this is taken to mean — that is, how the vagueness of OAIS compliance has been translated into a concrete set of criteria in a given case. These criteria might be something so complex as those laid down in the above-mentioned ISO standard. But, as Lavoie explains, to be OAIS-compliant could also quite simply involve using “OAIS concepts, terminology, and the functional and information models” when building a digital archive or preservation system; or OAIS-compliance can be the result of a mapping process in which “the various components in the archival system [are matched with] the corresponding features of the reference model” (2004, p. 17). But OAIS compliance could also mean “explicit application of OAIS concepts, terminology, and the functional and information models” or “that the OAIS concepts and models are ‘recoverable’ from the implementation – in other words, it is possible to map, at least from a high-level perspective, the various components in the archival system to the corresponding features of the reference model” (Lavoie, 2004, p. 17).

We would argue that there are good reasons to include the OAIS functional model in compliance testing as suggested by Lavoie. Thus, in particular, it can be assumed that in order to fulfill the OAIS mandatory responsibilities, an archival information system also has to perform the functions described in the standard. Accordingly, it is the second approach described by Lavoie that the GESIS Data Archive adopted in testing OAIS compliance; it thus followed in
the steps of the UK Data Archive and the ICPSR (see Beedham et al., 2005; Vardigan & Whiteman, 2007).

Mapping the GESIS Data Archive to OAIS

The GESIS Data Archive was originally founded in 1960 at the University of Cologne as the Central Archive for Empirical Social Research (Zentralarchiv für empirische Sozialforschung), Europe’s first data archive in the social sciences. In 1986, it became a member of the newly founded Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS). Since 2007, the Data Archive is one of five scientific departments of GESIS – Leibniz-Institute for the Social Sciences, Germany’s biggest research-based social sciences infrastructure institution. It is also a member of CESSDA, the Council of European Social Science Data Archives, dedicated to improving standardized access to social science research data in Europe (see http://www.cessda.org).

Since its foundation, the GESIS Data Archive has undertaken continual responsibility for preserving social science research. Collecting primarily digital data from empirical social research, the Data Archive currently holds more than 5,100 studies equaling over 600,000 files. To consolidate and demonstrate its status as a trustworthy provider of preservation services, the Data Archive has embarked on a series of self-audit and certification activities. The first step, now almost completed, is the application for the Data Seal of Approval (http://datasealofapproval.org/). From these activities resulted a decision to test OAIS compliance by carrying out a mapping of the GESIS Data Archive to the OAIS reference model. The objectives of this mapping are the following:

- gain a more structured overview of workflows and preservation/dissemination processes;
- identify and close possible gaps in these workflows and processes;
- introduce OAIS terminology and concepts to support communication within the Archive and with other organizations.

To achieve these goals, a mapping between the Archive and the OAIS functional model, as well as an application of the concepts from the OAIS information model, are currently being carried out.

In the following, we report briefly on the procedure and first results of our functional model mapping.

Functional Model Mapping

The main tool to carry out mapping was a simple spreadsheet listing OAIS functions and the different processes/responsibilities that these comprise. For each of these sub-processes we then determined the following:

- Who is responsible within GESIS (organizational unit down to team level)?
- Is the process carried out by a human staff member and/or is it supported by a technical system?
- How is the process incorporated into Archive workflows? (e.g. is it a routine activity carried out on a regular basis?)
- Are our activities sufficient?
- Any open questions or comments

At the same time we created a simplified diagram of the current archive workflow containing the main top-level functions performed as data are acquired, deposited, archived, and disseminated (see figure 1). This helped in creating a general overview of where and when processes were taking place, and to match these with the functional entities of the OAIS model. We then started increasing the granularity of the different sections of the overview diagram by spelling out the steps carried out in a given phase and by matching them to OAIS functions.

For the pre-ingest and ingest phase this resulted in the realization that the ingest process at the GESIS Data Archive (which as a social sciences archive puts a strong emphasis on extensive quality control, data processing and enhancement) cannot be adequately captured by OAIS in this form and detail (see also Vardigan & Whiteman, 2007). Thus, quality controls carried out during ingest, include: disclosure control; technical control of the files (format, readability, presence of malware, etc.); control of completeness; plausibility, consistency and weightings; as well

![Figure 1: GESIS Data Archive digital preservation workflow](image-url)
as format conversions. If any problems are discovered, further communication with the data depositor may be necessary in order to clarify the discovered issues and to correct mistakes. Although this does not pose a problem for OAIS compliance, as the standard itself acknowledges that “[t]he complexity of this ingest process can vary greatly from OAIS to OAIS, or from Producer to Producer within an OAIS” (CCSDS, 2012, p. 4-52), it does complicate the mapping process.

It further became clear that some of the activities performed during ingest at the Data Archive are placed in different functional entities in the OAIS model. This caused us to ‘re-allocate’ some of the functions to accommodate the actual Data Archive workflow. As a consequence, our ingest comprises of functions from the OAIS functional entities ingest and administration among others, which are performed by several members of archive staff. It should be noted that this, too, is accounted for by the standard, which clearly states with regard to the functional model: “However, this is not to be taken as a recommended design or implementation, and actual implementations are not expected to have a one-to-one mapping to the functions shown, and may for example choose to combine functions or break out functionality differently” (CCSDS, 2012, p. 4-3). Yet, this makes the mapping less straightforward and hence more time-consuming.

Benefits and Challenges
We primarily benefited from the mapping in three areas: communication, self-reflection, and process evaluation. As noted in Beedham et al. (2005, p. 82), OAIS – with its clearly defined vocabulary – can support communication within, or between organizations, by offering a common language. Thus, one stated purpose of the OAIS standard is to provide the digital preservation community with a vocabulary composed of terms “that are not already overloaded with meaning so as to reduce conveying unintended meanings. Therefore it is expected that all disciplines and organizations will find that they need to map some of their more familiar terms to those of the OAIS Reference Model” (CCSDS, 2012, p. 1-5). This mapping process has begun at the GESIS Data Archive, and while the introduction of this new vocabulary and its establishment in everyday communication is a gradual process taking its time, OAIS terminology’s potential to help ensure that staff are speaking about the same things is already apparent.

At the same time, mapping the elements and processes of the GESIS Data Archive to the OAIS functional model and vocabulary has fostered self-reflection. As mentioned above, the Data Archive has grown over decades, and while we are certain that our digital collections are in expert hands at the Archive, mapping to OAIS gives us the opportunity to – figuratively speaking – take a step back to analyze our daily routines and procedures. Looking at these routines through “OAIS glasses” and applying OAIS terminology to them, has helped us gain a more systematic understanding of the workflows that take place as we go about our daily work, the communication processes supporting them, and the roles and responsibilities of the actors involved.

Finally, undertaking the mapping helped us not only to identify and name processes, it also allowed us to evaluate them. Thus, we were able to spot gaps in our routines and to plan and implement strategies to close them.

On the other end of that scale, we find OAIS functional entities and functions which seem relatively simple and straightforward, but turn out to be highly complex in mapping. Such unexpected complexity can occur particularly when single functions are performed jointly by several teams and/or departments. In this case, mapping the function entails documenting all the communication processes taking place between the actors and the systems involved in this process. As Beedham et al. observe for the data management function, this leads to an explosion of mappings to all the different systems and processes that an archive performs (2005, p. 47). Another example of this is ‘Establish standards and policies’; which is part of the administration function, and which at the GESIS Data Archive cuts across different teams and departments (see table 1)

Formalism vs. pragmatism
As Beedham et al. observe with regard to the UK Data Archive and The National Archive, “the OAIS standard can sometimes be overly bureaucratic and over-concerned with processes. Realistically organisations like UKDA have to be more pragmatic in their approach to decision making. . . . The OAIS reference model only provides a formalised view of the functions of digital archiving; it does not prescribe implementation strategy or management style. Nevertheless, a real archival organisation never operates quite as cleanly as the OAIS model envisages” (2005, p. 53).
The same is certainly true for the GESIS Data Archive, as problem-solving, planning, and decision-making processes can be less formalized than those described in the OAIS functional model. Thus, we will often find that the Data Archive performs all the processes of which an OAIS function is composed. However, many of these processes take place as part of routine, team (or department) internal communication, which may take many different forms (as well as degrees of formality) and which will not always be explicitly labeled as pertaining to a given OAIS function. Thus, the fulfillment of a certain OAIS function becomes a side-effect of certain communication processes.

Without question, the mapping process helped us identify possibly critical processes where we need to introduce more formality; for example, in the case of the Monitor Technology function. Greater formalization may also be needed in cases where functions, which according to OAIS communicate with each other, are fulfilled by one and same person. As in this case no real communication (e.g. in form of requests and responses to these requests) takes place, the use of additional documentation tools (e.g. check lists) will have to be considered to create more transparency and to ensure that all necessary steps are taken.

However, as Beedham et al. point out, the OAIS reference model only lists and describes functions without specifying how exactly they should be implemented (see 2005, p. 53). This means that it is really up to an archive to decide how frequently, and in which form, a process should take place. Does fulfilling the “Establish Standards and Policies” function require regular meetings (on team, department or institutional level)? Or, can certain processes be adequately addressed in ad-hoc communication between the staff members involved?

As already mentioned, what helped in assessing our current practice in the course of the mapping process, was to identify the level of “formality” with which a certain OAIS function is fulfilled by the Data Archive. This was achieved, for example, by indicating

- whether a function is performed routinely (that is on a regular basis, or for every dataset submitted to the archive), or only on request/as required (and hence reactively rather than proactively; on this aspect see also Beedham et al., 2005, p. 53);
- whether checklists, minutes, or other records exist to document the process;
- and whether in our opinion the level of formality was sufficient or not.

However, in some instances the mapping may also result in a conscious decision to deviate from OAIS entirely for reasons dictated by the specific environment in which the archive operates.

**Degrees of compliance**

As the GESIS Data Archive’s experience with the ongoing mapping process illustrate, OAIS compliance – in contrast to compliance with a certification standard or criteria catalog such as ISO 16363:2012 – is not 1 or 0, Yes or No. Rather, we would argue, compliance comes in degrees.

Thus, what OAIS compliance means is really a matter of interpretation, an act of filling in the gaps that the reference model necessarily leaves with context information from of our own organizations. Sometimes, this context may even lead to the decision to not comply with a certain aspect of OAIS. We would argue that such decisions, as long as they are well-founded, do not necessarily compromise OAIS compliance. However, this illustrates once more that without providing enough of the context information specific to the archival system (which as a reference model OAIS must necessarily ignore), and without spelling out and making transparent the acts of interpretation performed in translating the reference model into something like a checklist, the statement “We are OAIS compliant” remains utterly meaningless.

**References**


NOTES

1. Natascha Schumann is affiliated at the Data Archive for the Social Sciences at the GESIS Leibniz Institute for the Social Sciences in Cologne. The main focus of her work is on digital curation of social science research data and audit and certification in this area. Her contact email is natascha.schumann@gesis.org.

2. Dr. Astrid Recker works at the GESIS Data Archive where she is responsible for the design and delivery of digital preservation workshops for the “Archive and Data Management Training Center.” She is also involved in the EU-funded project Data Service Infrastructure for the Social Sciences and Humanities (OASIS4). Her contact email is astrid.recker@gesis.org.

3. This paper is an updated version of a presentation given at the IASSIST 2013 conference in the session “Beyond Bits and Bytes: The Organizational Dimension of Digital Preservation.”

4. A similar observation is made in Beedham et al. for the administration function (2005, p. 48).

5. It is similarly clear, however, that OAIS terminology will not become the only vocabulary with which the Data Archive operates. For example, the communication with stakeholders – data producers...
and users in particular – requires the use of a different, less “technical”
vocabulary. Similarly, internal communication, too, will continue to
use non-OAIS terms and concepts where we regard these as (more)
appropriate. Yet, it is important to be able to use OAIS terminology
as a point of reference in cases of ambiguity or unclarity.
Abstract
Efforts towards internationalization have become increasingly important in scientific environments. As for content-based indexing of scientific research data, however, standards leading to internationally coherent indexing which is vital for retrieval purposes are not yet sufficiently developed. Even concerning the concrete use of indexing instruments, launched by initiatives on an international scale, there are still no binding policies and guidelines. Against this backdrop, essential criteria which internationally applicable indexing systems should meet will be outlined. These will be illustrated through the multilingual European Language Social Science Thesaurus (ELSST), originally based on the UK Data Archive’s (UKDA) Humanities and Social Science Electronic Thesaurus (HASSET) and ultimately developed by the Council of European Social Science Data Archives (CESSDA). Additionally, the general pros and cons of using international versus national indexing languages will be weighed using the ELSST and the Thesaurus for the Social Sciences (TSS) developed by GESIS – Leibniz-Institute for the Social Sciences. In this light, the benefit of vocabulary crosswalks for supporting a combined use of international and national indexing systems will be discussed.

Keywords: research data, cataloguing, thesaurus, internationalization, social sciences.

Introduction

Over the past several years, multiple efforts pertaining to the standardization of workflows, working instruments and working methods have been undertaken in various scientific domains. These efforts have been at national levels and, increasingly, on an international scale because standardization both supports and facilitates interoperability between cooperating institutions.

In the field of the social sciences the Data Documentation Initiative (DDI) metadata specification has been developed to serve as an international standard for describing data from the social sciences and related disciplines. By using this standard, coherent documentation across institutions in different countries is ensured and data exchange facilitated.

Internationalization and standardization efforts can also be observed in the context of subject indexing. The use of commonly applied indexing systems or the mapping of dispersed terminological resources is an attempt to support subject retrieval across distributed collections.

Subject Indexing of Research Data in the Social Sciences in Europe

The Council of European Social Science Data Archives (CESSDA), founded in the 1970s, is an umbrella organization for European social science data archives. Membership is comprised of data archives and other organizations which archive and provide social science data for secondary use. CESSDA currently provides access to 25,000 datasets with the collection growing
by approximately 1,000 datasets annually. Among other functions CESSDA is responsible for the development and maintenance of the European Language Social Science Thesaurus (ELSST) and the Topic Classification, both of which are used for subject indexing of research data by the member organizations. The CESSDA Catalogue enables retrieval of data stored at CESSDA archives throughout Europe and provides besides free-text search options for searches by topic of studies indexed with the Topic Classification or searches by keyword for studies indexed with ELSST.

The European Language Social Science Thesaurus (ELSST) is used for subject indexing of research data by the CESSDA member organizations. It is based on the subject thesaurus HASSET which is hosted by the UK Data Archive and is being further developed by the CESSDA thesaurus management team.

It is a multilingual thesaurus for the social sciences and has been translated from English into Danish, Finnish, French, German, Greek, Norwegian, Spanish and Swedish. It consists of approximately 3,300 concepts extracted from HASSET. These concepts aim to be culturally neutral thereby reflecting a European perspective instead of one that is country-specific, thus allowing international applicability of terms. Furthermore, ELSST allows for the addition of local extensions, which means that concepts of local importance can be added to meet institutional needs.

Currently, indexing practices using ELSST vary widely across the participating institutions due to the lack of binding indexing guidelines. For example, indexing specificity ranges from the description on a very general level with only some descriptors for one study to a very precise and deep indexing with more than a hundred descriptors for one study. This can lead to the result that the same issue is very differently described. This again has implications on retrieval as the same issue can only be retrieved by using different search terms – a fact that users will not be aware of.

Additionally, due to the requirement that concepts be internationally applicable, fine-grained local issues as well as historical, juridical, religious, political and other country-specific aspects cannot be displayed if using solely ELSST. Consequently, retrieval is limited to internationally valid concepts.

**Thesauri in Subject Indexing**

Thesauri are being used for verbal subject indexing in documentation. Consistently applying the same, controlled descriptor for specific issues results in consistent documentation and facilitates retrieval.

Indexing systems are usually based on specific collections, meaning that content and structure of systems even in the same domain can differ considerably. As well, levels of abstraction and hence specificity can vary among different thesauri depending on local conditions and needs. Moreover, different classification aspects following from a variety of perspectives on a topic can lead to different semantic relations between concepts.

**Requirements of an internationally applicable thesaurus**

One of the most important requirements for an internationally applicable thesaurus is that it be free of bias. The concepts it contains need to exist in every participating culture and have to be displayed in a hierarchical and semantic structure that fits all cultures and languages. Terms for concepts have to be multilingual to allow access in all of the languages in use.

However, the characteristics of an internationally usable system such as this include numerous limitations and constraints. Fine-grained issues at both the institutional and country-, respectively language-specific level cannot be displayed; thus retrieval is limited to internationally applicable concepts.

**Local indexing systems**

Local indexing systems are able to reflect the scope of the local collection very accurately and with respect to cultural characteristics. This allows for more precise indexing. Beyond that, they are easier to maintain as there are no cross-institutional agreements to follow.

On the other hand, the exclusive use of a local indexing system has its own deficiencies. Since it remains a solely locally applied system, without further measures there can be no access points for unified subject retrieval across dispersed collections that have been indexed using different terminological resources.

**Recommended Indexing Model**

One possible way to offer uniform subject access to heterogeneously indexed collections in a dispersed environment is the mapping of institutionally used indexing systems (Doerr 2001). Applied on retrieval, mappings aid the user in finding documents indexed no matter which indexing system was used by being able to only employ search terms in the system he or she is familiar with.

Though, mappings often carry a certain amount of intrinsic vagueness due to incomplete congruity between concepts of different indexing systems. For this reason we propose an aggregate of local thesauri with

![Figure 1: Interplay between local and international indexing system.](Image)
To reiterate, we would argue for a direct interplay between international and local indexing system in a way that will permit internationally applicable concepts standing for the core vocabulary to be fully integrated, i.e., represented, in all of the different local indexing systems attached to it. Consequently, the whole content of the core vocabulary is simultaneously part of every single local indexing system. This is the reason that the universal core system, i.e., the ELSST, must contain all central concepts which exist in all of the included languages. And, vice versa, these central concepts must be integrated into local indexing systems. For instance, the key indexing tool for German-language social sciences, the Thesaurus for the Social Sciences (TSS), translated into English and French, contains more than 8,000 concepts and, like the ELSST, includes a wide range of subdisciplines of the social sciences.

Looking at this direct interplay between local and international indexing system in practice, we explicitly advocate for further use of the local system as key indexing system. Taking “secondary school” as an example from the education sector, this concept, referring to the local indexing system introduced so far, needs to be, if not already the case, incorporated into the TSS. To summarize, missing concepts which are part of the internationally used core vocabulary must be created in the local indexing systems.

In addition to these internationally applicable concepts, the local indexing system must also contain any locally distinctive specificities. For example, the German concept “Gymnasium”, stands for a certain type of secondary school, one with a strong emphasis on academic learning. It is comparable to the British grammar school system or preparatory schools in the United States. Moreover, the local indexing system contains collection-specific concepts, e.g., the geographic subject heading “Nordrhein-Westfalen”, a federal state in Germany, indicating, for instance, the provenance of a dataset.

It becomes clear that a connection between internationally used core concepts and local indexing systems is necessary. In our judgment, this linkage could be best achieved by terminology mapping between international and local indexing system. Referring again to the two thesauri mentioned above, we like to hint at the major terminology mapping initiative conducted by GESIS - Leibniz-Institute for the Social Sciences as part of the project Competence Center Modeling and Treatment of Semantic Heterogeneity (KoMoHe) (Mayr & Petras 2008). Carried out shortly after the ELSST extension in the framework of the European Commission project Multilingual Access to Data Infrastructures of the European Research Area (MADERA) its main objective was to create crosswalks between various controlled vocabularies and also between the ELSST and the TSS. Approximately 2,300 equivalent relations were built up in each direction which could be reused when translating ELSST vocabulary for inclusion into the TSS. A search using the ELSST-concept “secondary schools”, would directly create a link to datasets indexed with the German term “weiterführende Schule”, and respectively into their English and French translations.

Additionally with the help of these crosswalks, the extent to which the international core system is already part of local indexing systems becomes apparent. Moreover, looking at these crosswalks from the opposite direction, in our case from the TSS to the ELSST, and looking at non-equivalent relations of this mapping, which had also been built up in the past, gives a hint at local specificities being part of the local indexing system. For example, the German concepts “Gymnasium”, “Realschule” and “Hauptschule” are narrower terms of the above-mentioned concept “weiterführende Schule”.

**Retrieval aspects**

We suggest there are significant information retrieval benefits to be obtained as a result of the direct interplay that occurs between internationally applicable concepts and local indexing systems. First of all, using an integrated retrieval system, e.g., the CESSDA Catalogue, the researcher is able to use proper terminology for core concepts included in the multilingual ELSST. Due to the hierarchical semantic structure of the thesaurus, the researcher is aided in the search for narrower subject terms. At this point the researcher has access to all the data collections of the CESSDA member organizations indexed with those commonly shared international core concepts. For fine-grained regional datasets with existing vocabulary mappings between international and local indexing system, a linkage between both vocabularies could be established. Similar to the CESSDA catalogue’s tree-like hierarchical search structure, additional local specific terms could be connected to the broader terms in the international indexing system. Thus, datasets indexed with the specific subject headings become searchable and accessible. It will prevent locally embedded information from being buried under broad general indexing terms.

![Figure 2: Application of the indexing model for retrieval.](image)

**Conclusion**

In a period when efforts towards standardization and internationalization of subject indexing of research data have become increasingly important, there is a pressing need to determine ways to integrate local indexing systems into widely launched internationally applicable vocabularies. Even though work on this began in the 1970s, there are still no binding and coherent indexing guidelines. Vast differences in cross-country indexing of research data remain.

With this as our backdrop, our aim was to present a concrete proposal on how to create an interconnection between local and international indexing system. Hence, we argued for an aggregate of local thesauri with a common core vocabulary of internationally applicable
key concepts. These concepts would exist in all of the participating languages represented in CESSDA. The vocabulary would be kept free of bias as much as possible. Concepts already integrated in local systems would be mapped to the core system and any missing concepts in the local systems would be added. Doing this would achieve a coherent and unified subject indexing of dispersed collections of research data. Information retrieval would be significantly improved. Fine-grained institutional and locally distinctive datasets that have been indexed with collection-specific subject headings will become searchable via crosswalks built up to the international indexing system. The local indexing system will remain the key tool as it is much easier to maintain.

Concrete efforts to move forward will include as a first step the review of ELSST to ensure no bias in the concepts and the interconcept-relations. Following this, the second step is to adapt the local systems. Subsequently, universally accepted indexing guidelines for the core system need to be developed.

References

NOTES
1. Katrin Baum is a librarian at GESIS – Leibniz-Institute for the Social Sciences in Cologne, Germany, where she works in the Data Archive department in the area of study descriptions. Her professional focus is on subject indexing and information retrieval. Katrin can be reached by email: mailto:katrin.baum@gesis.org.

2. Andreas Oskar Kempf is a research associate at GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany. He holds a PhD in sociology from Goethe University, Frankfurt am Main, and received a Master’s degree in Library and Information Science from Humboldt-University, Berlin, Germany. Andreas conducts applied research on GESIS authority data for content cataloguing (i.e. Thesaurus for the Social Sciences) of social science research literature, projects and data. He can be reached by email: mailto:andreas.kempf@gesis.org.
Abstract

Academic libraries are increasingly engaging in data curation by providing infrastructure and services to support the management of research data on their campus. Efforts to develop these resources can benefit from a greater understanding of the social factors that affect how researchers manage their data during and after their research projects. In particular, the age or amount of experience of researchers is often thought to be an important factor influencing their viewpoints on research data sharing and preservation. In this study, we categorized faculty members who responded to our campus-wide survey on research data management into four ranks—professor, associate professor, assistant professor, and non-tenure track—and analyzed differences in their patterns of survey responses. We found statistically significant differences among faculty ranks in familiarity with funding agency requirements for data management plans, reasons that might prevent data sharing, and interest in potential research data services. These findings reveal key distinctions among different ranks of faculty members in their outlook toward research data management, which can help guide academic librarians and data curation professionals to develop research data services that are tailored to the unique needs of specific populations of researchers.

Keywords: Data curation, research data management, data sharing, researchers, faculty rank, library

Introduction

Academic libraries are increasingly providing support for the management and dissemination of research data by offering infrastructure (e.g., institutional repositories), services (e.g., consultation on data management plans), and education (e.g., best practices in data management) to campus researchers (ACRL Research Planning and Review Committee, 2012; Fearon, et al., 2013; Heidorn, 2011; Monastersky, 2013). To build research data management support systems that are both effective and desirable by researchers, academic librarians and other information professionals have conducted surveys and interviews with researchers to further understand how they manage data throughout the research lifecycle and their opinions on issues such as data sharing (e.g., Bardyn, 2012; Jahnke & Asher, 2012; Scaramozzino et al., 2012; Wells Parham et al. 2010; Westra, 2010; Witt et al, 2009). These investigations indicate that researchers exhibit a myriad of approaches to managing their data depending on their discipline, research topic and methodology, source of funding, data privacy concerns, and collaborative networks.

The age or amount of experience of researchers is another factor that may influence data management actions and attitudes. Within conversations among information professionals, two assumptions are often made: (1) younger researchers have been raised in a culture of greater openness of information and therefore are more willing to share their data, and (2) researchers nearing retirement are concerned about their research legacy and therefore are more eager to preserve their data. These assumptions, however, are primarily based on anecdotal evidence or small numbers of researcher interviews (e.g., Office of Policy and Analysis, 2011). Moreover, the few formal investigations on this topic have yielded contradictory results. For instance, Kuipers & van der Hoeven (2009)
found that less experienced researchers (<10 years of experience) were more willing to deposit their data into a disciplinary repository than more experienced researchers (>20 years of experience). Likewise, Piwowar (2011) found evidence suggesting that younger researchers are more likely to share their data than older researchers. By contrast, Tenopir et al. (2011) found that junior-ranking economy professors were less likely to share their data than full economy professors. Therefore, the nature of the relationship between researcher age/experience and tendency to preserve or share their research data remains in question.

To further explore how the age or amount of experience of researchers is related to their views on managing data throughout the research lifecycle, we took faculty rank into account when analyzing the results of our campus-wide survey of researchers’ practices and perspectives on research data management. Specifically, we categorized faculty member respondents into four different ranks—professor, associate professor, assistant professor, and non-tenure track—and searched for differences in their patterns of survey responses.

**Methods**

In the fall of 2012, Emory University Libraries, in cooperation with the Office of Institutional Research, Planning, and Effectiveness, administered an online, 13-question survey on research data management practices and perspectives using Qualtrics software. A link to the survey was sent via email to all Emory University employees with faculty status according to Human Resource records (N = 5,590). The survey was open for 4 weeks, and three email reminders were sent at 1-week intervals. The survey was initiated by 456 faculty members (~8% response rate).

Our analysis focused on respondents who answered ‘yes’ to an initial question of whether they conducted research that generated some type of data (e.g., spreadsheets, text, images, videos, audio files, instrument files, photographs, physical samples/specimens, etc.; n = 330). Due to difficulties in equating rank among tenure, clinical, and research tracks, faculty members with clinical or research track designations were not included in the analysis. The remaining faculty members (n = 210) were divided into four groups based on Human Resource records: professor (professor, professor emeritus, or dean), associate professor, assistant professor, or non-tenure track—and searched for differences in their patterns of survey responses.

Differences in survey responses among the four ranks of faculty members were evaluated using chi-square (χ²) tests. Statistical significance was set at p < 0.05. Data are shown only for survey responses for which there were statistically significant differences among ranks. Complete survey results, including differences among arts & humanities, social science, basic science, and medical science domains, were previously reported (Akers & Doty, in press).

**Results**

**Data Management Planning**

We found no significant variations among different ranks of faculty members in the amount of research data they were storing or their methods for data storage and back-up (e.g., computer hard drive, external hard drive, instrument hard drive, university server, internet-based storage, lab notebooks, discs/tapes).

![Figure 1](image)

**Question:** How familiar are you with the requirement for data management and/or sharing plans as components of many funding applications?

**Response:** Somewhat or very familiar.

However, we did find variations among faculty ranks in their familiarity with federal funding agency requirements (e.g., National Science Foundation (NSF), National Institutes of Health (NIH), National Endowment for the Humanities (NEH)) for data management or data sharing plans as components of grant applications (χ² (3, n = 210) = 13.5, p = 0.004; Figure 1). The majority of full and assistant professors stated that they were either somewhat or very familiar with data management plans, and over half of associate professors also expressed familiarity with these requirements. By contrast, most non-tenure track faculty members were not familiar with data management plan requirements, which may reflect a greater focus on teaching and less reliance on research grants.

**Data Sharing**

Faculty rank did not predict faculty members’ willingness to share their research data with other people (e.g., researchers working on project, researchers outside of project, funders, instructors, general public) or their method of sharing research data (e.g., email upon request, supplementary material linked to journal article, data repository, university or personal website).
However, different ranks of faculty members expressed different opinions on why they might not share their data. Specifically, full and associate professors were more likely than assistant professors and non-tenure track faculty members to state that it takes too much time or effort to share their research data ($\chi^2 (3, n = 199) = 10.1, p = 0.018$; Figure 2). This finding may reflect that senior-ranking faculty members may simply feel they are too busy to organize, document, and compile their data into shareable data packages that can be understood and used by others. Alternately, junior-ranking faculty members may feel that sharing their data with others is an expected part of the research process and thus may not perceive preparation for data sharing as an imposition on their time.

There were no differences among faculty ranks in other reasons that might prevent data sharing, including having data that contain private or patentable information, having data that require restricted access, fear of not getting credit for their data, fear of possible misinterpretation or misuse of their data, or belief that their data are of little use to others. Different ranks of faculty were also equally likely to deposit their data in data repositories or express familiarity with data documentation and metadata.

**Discussion**

It is often assumed that younger researchers are more supportive of open data and therefore more likely to share their research data with others via websites or data repositories/archives (e.g., Johnson, 2008; Lin, 2013; Boulton, 2013). However, empirical studies have not consistently provided support for this assumption. Although evidence from Kuipers & van der Hoeven (2009) and Piwowar (2011) suggests that younger researchers are indeed more willing to share or archive their data than older researchers, our survey failed to find differences among ranks of faculty members in their willingness to share research data or their preferred method of data sharing. Moreover, Tenopir et al. (2011) and Andreoli-Versbach & Mueller-Langer (2013) found that younger researchers were less likely to share their data than older researchers. Therefore, younger age or less research experience may not always be a predictor of data sharing.

Rather than a simple correlation between researcher age and willingness to share data, findings by Tenopir et al. (2011) suggest that the situation is more complex. Their survey revealed that younger scientists were less likely than older scientists to place their data in a central repository without restrictions. However, younger scientists were slightly more likely than older scientists to make their data available if they could place conditions on data management plans, consultation on data confidentiality and/or legal issues, personalized consultation on research data management for specific researchers or research groups, an institutional repository for research data, assistance with data documentation or metadata creation, research data management workshops for trainees (i.e., graduate students or postdocs), digitization of physical research materials, assistance identifying appropriate disciplinary data repositories, or methods for data citation.

**Interest in Data Services**

In the final survey question, we offered a list of ten potential research data services and asked faculty to select which services they would use if available. The service garnering the most interest was faculty workshops on general data management. This service was desired by non-tenure track faculty members more than by assistant, associate, or full professors ($\chi^2 (3, n = 191) = 11.6, p = 0.009$; Figure 3).

There were no rank-related differences in interest for the other potential services, including assistance preparing data management plans, consultation on data confidentiality and/or legal issues, personalized consultation on research data management for specific researchers or research groups, an institutional repository for research data, assistance with data documentation or metadata creation, research data management workshops for trainees (i.e., graduate students or postdocs), digitization of physical research materials, assistance identifying appropriate disciplinary data repositories, or methods for data citation.
re-use, such as requiring legal permission for re-use of their data or the receipt of a complete list of products that make use of their data. Andreoli-Versbach & Mueller-Langer (2013) speculate that young researchers might be hesitant to openly share their data because this could enable other researchers to use their data before they can be fully exploited for additional publications. In other words, young faculty members who have not yet secured tenure may act more competitively than their tenured counterparts, choosing to withhold their research data or impose restrictions on data re-use. As we found no rank-related differences in faculty members’ tendency to state that they might not share their data due to fear of not getting credit, our results do not directly support this possibility. Nevertheless, junior-ranking faculty members may be the ideal target population for outreach on ways of turning datasets into citable outputs of scholarly research to increase personal research impact, including assigning digital object identifiers (DOIs) to datasets, depositing data into disciplinary or institutional repositories, or publishing data papers. Younger researchers might also be particularly receptive to evidence indicating that openly sharing research data increases the citation rate of associated journal articles (Bueno de Mesquita et al., 2003; Dorch, 2012; Henneken & Accomazzi, 2011; Ioannidis et al., 2009; Piwowar et al., 2007; Piwowar & Vision, 2013; Sears, 2011).

Our survey did not contain questions about data re-use, but previous studies indicate that the age of researchers may be an important indicator of their likelihood to re-use or re-purpose other people’s data. Kuipers & van der Hoeven (2009) found that less experienced researchers are more eager to re-use data from other disciplines than more experienced researchers. Similarly, Tenopir et al., (2011) found that younger scientists are more likely to consider lack of access to data as a barrier to scientific progress that has restricted their ability to answer research questions. These findings underscore our suggestion that junior-ranking faculty members, in particular, could benefit from learning about ways to publically disseminate and thereby open their research data for re-use. Also, younger researchers may be more interested in receiving assistance with discovering and accessing pre-existing datasets.

Although we observed no differences among faculty ranks in willingness to share data, we found that senior-ranking faculty members were more likely to state that they might not share their research data due to the amount of time and effort involved. Indeed, Tenopir et al. (2011) found that insufficient time was the top reason that scientists did not make their data available to others, and others have also recognized this potential barrier to data sharing (Cragin et al., 2010; Peters & Riley Dryden, 2011; Williams, 2013). However, Vickers (2006) makes the case that it takes too much time suggests that researchers do not always develop clean and well-annotated datasets (Savage & Vickers, 2009). We speculate that this is because researchers often have no clear incentive to do so; instead, they may be motivated only to develop datasets that are sufficient to support their own analyses for a particular project and not to invest time to make their datasets understandable to other researchers or to themselves at future dates. Therefore, our results suggest that senior-ranking faculty members, in particular, may benefit from increased university investment in the management and dissemination of research data, including the creation of electronic research data management systems that could automatically organize data, generate metadata and documentation files, and push data packages into open repositories after project completion.

Finally, we found that non-tenure track faculty members were least familiar with funding agency requirements for data management plans but most interested in taking advantage of faculty workshops on general data management practices. At our university, faculty members are hired onto one of three different tracks: tenure track, research track, or clinical track. Due to difficulties in equating rank status across tracks, however, we removed research and clinical track faculty respondents from our analysis, meaning that the professional focus of the remaining non-tenure track faculty members was more likely to be teaching than other types of scholarly activities such as research. Therefore, their unfamiliarity with data management plans may be a result of their lack of dependence on external funding for research projects. Nevertheless, these faculty members also answered ‘yes’ to a question of whether they performed research that generated some type of data, indicating that they were indeed engaged in research to at least some extent. As such, our finding that non-tenure track faculty members expressed the highest desire for learning about “best practices” in research data management is very interesting and suggests that this faculty contingent, which is continuing to increase in size across academia (Curtis & Thornton, 2013), may be an overlooked population of researchers that might welcome greater outreach from academic librarians and other data curation professionals.

Acknowledgments
We thank Vincent Carter for his critical role in creating and administering the survey. We also thank members of the Emory University Libraries Research Data Management Team for their feedback on survey design and analysis.

References


Notes
1. Katherine G. Akers was the eScience Librarian and a Council on Library and Information Resources (CLIR) Postdoctoral Fellow at the University of Michigan Libraries. She can be reached by email: kgakers@umich.edu.
2. Jennifer Doty is the Data Management Specialist at Emory University Libraries. She is now the eScience Librarian and a CLIR Postdoctoral Fellow at the University of Michigan Libraries. She can be reached by email: jennifer.doty@emory.edu.
Toronto welcomes you to the 40th anniversary conference of IASSIST — the International Association for Social Science Information Services & Technology — to take place on June 3-6, 2014. The “official” hashtag is #iassist2014 #iassist40 — spread the word!

**Call for Paper and Session Proposals**

This year’s conference theme — “Aligning data and research infrastructure” — touches upon the international and interdisciplinary requirements of aligning data and research infrastructure. The 2013 OECD Global Science Forum report on New Data for Understanding the Human Condition identifies key challenges to international data collaboration that beg for new solutions. Among these challenges is the mounting pressure for new forms of social science data. In today's abundance of personal data, new methods are being sought to combine traditional forms of data collection with new forms of data collection from personal devices.

The international research community is in the midst of building a global data ecosystem that consists of a mixture of domain data repositories, data archives, data libraries, and data services and that seeks ways to facilitate data discovery, integration, access, and preservation. Evidence of this transformation is found in the recently established ICSU World Data System and in the Research Data Alliance. Like IASSIST, these organizations are contributing to the development of a global data ecosystem. Alignment, or unification of strategies, must take place at many levels to achieve this. How do we proceed? What advancements are needed in research data management, research infrastructure, and the development of new expertise?

- For more details, see the [complete call for papers document](#).
- If you're all ready to submit your proposal, use the [online submission form](#).

Your program chairs (Johan Finn, Jen Green, Chuck Humphrey) and Local Arrangements Committee (Walter Giesbrecht, Kevin Manuel, Berenica Vejvoda) and a whole host of volunteers will do their best to make this the best IASSIST conference ever!
The International Association for Social Science Information Service and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, database librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST. Join today by filling in our online application:

http://www.iaassistdata.info/