Data Service Infrastructure for the Social Sciences and Humanities

EC FP7
Grant Agreement Number: 283646

Deliverable Report
Deliverable: D5.2A & D5.2B
Deliverable Name Part A: Metadata Quality Improvement
Deliverable Name Part B: Portal Progress report

Responsible Part A: DANS
Authors: Hervé L'Hours (UEssex- UK Data Archive), Lene Offersgaard (UCPH), Marion Wittenberg (KNAW-DANS), Bartholomäus Wloka (OEAW)
Contributing: Lucy Bell (UEssex), Emily Ekstrand-Brummer (KNAW-DANS), Tom Ensom (UEssex), Mike Priddy (KNAW-DANS)

Responsible Part B: MPG-PL
Contributing: Catharina Wasner (GESIS), Matej Durco, Bartholomäus Wloka (OEAW), Stephanie Roth, Olof Olsson (UGOT), Przemek Lenckiewic, Kees Jan van de Looij, Binyam Gebreke, Daan Broeder (MPG-PL).

Work Package Leader: Daan Broeder (MPG-PL)

SEVENTH FRAMEWORK PROGRAMME

Table of Contents

List of figures

Executive Summary – Part A

The aim of this task was to analyse and compare the different metadata strategies of CLARIN, DARIAH and CESSDA, and to identify possibilities of cross-fertilization to take profit from each other solutions where possible. To have a better understanding in which stages of the research lifecycle metadata comes to the fore, we looked at several research data lifecycles and business process models. However the current research data lifecycle models have the 'static' data object as basis, whereas metadata design, redesign, creation and management can continue to be 'live' issues within the research lifecycle. We therefore developed a metadata lifecycle based closely on familiar lifecycle models but extended to support the more dynamic metadata issues.

To describe the metadata management of the different infrastructures we took a double approach. We looked on a more general level and outlined the policies and strategies regarding metadata of the three infrastructures. We evaluated these strategies on metadata quality issues with the Bruce and Hillmann criteria. On the other hand we looked with more detail how the work on metadata management is done by the individual data repositories.

The infrastructures of CESSDA, CLARIN and DARIAH differ in visions, strategies and initiatives regarding metadata issues; similarly there is a difference in metadata management among the various repositories. Despite these differences, cross fertilisation by coordination on common lists of metadata elements, sharing of knowledge, and linking resources would leverage the overall metadata quality. Evaluation of the prototype of the joint CLARIN, DARIAH and CESSDA metadata portal endorses the opinion that more coordination is needed.

Metadata quality must be discussed in relation to the activities for which they are used. We suggest that the infrastructures DARIAH and CLARIN prioritise future collaboration about standardisation efforts, which have already been initialised in dialogue between the CLARIN Standards Committee and the DARIAH representatives. Similar initiatives could be established with CESSDA.

Executive Summary – Part B

This document reports on the progress of the DASISH Joint Metadata Domain JMD, task 5.4.

Partners in this task group were: DANS, GESIS, MPI-PL, OEAW and UGOT. As a task division, MPI-PL was responsible for the task coordination and the technical infrastructure (catalogue software, metadata harvesting etc.) while the other partners contributed with their expert knowledge especially on the metadata infrastructure used in their respective communities: DANS & OEAW for DARIAH, GESIS and UGOT for CESSDA, MPI-PL for CLARIN.

The Joint Metadata Domain was implemented as a SSH metadata catalogue

software, filled with metadata harvested from metadata providing centers from the three participating infrastructures. The harvested metadata is mapped on a set of community-discussed facets and the facet values are normalized. As regards extending and configuring the catalogue software as well as developing the mapping and normalization software, considerable effort went into finding and documenting suitable metadata providers as well as mapping and normalization rules.

The original proposal in the DASISH Description of Work (DoW) also recommended using RDF technology. However, having studied the problem further and finding which technologies were already at our disposal, we decided to use semantic mappings that are implemented as schema-derived XPath specifications.

We decided on an approach with proven technology to look for and investigate the availability of SSH metadata that - according to our information - should exist, rather than experimenting with metadata search technologies. Providing a tool that can be used to inspect the metadata of the participating infrastructures. Our reasoning is that reporting on the availability of metadata is an important part of this task.

When initially confronted with the task to obtain information and documentation about existing metadata providers and metadata schemas used in the SSH, it turned out to be a more difficult task than predicted. In the end we managed to get a sufficient overview of the available metadata within the three research infrastructures: CESSDA, CLARIN and DARIAH.

PART A – METADATA QUALITY IMPROVEMENT

Guide to the Reader

This deliverable is the outcome of the work of task 5.3 of the DASISH project. The aim of this task was to analyse and compare the different metadata strategies of CLARIN, DARIAH and CESSDA, and to identify the possible mutual benefits from cross-fertilization of approaches. To support this analysis the context was defined in terms of metadata types and quality criteria and a structure was created which extended common lifecycle models to address metadata issues. This reading guide summarises the main topics of this work.

Metadata lifecycle

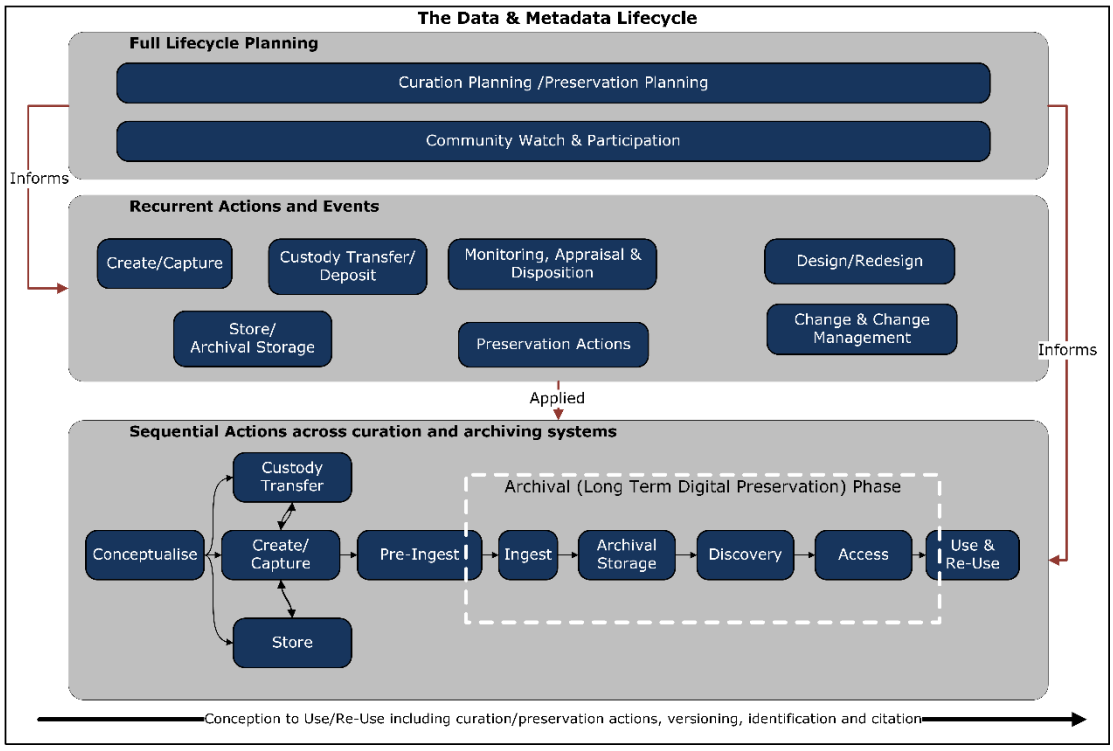


Figure 1: Data/Metadata Lifecycle to support metadata quality

Metadata design, redesign, creation and management can continue to be 'live' issues for those preserving or providing access to data even when the data itself remains unchanged. Most views of the research data lifecycle tend to treat data as fairly 'static' from the point of ingest into an Archive until the next Access/Use/Re-use cycle but repositories must apply new or update existing standards and re-enrich metadata to meet the changing needs of their target community. To support these more dynamic metadata issues we adapted existing research data lifecycles.

This metadata lifecycle may serve as a baseline, which, alongside an understanding of metadata quality evaluation and metadata types, can be used to design and benchmark a local approach to describing, delivering and

improving metadata quality. The metadata lifecycle aligns with the OAIS model, but places it in a wider context. It consists of three levels of activities:

Full life cycle planning
Communication is a key aspect within the metadata lifecycle. To deliver planning across the lifecycle the outcomes of *Community Watch and Participation* must be integrated into *Curation/Preservation Planning* processes. If your goal is to serve a community then the starting point is to engage with and understand that community. Good planning, communication and practice throughout the lifecycle reduce costs and complexity and contribute to improved quality.

Recurrent actions and events
A number of data/metadata related activities occur numerous times during the lifecycle of a digital object; these benefit from centralised design and planning so they can be implemented coherently, thereby supporting consistency and quality. These activities are defined by *Curation/Preservation Planning*, often influenced by *Community Watch and Participation*.

Sequential Actions across curation and archiving systems
Despite the fact that 'circular' approaches display more of the innate complexities of the process, the 'birth to re-use' sequence is commonly understood and support communication in day-to-day business processes. This follows the traditional research data lifecycle stages.

A more detailed description of the metadata lifecycle you find in section 3 of this deliverable. The concepts of this lifecycle are applied in the UK Data Archive case study, see appendix D.

Metadata strategies of CLARIN, DARIAH and CESSDA

We looked at the metadata policies and strategies of the three infrastructures and evaluated these in terms of metadata quality against the Bruce and Hillmann criteria. See section 4-8. Additionally we described in more detail how the individual data repositories within the different infrastructures implemented metadata management. See the case studies in Appendix D-G.

The infrastructures of CESSDA, CLARIN and DARIAH differ in visions, strategies and initiatives regarding metadata issues; similarly there is a difference in metadata management among the various repositories. Despite these differences, cross fertilisation by coordination on common lists of metadata elements, sharing of knowledge, and linking resources would leverage the overall metadata quality. Evaluation of the prototype of the joint CLARIN, DARIAH and CESSDA metadata portal endorses the opinion that more coordination would be beneficial for the metadata quality.

Background information on metadata

In the appendix to this deliverable an extensive glossary and more general background information on metadata issues is included.

1. Introduction

Researchers have access to more data now than ever before, there is an increasing number of digital repositories and more researchers make their data public. In the midst of all this information, metadata plays a very important role in organising, sharing, and finding research data. In fact, without metadata, some data within the Social Sciences and the Humanities (SSH) would be practically useless. It would be impossible to find relevant data in a repository without descriptive metadata to identify it, and without contextual metadata, some data would appear to be nothing more than a seemingly random set of numbers, images, or words.

High quality metadata (and data) is the basis for smart eScience based computation. DASISH wants to start a campaign to improve the metadata quality. In the Description of Work (DOW) of the DASISH project the aim of this task was to analyse and compare the different metadata strategies of CLARIN, DARIAH and CESSDA, and to identify possibilities of cross-fertilization to take profit from each other solutions where possible. In addition training material should be developed to raise awareness.

The relationship between metadata and data is a fluid and shifting one, and even in some disciplines the boundary between the two is very vague. However to be able to place the content of the report in the correct context we start in section 2 with a definition of the different types of metadata followed by characteristics of metadata quality. In appendix A a glossary of the terms used in the report is given. Appendix B goes into more detail on the different types of metadata, it contains the components for the training material, which is not part of this deliverable.

It is helpful to think about metadata in terms of the research data lifecycle, since it plays different roles throughout the lifecycle, and different types of metadata are used at different stages. However during our research it became clear that the research data lifecycles we investigated don't take in account the more dynamic nature of the metadata in comparison to the data. Metadata design, redesign, creation and management can continue to be 'live' issues for a repository even when the data itself remain unchanged. To support the more dynamic metadata issues we developed a metadata lifecycle. Section 3 describes this metadata lifecycle in detail; the data lifecycles on which we based our model are referenced in appendixes C1 to C6.

For the purpose of considering metadata quality the infrastructures of CLARIN, DARIAH and CESSDA can be simplified to identify key agents, systems and information (data and metadata) flows. Section 4 outlines a general infrastructure model, while sections 5 to 7 offer a description of each of the three infrastructures addressed by this deliverable. Detailed case studies describing metadata management at repositories belonging to these infrastructures are given in Appendix D to G.

In section 8 we focus on the cross fertilization between the infrastructures. The infrastructures differ in visions, strategies and initiatives; that at a first glance they might not seem to have a large overlap. However, despite this variation, how can the infrastructures cooperate to improve the overall metadata quality?

Section 9 describes the challenges of metadata quality when looking at the actual aggregation of metadata. In DASISH task 5.4 a joint CLARIN, DARIAH and CESSDA metadata portal is under development. We evaluate the preliminary results of the aggregation harvested by this prototype.

In section 10 we summarize our research and draw conclusions from our findings.

2. Metadata and metadata quality

Within this report we will assume that data and metadata may pass through the custody of several actors from the original researcher, to intermediate repositories to archives with long term digital preservation (LTDP) responsibility, the curation of data and therefore the creation and management of metadata continue throughout.

Metadata comprises an important part of the research process, and is essential for a repository to preserve and manage data. It is often defined as "data about data" (EDINA and Data Library, University of Edinburgh, n.d., Documentation and Metadata section, slide 9; National Information Standards Organization [NISO], 2004, p.1), but this definition is vague, and does not accurately represent the importance of metadata and its role in data management (Bargmeyer & Gillman, 2000).

The relationship between metadata and data is a fluid and shifting one. In fact, in Bargmeyer and Gillman's article about metadata standards, they state that, "We don't know when data is metadata or just data. Metadata is data that is used to describe other data, so the usage turns it into metadata," (2000). Metadata and the data it describes are connected to each other. Therefore, when this report refers to data in terms of the research lifecycle and data management, it is also referring to the accompanying metadata, and vice versa.

Many of those involved in the curation of data will focus on those metadata directly involved in the management of the 'digital object' which is the focus of their work, It is important to note that these assumptions may vary from data producer, to rights holder, to archivist. Metadata is critical to managing both digital objects and the processes, which surround those objects whether directly (format migration) or peripherally (management and administration of repositories). Many information professionals may tend to consider metadata primarily in the structured, controlled terms of elements and attributes presented in a structured and controlled format like XML but less structured metadata in the form of prose documentation or the formal records supporting business processes (from meeting agenda and minutes to strategies, policies and procedures) play a vital role as supporting metadata around digital objects and curation processes.

2.1. The Research Data Lifecycle and Metadata Lifecycle

It is helpful to think about metadata in terms of the research data lifecycle, since it plays different roles throughout the lifecycle, and different types of metadata are used at different stages. The actors involved in metadata and data management also change throughout the lifecycle; quality metadata cannot be single-handedly created at the end of a research project.

Information gathered throughout the entire research lifecycle is needed to create concise metadata. Therefore, everyone involved in the research lifecycle, from the researcher to the repository to the funder, must communicate with each other about their expectations and responsibilities regarding metadata (EDINA and Data Library, University of Edinburgh, n.d.; Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011; Mohler et al., 2010; Research Information Network [RIN], 2008; Vardigan, Heus, & Thomas, 2008; Wayne, 2005).

While the research data lifecycle is a valid common reference point it is important to consider that much modern data of interest to researchers was not conceived and captured with its subsequent research in mind, These data, often administrative in nature, are important resources for researchers but the reduced level of control over them, especially in the early phases of their lifecycle, presents significant challenges around data and metadata quality assurance.

During the preparation of this deliverable and the associated use cases a number of business process/lifecycle views were presented as possible structures. It became clear during the process that there is a logical and natural tendency to assume that the lifecycle is primarily concerned with the 'data' contained within the digital object itself, usually the 'study' at the centre of the research process. There is an implicit assumption that the primary activities of the lifecycle are intended to convey the canonical output of research from producer to final consumer and while this is generally accurate such lifecycles don't explicitly take account of the more dynamic nature of the metadata in comparison to the data.

Metadata design, redesign, creation and management can continue to be 'live' issues for repositories, archives and other access providers even when the data itself remains unchanged. Repositories continue to update to new standards and re-enrich metadata to meet the changing needs of their target community whereas the research data lifecycles tends assume a fairly 'static' data object (barring preservation/admin metadata etc) from the point of ingest into an Archive until the next Access/User/Re-use cycle.

Section 3 of this document is structured to represent the full lifecycle based closely on familiar lifecycle models but extended to support the more dynamic issues surrounding metadata management.

2.2. Types of Metadata

In 1997 the IFLA Study Group on the Functional Requirements for Bibliographic Records published a final report. As part of that effort, the group identified four generic user tasks to be accomplished using bibliographic records:

> "1. To *find* entities which correspond to the user's stated search criteria (i.e., to locate either a single entity or a set of entities in a file or database as the

result of a search using an attribute or relationship of the entity)

2. To *identify* an entity (i.e., to confirm that the entity described corresponds to the entity sought, or to distinguish between two or more entities with similar characteristics)

3. To *select* an entity that is appropriate to the user's needs (i.e., to choose an entity that meets the user's requirements with respect to content, physical format, etc., or to reject an entity as being inappropriate to the user's needs)

4. To *acquire* or obtain access to the entity described (i.e., to acquire an entity through purchase, loan, etc., or to access an entity electronically through an online connection to a remote computer)"

These are the key functions of metadata about a digital object from the end user perspective.

From the curators' perspective metadata serves different functions depending on the lifecycle phase, and the actors that create or use it. The terms associated with metadata are also used in varying ways, they often carry different meanings depending on the context in which they are being used, and the functions of some types of metadata may blend in with the functions of others. The tables "Functions and Schemas for Different Types of Metadata" in Appendix B illustrate some of the relationships between these types of metadata:

- Descriptive metadata
- Contextual metadata
- Technical metadata
- Preservation metadata
- Administrative metadata
- Structural metadata

This list is not meant to be a definitive list of terms, but it provides a standard vocabulary that can be used for the purposes of this report. Additionally, a glossary of common metadata terms can be found in Appendix A.

2.3. Metadata Quality

In 2004 Bruce and Hillmann published a paper about metadata quality in which they developed a domain-and method-independent model of quality indicators (Bruce & Hillmann, 2004). The literature has continued to grow, but most papers still reference Bruce and Hillmann framework. See e.g. Shreeves et.al (2005), Park (2009), Palavitsinis (2014). Their approach is applied in the evaluation of the metadata strategies of CLARIN, DARIAH and CESSDA, see sections 5,6 and 7.

Bruce and Hillmann define seven general characteristics of metadata quality. These characteristics are necessarily abstract, so that they can be applied domain independently.

Completeness: The metadata elements should describe the objects as completely as economically feasible. Moreover the elements should be applied, as much as possible, to the whole collection.

Accuracy: The information provided in the values of the elements needs to be correct and factual.

Provenance: Knowledge about the creation (how and by whom) and following transformations is necessary to make a judgement about the quality.

Conformance to expectations: Element sets and application profiles should in general contain those elements that the community would reasonably expect to find. Controlled vocabularies should be chosen with the needs of the intended audience in mind and explicitly exposed to downstream users.

Logical consistency and coherence: A need to ensure that elements are conceived in a way that is consistent with standard definitions and concepts used in the subject or related domains and presented to the user in consistent ways.

Standard mechanism like application profiles and common crosswalks enhance the ability of downstream users to assess the intended level of coherence.

Timeliness: This characteristic comprises two aspects; currency and lag. Both refer to the issue that metadata needs to be in synchronization with its target object. The end user should be able to judge that the description is not out-dated or the described object not yet available. "The aging of metadata presents obvious problems in the form of potentially broken URI's, drifting controlled vocabularies, and evolving, sometimes divergent, conceptual maps of the underlying corpus".

Accessibility: Metadata that cannot be read or understood by users has no value. In particular with diverse audiences for heterogonous federated collections, it is important to consider carefully the potential differences when designing and documenting metadata implementations.

Furthermore they define a system of tiered quality indicators. The first tier comprises three indicators that can be validated automatically:

- The ability to validate against a schema
- The use of appropriate namespace declarations
- The presence of an administrative wrapper

At the second level the following aspects improve the quality of the metadata, which also can be confirmed by automated means:

- The presence of controlled vocabularies
- The definition of elements by a designated community, with a publicly available application profile
- Provenance information at a more detailed level.

The third level of quality indicators is less likely to be determined automatically; it comprises information on conformance, trust and full provenance information.

Bruce and Hillmann applied the system of tiered quality indicators to the seven metadata aspects, creating a table which supports metadata creators or aggregators who might look for weaknesses in generated metadata including legacy and multiple-source data.

3. Metadata lifecycle

Metadata quality depends on an understanding of the business processes undertaken which in turn depends on an agreed representation of the metadata lifecycle and the Actors and Software agents (See section 3.2) and the roles they play. For the purposes of this deliverable and the lifecycle defined in this section we have defined a categorisation of actors. The definition of this categorisation is given in the glossary (Appendix A). Those working within data curation, from the point the creation or collection of data is conceived, to the point that it is used or re-used often employ some form of abstracted simplification of their high level activities, a 'lifecycle model' to support communication and planning.

The lifecycle structure used here cannot hope to provide a definitive description of lifecycle stages, of metadata types or of metadata quality criteria, which represents all of the metadata curation environments in scope. Instead the lifecycle is presented as a baseline which, alongside an understanding of metadata quality evaluation and metadata types, can be used to design and benchmark a local approach to describing, delivering and improving quality metadata.

The classic reference point for archivists is the OAIS model but this focusses almost exclusively on the repository phase of the lifecycle.

The OAIS remains a critical reference point but this section tries to reduce the somewhat artificial separation of the 'Archival Phase' of the lifecycle from wider data curation activities. The data itself remains the primary focus of the data producer/researcher/user but we incorporate the critical supporting metadata/documentation issues in a way, which more explicitly demonstrates their dynamic role in the lifecycle.

A number of alternate models, more focussed on the full lifecycle of digital objects were evaluated during the development of use cases for this deliverable (see appendices). These all fulfil some sort of communications purposes but from the perspective of this deliverable there remained a clear focus on the data which is the subject of curation and preservation with less focus on the numerous types of metadata necessary to support the management and movement of those data.

Data are the subject of research and they sit at the centre of most sequential/circular descriptions of the 'data' lifecycle. The data-centric focus is on the integrity and fixity of that data and while it may be validated and enriched in some ways the data points themselves are perceived as protected. For those engaging in curation (including those collecting, creating and using data) the metadata remains more dynamic than the original data.

Metadata design, redesign and implementation continue to be 'live' issues for

an Archive or other curators and access providers even when the data themselves remains unchanged. Those managing metadata, or using metadata to manage, continue to update to new standards and re-enrich metadata to meet the changing needs of their user communities. This contrast somewhat with the research data lifecycles' tendency to assume a fairly 'static' data object (barring preservation/admin metadata etc) from the time of ingest into an Archive to the next Access/User/Re-use event.

This section is based on an evaluation and mapping of the models identified as part of use case work, they also reflect ongoing work at the UK Data Archive to more coherently define the repository role within the wider digital curation lifecycle.

The lifecycle which provides the structure for the following subsection aligns with the OAIS model, but places it in a wider context by considering whether some derivation from existing lifecycle models would help support:

- Communicating issues of metadata quality to all actors in the data/metadata lifecycle
- Managing the stages of the lifecycle from a Curation/Preservation Planning perspective

It is not desirable to deliver yet another model to the curation and preservation community, instead existing models are evaluated and structured in a way that extends the focus to metadata issues. Actions are described in terms of data/metadata where both apply.

A number of activities undertaken during curation involve continuous interaction with and monitoring of the stakeholder ecosystem (see Community Watch & Participation) and planning (see Data Curation/Preservation Planning) in response to that monitoring. Other activities recur repeatedly through the lifecycle of digital objects (see Recurrent Actions and Events) and can usefully be designed once and applied several times. . The third component is the broadly sequential actions that are commonly understood to be the journey of data and metadata.

To improve data and metadata quality a common understanding of the lifecycle stages is a necessary precursor to understanding their component activities and the quality criteria of those activities. Further, we need to understand the actors (and sometimes software agents) playing a role in the lifecycle to effectively manage communication between them.

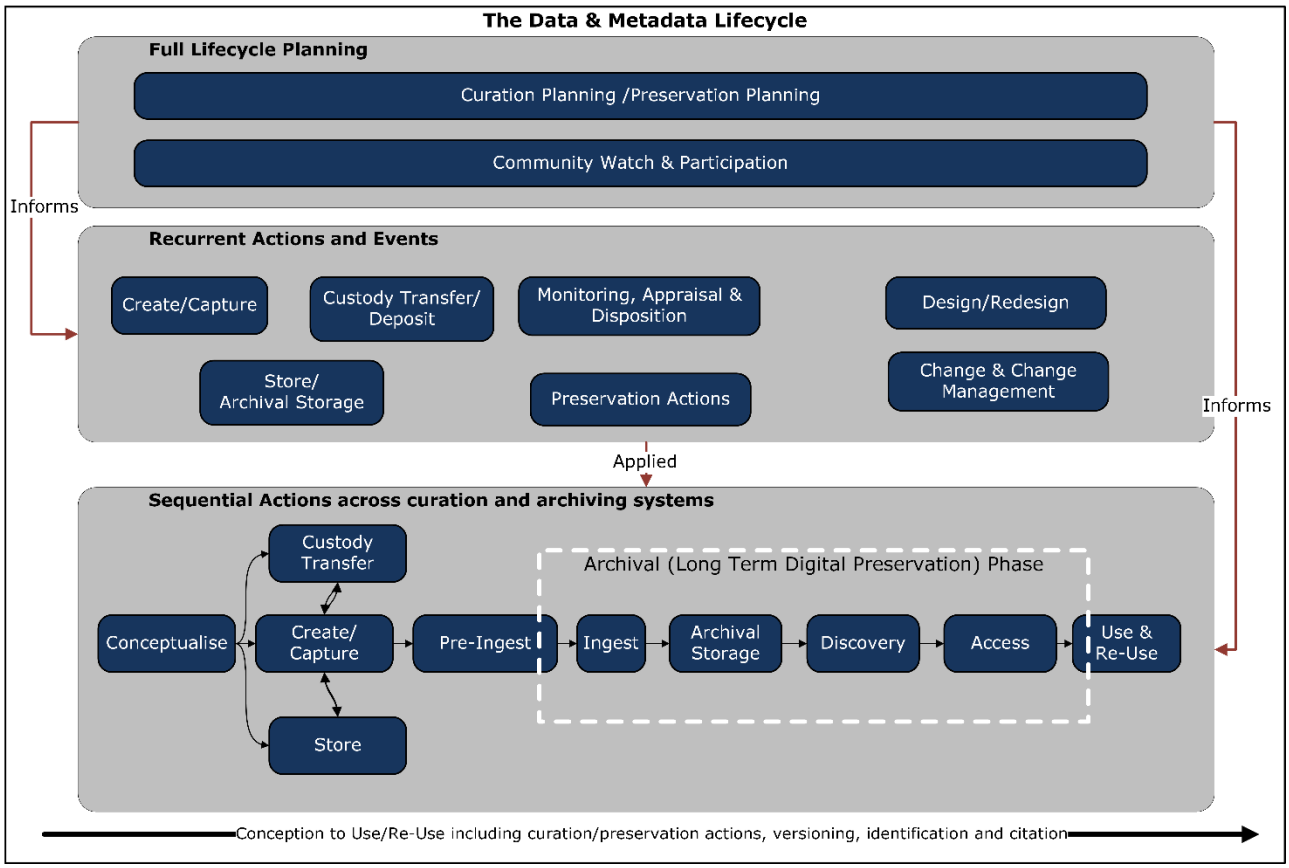


Figure 2: Data/Metadata Lifecycle to support metadata quality

3.1. Lifecycles Referenced

The lifecycle-related models below were referenced. Key concepts mapped from the lifecycles examined to the subsections below are cross-referenced preceded with "Map:". Mappings are indicative of alignment to support communications of the concepts rather than detailed comparisons of component activities. In some cases (e.g. Migration) mappings reflect the extension of the focus from purely 'Data' related concepts to also include metadata related concepts.

Mappings to OAIS and Lifecycles referenced are bordered in blue. Definitions from the DCC model are bordered in orange. Lifecycle diagrams are contained within Appendix C.

C1: OAIS Model

The reference model for designing an organisation of systems and people which will take responsibility for the preservation of information and providing access to that information for a designated community. This details the repository phase of the lifecycle by describing the activities within several 'functional entities': ingest, data management, archival storage, access, administration and preservation planning.

C2: DCC Curation Lifecycle Model

A familiar and well respected model and a well-recognised diagram for

examining the full curation lifecycle.

B3: DDI-L: Combined Lifecycle Model
The model applied to the most recent versions of the Data Documentation Initiative commonly used by social science archives. Previously 'study' or data collection focussed the standard/schema now addresses the full data/metadata lifecycle.

B4: Generic Longitudinal Business Process Model (GLBPM)
Influenced by the DDI-L lifecycle the GLBPM provides more granular sub-activities. Though a linear lifecycle approach is relevant and has its place the supporting documentation[1] makes it clear that we need to take into account that the various actions are repeatedly taken during any real longitudinal process. Longitudinal is of interest because it has a higher number of iterations and revisions of metadata (and data) including both planned and unplanned revisions.

B5: The Research Lifecycle: Traditional Model (DWB)
The traditional model presented as a reference point by the metadata quality work for T5.3 under the DASISH project.

B6: Steps in the Research Life Cycle (DMConsult)
A more project/publication focussed view of the data/metadata lifecycle.

B7: Authenticity Protocol Information from APARSEN WP24
A subset of the activities proposed as priorities for metadata collection under WP24 of the APARSEN project. Highlights the issue of data/metadata management in non-Archival 'Keeping Systems', which may not have a Long Term Digital Preservation (LTDP), remit but remain critical stages in the lifecycle. Within this document these 'keeping systems' are referred to as Curation Systems, see below.

3.2. Actors and Communications across the Lifecycle

The individuals and organisations with an influence on, or a more direct role in, the lifecycle are often referred to in broad terms as the 'stakeholders'; within the lifecycle and, more explicitly when undertaking specific business processes the term 'actors' is used.

The task of creating and managing quality metadata cannot be left up to one person or organization in the research lifecycle. Actors involved in all levels and disciplines of data management need to communicate and work together to create quality, compatible metadata, while avoiding duplication and inconsistency (RIN, 2008). The relevant stakeholders must be identified and managed and clear lines of communications developed.

[1] http://www.ddialliance.org/system/files/GenericLongitudinalBusinessProcessModel.pdf

The key word here is *communication*. A repository can invest time and money into double-checking the accuracy of metadata (Atlas of Living Australia [ATLAS], 2011) but this approach cannot reach its full effectiveness and efficiency without clear lines of communication throughout the data curation process. All parties involved must understand their responsibilities and have the appropriate knowledge and training to do their part in creating quality metadata (RIN, 2008).

When referring to the lifecycle below it must be understood that though simplifications can be made to support communications the portrayal of the lifecycle as a rigid sequence or a continuous circle must be understood to be representative. Some steps may be repeated several times or in a different order from that portrayed. Relevant actors involved at each lifecycle stage must be identified as part of the stakeholder analysis process.

Until the point of ingest into an archive the communications between the researcher, the repository, and the funder are crucial. For archives the need to correct or create metadata, which could have been collected earlier in the lifecycle, is costly in both time and human resources as well as being more error prone and potentially less rich. The archive would ideally communicate with the researcher about creating high quality metadata *before* the research process begins. It is important to remember that most researchers are *not* information specialists and may not have much experience in creating metadata (Campbell, 2007; Hillman et al., 2004). Repositories could benefit from providing education and support to researchers in order to simplify the metadata creation process for them. While this type of outreach may initially present extra costs for the repository, it could eventually save time and money by eliminating the duplication and double-checking of work that occurs when a repository receives insufficient metadata (RIN, 2008).

The funder's role is to make its expectations clear in terms of metadata and data sharing. It should also provide sufficient funds to the researcher to be allocated for the time and resources it takes to implement a good data management plan (RIN, 2008). Some funders, such as the National Science Foundation and the Economic and Social Research Council, require a data management plan as part of the research proposal, which conveys the funders' expectations for quality metadata and data management to researchers (Economic and Social Research Council, 2013; National Science Foundation [NSF], 2012).

Although automatically captured metadata has the potential to be more consistent than human created metadata (though sometimes less rich), it may present a barrier to interoperability if there is a lack of communication between the researcher and the repository. This can result in wasted time and money. For example, Mize and Robertson introduce a scenario in which a research organization has to take the time to adapt its carefully documented metadata to a different schema in order to comply with the standards of a data-clearinghouse (2009, Introduction, para. 5). This process could have been eliminated had the organization and data-clearinghouse established clear lines

of communication before the organization chose their incompatible metadata schema. Another issue with automatically captured metadata is a lack of interoperability standards between the different types of software used for capturing metadata because there is little coordination between the software manufacturers (UKOLN, 2007).

Essentially, the movement of information around the research lifecycle does not only represent a transfer of data, but also a transfer of *knowledge:* that is, "…a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information," (Liyanage, Elhag, Ballal & Li, 2009, p. 119). Keeping this in mind, it is crucial that the repository and researcher establish clear lines of communication so that both are aware of the expectations for complete and quality metadata, and to ensure that research results can change hands without any loss of information.

It is imperative to recognize that all communication that occurs during the research lifecycle is a two-way interaction. Each actor has not only something to learn, but also information to contribute (Eppler, 2008; Liyanage, et al., 2009). For example, a repository may guide researchers in how to correctly and completely document their descriptive and contextual metadata during the research process but on subject-specific issues it is unlikely the repository will have the same level expertise. . Both actors would benefit from a collaborative approach. Eppler sums this up, stating, "The process of knowledge communication hence requires more reciprocal interaction between decision makers and experts because both sides only have a fragmented understanding of an issue and consequently can only gain a complete comprehension by iteratively aligning their mental models," (2008, p. 326).

Of course it is unrealistic to expect a repository to personally interact with every single individual researcher who contributes data, but it would be advantageous for a repository to provide a forum where contributors could ask questions or give feedback. The UK Data Service is one example that offers this kind of support. In addition to providing easy to use forms for recording metadata, they have a complete list of commonly asked questions and a query service for data contributors on their website (UK Data Service, 2014).

Communication is crucial in any transaction, and the knowledge transfers that take place in scholarly research and data management are no exception. This is true whether you are creating a data management plan, creating metadata, or managing someone else's data. Humphrey explains the importance of all actors in the research lifecycle when he says, "In the same way that it takes the proverbial village to raise a child, it takes the commitment of the knowledge community to preserve and provide access to research data," (2006, p. 2).

3.3. Full Lifecycle Planning

Activities which take place throughout the lifecycle. Lifecycle Planning informs both recurrent and sequential actions.

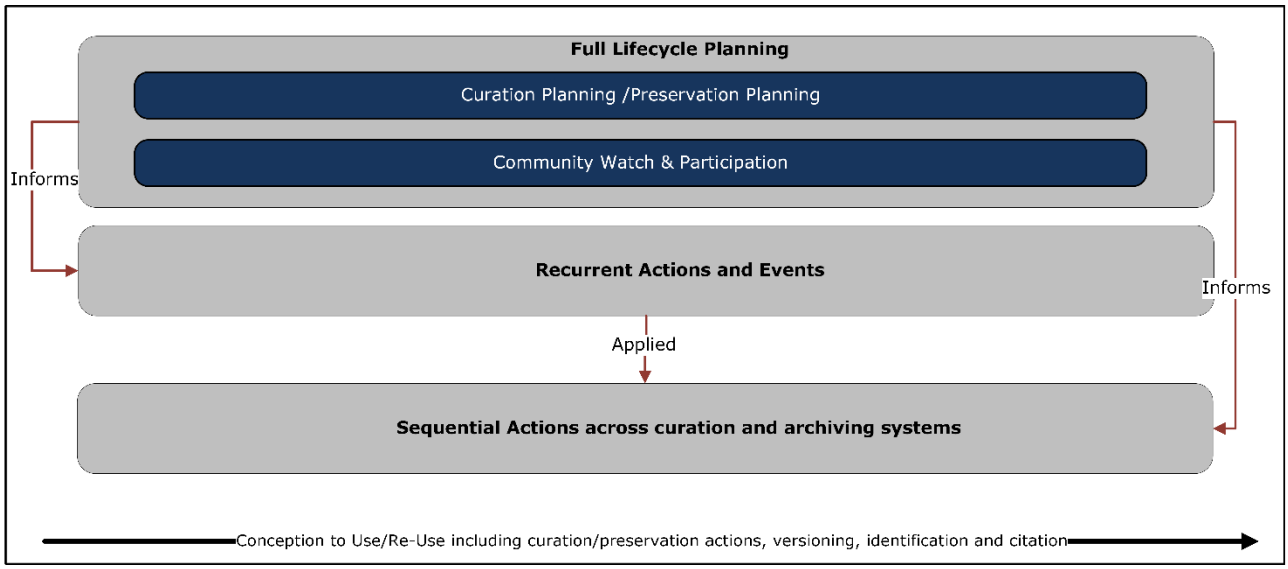


Figure 3: Full lifecycle planning

To deliver planning across the lifecycle the outcomes of community watch and participation must be integration into curation/preservation planning processes.

It is clear that data/metadata may progress through numerous systems during the lifecycle so Lifecycle Planning will seldom, if ever, be a unitary role. It is presented as a single grouping here on the assumption that it is understanding and communication between the actors and systems in play, which delivers the best quality data and metadata to researchers.

Map: DCC Full Lifecycle Actions

Map: Collaborate & Communicate (DwB)

Community Watch and Participation

If your goals are to serve and communicate with a community (for Archives a 'designated community' in OAIS terms) then the starting point is to engage with and understand that community.

Data/metadata related to the identification and management of stakeholders and records related to the outcome of those interactions must be managed.

The outcomes of Community Watch and Participation guide decisions taken during Curation/Preservation Planning including the design of recurrent and sequential actions.

"Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software." (DCC)

Map: Monitor Designated Community (OAIS)

Map: Community Watch and Participation (DCC)

Curation/Preservation Planning

Curation/Preservation Planning incorporates knowledge from Community Watch and Participation to guide a standardised approach to Recurrent and Sequential actions in the Lifecycle.

Archives may have a more explicit LTDP remit and there may be variations in practice but advice on best practice activities, which ensure the quality of data/metadata, don't vary greatly through the digital object lifecycle. Good planning, communications and practice earlier in the lifecycle reduce cost and complexity and increase quality later in the lifecycle.

This activity develops plans for curation and preservation practices and actions, including their management and administration throughout the curation lifecycle of digital material. This includes tools for the capture and management of relevant data/metadata.

At the curation/preservation planning level the target digital object for curation must be defined whether this consists of a digitised book or an extensive collection of complex interrelated research files from a longitudinal study. This core object data and associated metadata is further supported by metadata as records, database fields and documentation, which support the digital object management process. Some of this peripheral metadata may not be retained alongside the object in an archival information package (AIP) but it remains critical to the management of the object through the lifecycle.

The implementation of these curation/preservation plans is contained within the Recurrent and Sequential lifecycle headings below. These include activities (defined by the DCC as 'Full Lifecycle') to assign administrative, descriptive, technical, and structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term and to collect and assign representation information required to understand and render both the digital material and the associated metadata. These activities include the implementation of appropriate versioning, identification and citation practices as defined by the Curation/Preservation Planning activities.

Map: Preservation Planning (OAIS)

Map: Preservation Planning (DCC)

Note: Preservation Planning also defines the Preservation Actions to be taken. Even though formal 'preservation actions' are LTDP actions all curation

systems taken managed actions to maintain data and metadata. For this reason Preservation Actions are included under Recurrent Actions and Events.

3.4. Recurrent Actions and Events

The design and implementation of recurrent actions and events is informed by Lifecycle planning. These actions are designed to ensure consistency and efficiency and are then applied at various points within the sequential actions.

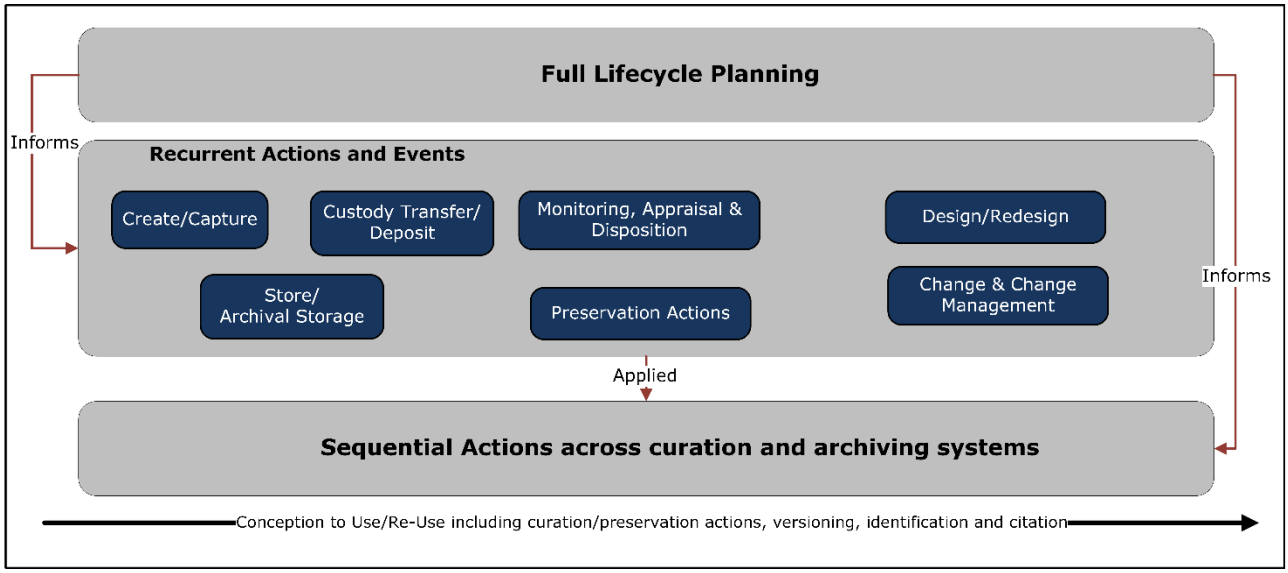


Figure 4: Recurrent actions and events

A number of data/metadata related activities occur numerous times during the lifecycle of a digital object. These activities and their application are defined by Curation/Preservation Planning, often influenced by Community Watch and Participation. Recurrent actions may not be applied identically at each of the relevant stages of the Sequential lifecycle but considering them as a whole supports improved full-lifecycle planning and communications.

Curation/Preservation Planning can never define the exact implementation of actions for these recurrent actions but good practice and standardised approaches can be defined.

Map: Occasional Actions (DCC)

Monitoring, Appraisal and Disposition

Monitoring, appraisal and disposition will be terms familiar to those dealing with records management issues. Records Management is increasingly being aligned with curation and preservation practices.

Data/metadata are monitored throughout the lifecycle to support their appraisal at key points. The outcome of an appraisal defines the 'disposition' action to be taken: to retain, reject, transfer or destroy data/metadata.

The outcome of an appraisal decision on a monitored metadata record could indicate that the retention period for personal data held by the repository had expired and that the metadata should be deleted.

The outcome of an appraisal decision on a monitored file format could indicate that the format has reached an unacceptable risk threshold and that a format migration to a lower risk format should be undertaken.

- **Monitoring** requires clear data/metadata identification and management
- **Appraisal** requires a decision process with clear criteria
- **Disposition**:
    - **Retention** may be applied in line with a retention schedule to define the period of retention (for legal reasons) or the period until the next appraisal.
    - **Rejection** (e.g. at the point data/metadata are 'offered' to an archive or other keeping system,) implies that there is no custody transfer, or, if a custody transfer has taken place that destruction is necessary.
    - **Destruction** requires a standard process for the destruction of data/metadata whether as a 'tidying' process or for more formal information security reasons.
    - A **Transfer** decision requires a standard process with clear criteria, implemented through a standard custody transfer (see Custody Transfer/Deposit).

Pre-Archive decisions may be more ad hoc as to whether data/metadata should continue to be maintained to the end of a process or project. During the Archival phase of the lifecycle an Appraisal is likely to trigger a rejection of an offer of data/metadata (often based on the remit of the Archive), for Archives it is more likely that metadata will be changed and versioned and older versions retained but there are cases where for space, policy or security reasons the complete 'Deaccession' of data/metadata is necessary.

Community Watch and Participation information may be drawn into the monitoring process via curation/preservation planning decisions to help identify changes in user requirements. These may be driven by practical research requirements such as increased descriptive metadata to support interdisciplinary research or it may involve the sharing of more administrative and preservation metadata with dissemination information packages (DIP) as users demand greater evidence that the curation process can be trusted.

Monitoring, appraisal and disposition may be impacted by the time since the data and metadata were last updated. User expectation will change and descriptive metadata may need to be updated to reflect current best practice even if the underlying data is static. Data which present a disclosure risk at their time of creation may need to have rights and access metadata amended over time.

Structural metadata may need to be updated to support more advanced user

interfaces and, along with preservation and administrative metadata, is likely to be updated as technologies to support preservation change.

Map: Reappraise (DCC)

"Return data which fails validation procedures for further appraisal and reselection." (DCC)

Map: Dispose (DCC)

"Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction." (DCC)

Design/Redesign

There are several points in the lifecycle where different actors with custody over the metadata may have the power to design/redesign metadata for their own purposes or to meet the needs of others.

Ideally metadata remains as fixed as possible after initial design throughout the digital object lifecycle but redesign at some level is inevitable. Metadata entered should be based on a metadata design or 'profile' of acceptable and/or required metadata elements applicable in the local environment.

Metadata Profile

The full list of metadata elements supported by an organisation or system may be characterised as a 'Metadata Profile'. A Design/Redesign implies a change to the metadata profile, which must be subject to some level of change management to plan the process and identify the impact. An updated metadata profile may provoke a metadata migration.

A managed metadata design/redesign process should take place (i.e. a new version of the metadata profile should be created) before changes are made to digital objects or their metadata.

Many metadata standards, including METS and the DDI use the concept of a metadata profile to define the subset of elements applied in a particular scenario by a particular organisation or system.

A researcher or data producer may revise their metadata design during the create/capture process, or to better handle storage and custody transfer issues. Archives will change their Deposit metadata at the 'Producer/Archive Interface' and may change current and past metadata as ingest practices are changed. Design/Redesign approaches are set by Curation/Preservation planning and may be intended to meet internal goals (new best practice, information security or trusted digital repository issues) or to meet the needs

of the community at the Discovery, Access and Use/Re-Use stages. Archival Storage metadata design may be more complex or rigorous than at earlier stages in the lifecycle but it is in the interests of all to align approaches to storage and storage metadata.

The selection of metadata schema/elements and controlled vocabularies includes the descriptive, contextual, technical, preservation, administrative and structural metadata and each design must be evaluated in terms of quality requirements. Criteria for quality assurance must be designed and applied during any change management process.

Completeness must be ensured by the selection of appropriate schema/elements based on the needs of the repository (for technical/administrative/preservation purposes) and the needs identified during community watch and participation. Standards for accuracy must be set and either automatic

Metadata design/redesign must include an understanding of the relationship between Metadata, Documentation and Data at each stage so depends on the provision of object model(s) by curation/preservation planning. File naming, version procedures and structure of Documentation and Data may be key metadata.

- Define the purpose
- Understand the lifecycle
- Evaluate existing standards
- Evaluate existing best practice
- Define processes including capture, validation, quality assurance
- Identify infrastructure requirements
- Understand likely storage environments
- Understand likely delivery environments (resource discovery systems, visualisation systems)
- Identify costs of metadata creation
- Identify rights issues (who owns the rights to metadata once deposited in an Archive)
- Identify security issues (including who has the rights to edit metadata)
- Identify privacy/disclosure issues.
- Specify metadata
- Plan timetable

And in the case of Redesign

- Identify impact of change

The design/redesign of metadata implies the need for Change Management.

Map: Design/Redesign (GLBPM 2)
Map: Build/Rebuild (GLBPM 3)

Change and Change Management
Metadata around digital objects or processes being managed may change as a result of undertaking standard actions or as a result of the design/redesign of metadata.

Versions
Both the metadata profile and the metadata applied to a particular digital object should be subject to standard versioning procedures; ideally previous versions of the metadata or metadata profile are retained.

After the initial capture of metadata the notional data/metadata object exists and further metadata changes can be characterised as migration, integration, deletion or aggregation/extraction.

Metadata Migration
Updating metadata to apply a revised metadata profile e.g to align with a new metadata schema version, a new metadata schema entirely or a new or revised controlled vocabulary.

Map: Migrate (DCC)

"Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence." (DCC)

Integration
New data/metadata is added

Deletion
Data/metadata is deleted as a result of an appraisal or of reaching the end of the agreed retention period

Aggregation/Extraction
Different data/metadata objects are merged (aggregated) or extracted (separated)

The collection of metadata related to change management of data/metadata is critical to a complete provenance record for digital objects. Most commonly only changes which impact the use/re-use of resources is shared with end users but increased awareness of the Trusted Digital Repository (TDR) agenda may drive repositories to share a larger proportion of such technical/administrative/preservation metadata.

Map: Build/Rebuild (GLBPM 3)

Map: Transform (DCC)

"Create new data from the original, for example

- By migration into a different format.

- By creating a subset, by selection or query, to create newly derived results, perhaps for publication" (DCC)

The 'Transform' action defined by the DCC may be a Preservation Action (migration to an improved data/metadata format triggered by a Design/Redesign) or may occur during Use/Re-Use.

Custody Transfer/Deposit

Transfer

The custody transfer of data/metadata between two Curation systems

Deposit

The custody transfer from a Curation system to an Archive system. This does not include the full Ingest process of normalisation and enrichment in preparation for archival storage and access, rather this equates to a Pre-Ingest process which validates that the digital object is acceptable and deals with the administrative matters around validation and transfer, as covered by the 'Producer-Archive' interface from the PAIMAS.[2]

All custody transfers are a point of risk for data and metadata through incomplete copies, copy errors or unclear transfers of responsibility. LTDP Archives may have specific appraisal and selection criteria to be checked and additional custody transfer validation processes but many of these good practices could usefully be applied in the earlier stage of the data/metadata lifecycle.

Contextual data is particularly at risk during a custody transfer (Humphrey, 2006). An example of this is a researcher withholding the sources of questions asked in an interview. The researcher may not have formally documented this information, or may not think it will be useful to future researchers, but no one should make assumptions about how other people will use data (Nelson, 2009). Additionally, one study found that people are more likely to share information about what they know best (Beckett & Hyland, 2011). Placing this in the context of research data, if researchers limit the research materials they share to those relevant to their area of expertise and omit elements of contextual metadata, which may be of interest to other researchers, they are reducing the likelihood of their data being re-used, especially in interdisciplinary research.

Formal PAIMAS-like transfer 'projects' can be designed to ensure completeness in terms of machine-actionable metadata and some degree of accuracy if those managing the process have sufficient expertise. Provenance metadata from the previous custodian should be captured wherever possible.

[2] http://public.ccsds.org/publications/archive/651x0m1.pdf

Map: Receive (DCC)

"Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata" (DCC)

The DCC description here is from an Archive perspective but we need to take account that this is a subset of the 'Custody Transfers' of data in its lifecycle.

Store/Archival Storage
The technical and administrative mechanism for storing and maintaining the data/metadata. Requirements for Archival Storage may be more rigorous but it is in all parties' interest to adopt good storage practices at all stages of the digital object lifecycle.

Technical, Preservation, administrative and sometimes structural metadata form the basis of storage metadata with the need to validate the identify and integrity of multiple copies in multiple locations before and after each copy is taken and in the case that a copy is compromised in some way and must be replaced.

Map: Store and Archive (DwB)

Map: Archival Storage (OAIS)

Map: Data Archiving (DDI-L)

Map: Archive/Preserve/Curate (GLBPM 6)

Map: Store and Archive (DwB)

Map: Data Archive (DMConsult)

Preservation Action
Even though preservation actions are, technically, just another class of managed change (based on design/redesign by preservation planning) they are included here separately as a high risk point for data and a critical area of metadata creation and management. Actions undertaken to ensure continued access to data may include changes like file format migration which are high risk points in the lifecycle, Metadata to record, justify and validate changes made are critical.

The vast majority of actions to ensure access to data and metadata in the long term impact systems throughout the digital object lifecycle, not just within an archive. Curation/Preservation Planning should specify the application of appropriate monitoring and actions throughout the relevant phases of the lifecycle.

As a simple example, appropriate standard or format selection for LTDP needs at the Conception phase will always be an improvement over the application of

such standards and formats at the point of Ingest. Similarly the need to migrate to new standards and formats may need to be addressed in curation systems with no formal LTDP mission.

Map: Preservation Action (DCC)

"Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats." (DCC)

Map: Transform (DCC)

"Create new data from the original, for example

- By migration into a different format.

- By creating a subset, by selection or query, to create newly derived results, perhaps for publication" (DCC)

The 'Transform' action defined by the DCC may be a Preservation Action (migration to an improved data/metadata format triggered by a Design/Redesign) or may occur during Use/Re-Use.

3.5. Sequential Actions

A key reason for presenting a 'sequential' actions section is that despite the fact that 'circular' approaches display more of the innate complexities of the process the 'birth to re-use' sequence is commonly understood and supports simple communication, especially when drilling down into the design of more detailed business processes.

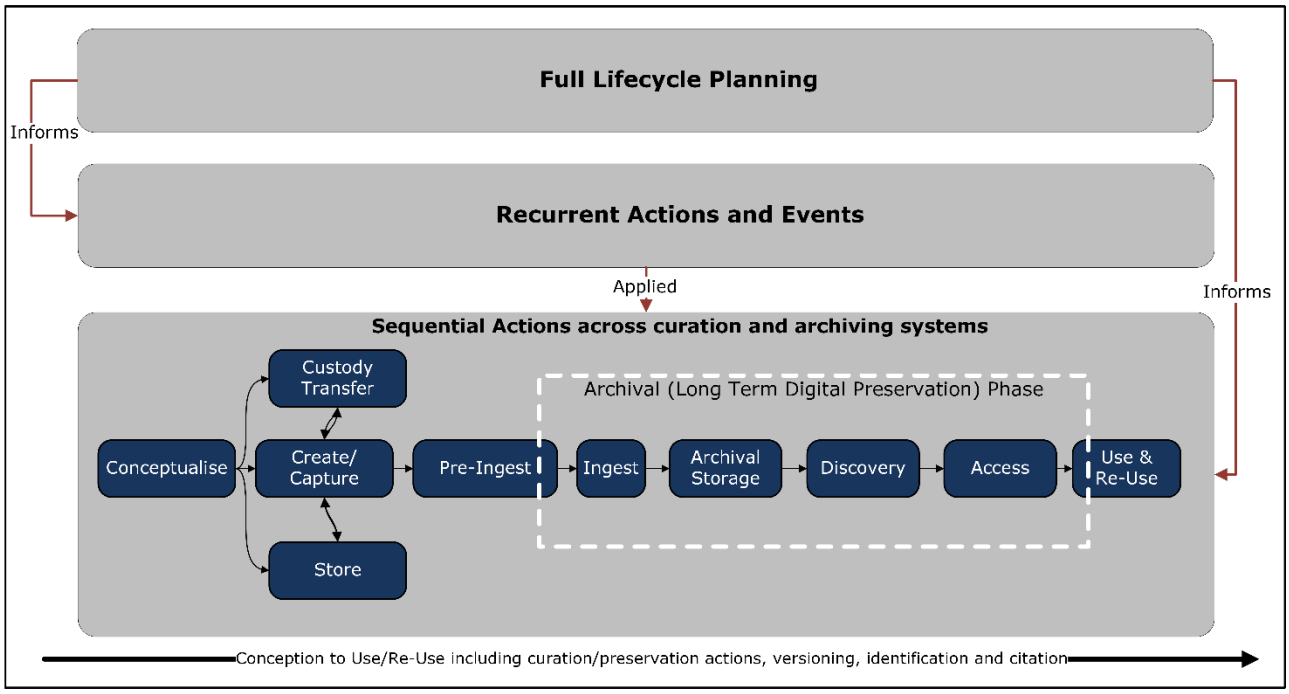


Figure 5: Sequential Actions

Map: Sequential Actions (DCC)

Conceptualise

During the conceptual phase of the research data lifecycle the researcher is the primary actor. In the circular nature of data creation and management researchers are often dependent on the quality of descriptive metadata within a repository for the search and discovery process (NISO, 2004) (see **Discovery**, the relevant data are accessed (see **Access**) and evaluated. Data may be re-used (see Use and **Re-Use**) or the researcher may conclude that new data is required to answer the research question. In either case, it is often the researcher's responsibility to create contextual and descriptive metadata to accompany the new data created during research (Hillman, Dushay, & Phipps, 2004).

Map: Conceptualise (DCC)

Map: Study Concept (DDI-L)

Map: Evaluate/Specify Needs (GLBPM 1)

Map: Proposal Planning Writing (DM Consult)

Map: Develop Proposal (DwB)

"Conceive and plan the creation of data, including capture method and storage options" (DCC)

Should be guided by Curation/Preservation Planning and by standards agreed for Recurrent Actions and Events.

The DCC description helps to clarify that defining capture methods which, may include both human factors and tools are in scope for conceptualisation. Storage options imply, not only physical storage but also security of storage and the basic, but vital, concepts of structural, title/naming and identifier metadata, Asnoted above these challenges appear at many stages in the lifecycle and may be better designed separately and then applied throughout as standard.

Metadata at this stage is ideally designed with the full future lifecycle of the data/metadata in mind, but of course that future journey of the metadata is often unknown. Without an understanding of the future data/metadata lifecycle the metadata design will often focus on that required to support the administration of the data to the point that it has served its purpose (analysis and publication in the case of data which is the product of research funding). Even with an understanding of the later stages of the lifecycle of the data/metadata there is a need to incentivise the provision of metadata to support archiving, resource discovery and secondary use, whether financial (see Funding below) or by promoting the benefits to the data/metadata creator of re-use and citation.

For data/metadata to be generated as a product of research there is often a need to seek funding which may imply additional steps (see below) but for other data which are of relevance to research there may have been little or no consideration of the wider metadata needs or the lifecycle of the data beyond original conception.

Examples
Social Media data: metadata is designed to support the functions of the service and to report on behaviours, which will inform future service provision or monetisation of the service. Metadata to support later research may not be considered.

Administrative data: metadata is designed to support the business processes identified and any critical analysis and reporting at a higher level. Metadata to support later research may not be considered.

Curation/Preservation Planning should provide some guidance for Conceptualisation.

Conceptualisation should set standards for each of the actions within the curation system(s) involved. After creation/capture data/metadata may undergo multiple custody transfers to multiple storage systems before deposit in an archival system.

Funding: for data derived from funded research it is increasingly likely that there will be a Data Management Plan including plans for metadata, which are a requirement for funding. Data Management Plans ideally consider the full lifecycle of data, metadata through to long term digital preservation.

As noted above the metadata will simply be a by-product of the immediate need to support the collection/reporting processes so funding (and planning) is less likely to be addressed independently of planning around the data. The cost of metadata creation is usually part of the overall costs and not split out.

Curation/Preservation Planning should provide some guidance for Funding. Note that metadata creation has a cost implication throughout the lifecycle.

Map: GLBPM 1.6 Prepare proposal and get funding

Create/Capture
Initial collection of data/metadata. Conceptually initial creation/capture is a one off activity as subsequent actions are (in the strictest sense) managed changes which are 'integrated' into to the original data/metadata captured.

For longitudinal work multiple 'Created/Capture' events over time may be 'aggregated' into a single digital object.

Map: Data Collection (DDI-L)

Map: Data Processing (DDI-L)

Map: Collect (GLBPM 4)

Collect4Map: Gather Resources (DwB)

Map: Data Collection (DM Consult)

Map: Data Analysis (DM Consult)

Map: Create (DCC)

"Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation." (DCC)

Custody Transfer
See: **Custody Transfer/Deposit** in Recurrent Actions and Events

Transfers will differ based on the custody transfer protocols in place for each sending and receiving system up to the point of Deposit in an archive.

Pre-Ingest
The Pre-Ingest phase covers the lifecycle from the first point of contact with a potential depositor (who may be the Data Producer or another custodian), through negotiation to deposit. See the 'Producer-Archive' interface from the PAIMAS (http://public.ccsds.org/publications/archive/651x0m1.pdf).

Contact detail metadata may be the first collected, then sufficient metadata to support Appraisal and Selection. The Pre-Ingest process is effectively a negotiation with the depositor which includes descriptive and rights metadata as well as the appropriate licences at the deposit stage.

As the responsibility for data and metadata is transferred the administrative, technical, and preservation metadata become critical. The archive takes on a clear role as it manages and preserves the data for future use (UKOLN, 2007). However, while the archive is the primary creator of metadata from pre-ingest through to archival storage and access stages, it is still necessary to keep open lines of communication between the archive, funders, and researchers, to ensure that everyone has the same expectations about data access and preservation.

Appraisal and Selection for Collection

Appraisal and Section (DCC)

"Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements." (DCC)

Deposit
See: **Custody Transfer/Deposit** in Recurrent Actions and Events

The formal custody transfer protocols covering the move of data/metadata into an Archival system.

Note that from the point of Deposit custody transers between actors within the archival business processes may usefully apply standard Custody Transfer/Deposit protocols to support data/metadata integrity

See: **Custody Transfer/Deposit** in Recurrent Actions and Events

Ingest
The Ingest process is where the bulk of metadata for discovery and context is validated, enriched or created.

"Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements" (DCC)

Map: Ingest (OAIS)

Map: Ingest (DCC)

Map: Data Processing (DDI-L)

Map: Archive/Preserve/Curate (GLBPM 6)

Archival Storage
Extensive metadata is in place within the Archival Storage system included that required to manage multiple copy integrity.

Store/Archival Storage section

Map: Store (DCC)

"Store the data in a secure manner adhering to relevant standards." (DCC)

Map: Archival Storage (OAIS)
Map: Data Archiving (DDI-L)
Map: Archive/Preserve/Curate (GLBPM 6)
Map: Data Archive (DMConsult)
Map: Store and Archive (DwB)

Discovery
Metadata to support data discovery is a critical area of metadata development. This will include local archival discovery systems and opening metadata records for harvesting by wider data/metadata portals.

Map: Data Dissemination/Discovery (GLBPM 7)
Map: Search and Discovery (DwB)

Access
Critical metadata to support suitably granular access controls dependant on the requirements of the data producer and the sensitivity of the data/metadata. Access systems must integrate the access criteria associated with data/metadata, often interacting with metadata relating to those requesting access to data such as affiliation or intended use.

Map: Access, Use, Re-Use (DCC). See definition under
Access, Use and *Reuse*
Map: Data Distribution (DDI-L)
Map: Data Dissemination/Discovery (GLBPM 7)
Map: Publish and Disseminate (DwB)
Map: Data Sharing (DMConsult)

Use and Re-Use
Metadata including contextual information, version, identification and citation

provided at the point of Access are critical to effective use and re-use of data/metadata. The original researcher or the host repository may be able to demonstrate the impact of their work by monitoring citations during re-use.

Map: Access, Use and Re-Use (DCC)

"Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable." (DCC)

Map: Data Analysis (DDI-L)

Map: Research/Publish (GLBPM 8)

Retrospective Evaluation (GLBPM 9)

Map: Analyse and Experiment (DwB)

4. Research Infrastructure Model

This section provides a generic description of the properties of repositories (including archives) which form part of wider cooperative aggregation efforts (infrastructure in European research terms) where data and metadata may be shared. It further considers the evolving possibility of 'super-infrastructure,' which may aggregate content from one or more infrastructures and/or their component repositories.

For the purposes of considering metadata quality the infrastructures addressed by this deliverable can be simplified to identify the key agents, systems and information (data and metadata) flows.

The different metadata propagation and use strategies within CLARIN, DARIAH and CESSDA are respectively covered below under section 5, 6 and 7. For the sake of further simplifying the examples we will assume that only metadata and not data propagates between systems.

At the most basic level metadata about a digital object propagates from a repository system to an infrastructure system. The initial information transfer may be triggered by either party i.e. a push or a pull.

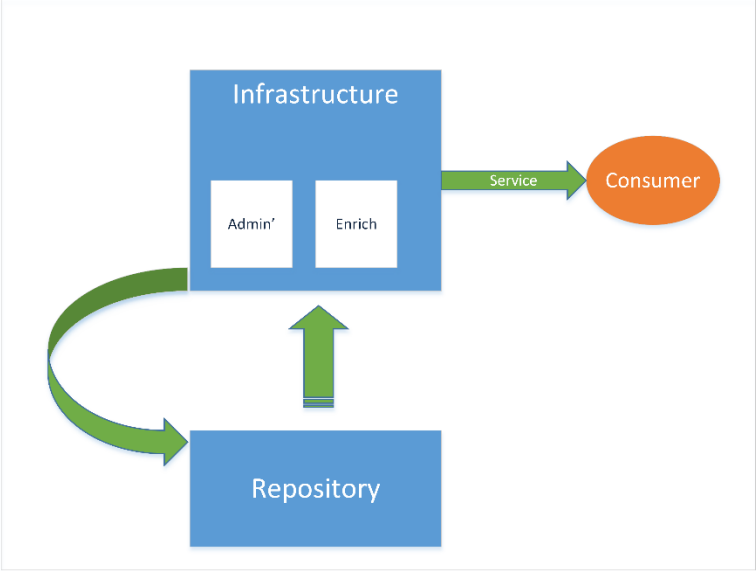


Figure 6: Interactions between sending and receiving system

The receiving system receives the metadata from multiple sources in order to offer services based on the aggregated resources.

Even if the digital object metadata is used without change at the infrastructural level there will be additional administrative metadata needed to support the services offered to consumers.

If the digital object metadata is enriched or changed in any way this creates a new version of the metadata. The sending and receiving systems must consider whether enrichments at the infrastructure level should be passed back

down to the repository level to ensure the metadata remains synchronised. A common form of enrichment at the infrastructure level would be the creation of defined relationships between aggregated resources whether objects, authors or publications.

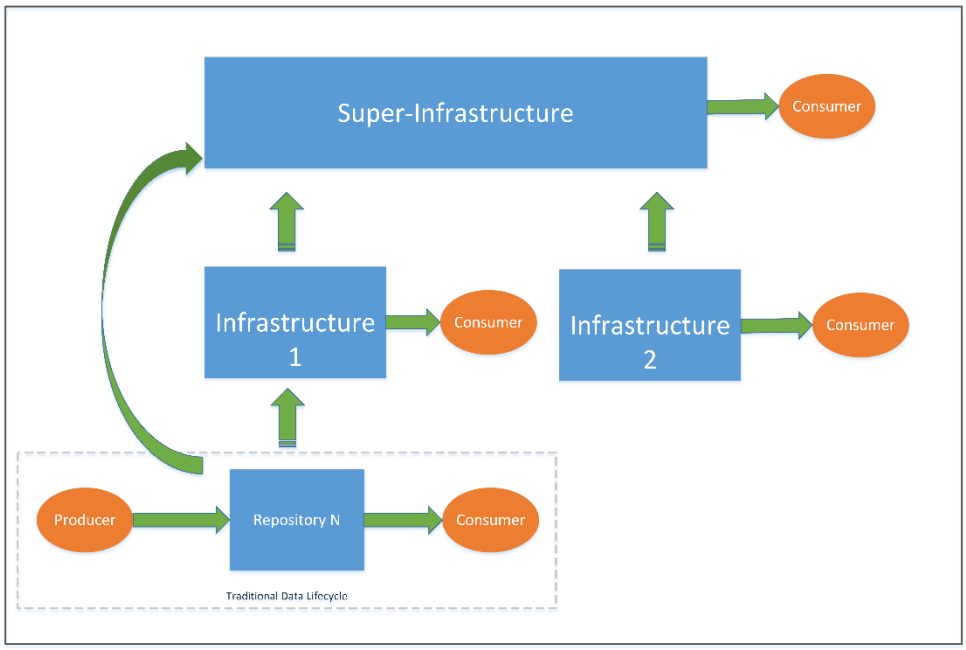


Figure 7: Super-Infrastructures

As the Infrastructures themselves off their own resources, either through a managed propagation or via harvesting through OAI-PMH or through making an API available an additional tier of change is encountered.

A super-infrastructure receiving metadata from multiple infrastructures may harvest from the original repository and/or via the intermediate infrastructures. As infrastructures will often specify a simplified metadata schema for interoperability harvesting from the original repository may offer richer metadata from closer to the source but super-infrastructures may also want to harvest the enrichments undertaken on the aggregated resources at the infrastructure level.

While these transfers of metadata are not formal 'custody transfers' each propagation event may involve some simplification of the source metadata and each enrichment event may not be synchronised back to the source.
In addition to the potential benefits these rich interactions between systems present a potential risk to metadata quality in terms of depth and accuracy.

5. Metadata Strategies of CLARIN

5.1. Organisation of CLARIN

CLARIN[3] is a distributed data infrastructure, with nine 'CLARIN Centres' in Europe. The centres are of different types, covering universities, research institutions, libraries and public archives. The CLARIN Centres all provide access to digital language data collections, and to different extents also to digital tools to work with those resources, and to the necessary expertise to support researchers working with them. The CLARIN Governance and Coordination body at the European level is CLARIN ERIC and its members are governments or intergovernmental organisations. The following eight countries are at this moment members of CLARIN ERIC: Austria, Bulgaria, Czech Republic, Germany, Denmark, Estonia, the Netherlands and Poland. The ninth member is the Dutch Language Union, which is an intergovernmental body created by the Dutch and the Flemish government, responsible for the maintenance and promotion of the Dutch language. These nine were the founding members. Norway, currently an observer, will also join CLARIN ERIC, and more countries are expected to join in the coming years. CLARIN ERIC Members ship is not limited to countries of the EU and Associated States.

CLARIN ERIC's main task is to build, operate, coordinate, and maintain the CLARIN infrastructure;it neither conducts nor funds research activities. CLARIN has a coordinating office which is aligns activities between the centres and is also financing some common infrastructure facilities. The majority of work isn undertaken out in the CLARIN Centres in the member countries and is financed solely by the member states.

5.2. CLARIN metadata strategies

Since the beginning of the Preparatory Phase Project (2008-2011) CLARIN has focussed on collecting and streamlining metadata, with the focus that metadata from digital language data resources from all member countries should be harvestable to a central resource discovery service - the *Virtual Language Observatory(VLO)*[4] - and that the metadata should be searchable and contain links to the resources at the different centres.

[3] See clarin.eu for more information.

[4] http://clarin.eu/content/virtual-language-observatory

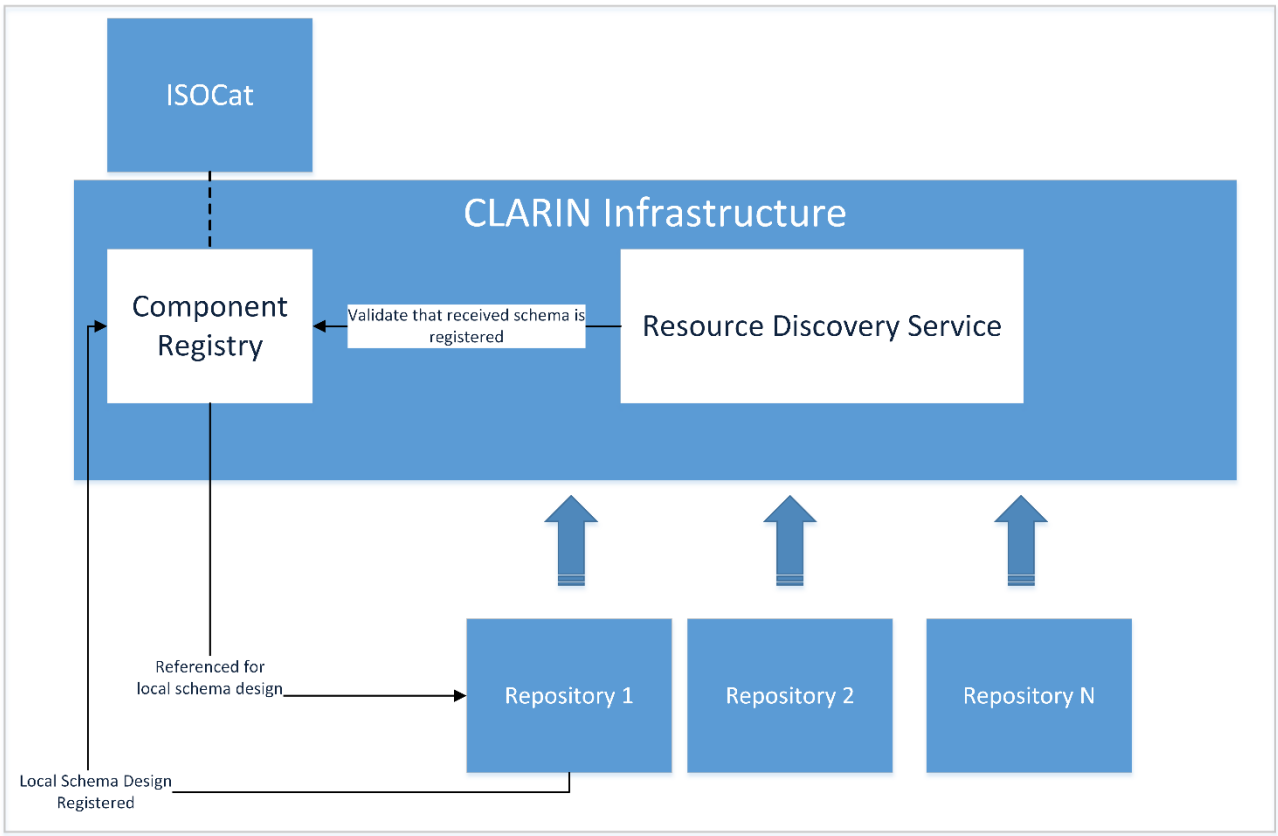


Figure 8: Overview of the CLARIN infrastructure.

One of the goals for CLARIN is to make it easier to document and share metadata xml schemas.

Therefore, CLARIN has developed a platform where xml schemas specifying metadata can be stored and shared. This is not limited to full schemas but also allows users to store and share component parts of schemas covering a small or large part of a metadata structure, e.g. a license-component and an organisation component.

The platform is called the Component Registry[5] and is both a metadata schema development tool and a repository, where metadata schemas following the CMDI guidelines[6] can be defined and stored. The CMDI guidelines are flexible in relation to the the part of the metadata schema that describes the metadata of the content of the resource (cmd.Components part), but has a strict specification for describing the links to the items in a resource in the cmd.Resources-part, and also a strict specification for a few overall administrative metadata specified in the cmd.Header-part.

To ease the documentation of the schemas and in this way facilitate reuse of components, the Component Registry recommends links for each metadata element to the ISOcat registry[7], which allows all to define a metadata element,

[5] http://catalog.clarin.eu/ds/ComponentRegistry/#
[6] http://clarin.eu/content/component-metadata
[7] http://www.isocat.org/

by a definition, translated to multiple languages and with a description of the content of the element. CMDI also suggests that you link to standards specifications of the metadata elements if a commonly used standard exists and can be used.

Furthermore the CLARIN ERIC specifies that the CMDI metadata should include persistent identifiers (PID's) to the resources to allow for a persistent access to the resources that the metadata cover. This will lead the users of the VLO or of the harvested metadata to the resource even it is placed in a repository in a national CLARIN centre, and if it is moved to a new server.

The national repositories can – even using CMDI metadata - have very different information stored about a resource, as the elements defined in the CMDI components are completely configurable by the owner of a CMDI schema. To make the search in the VLO useful, a mapping is carried out between the metadata used in the CMDI components to the search categories in the VLO using the ISOcat references. If the mapping cannot be carried out automatically it can be configured manually in collaboration between the VLO developers and the centre.

Metadata categories and details are very different in the different repositories, as they have individual needs and focus areas to comply with, but as far as the mapping to the VLO is successful, a harmonisation has been carried out, to make it easy for the user to search for and find the resources.

5.3. Metadata in the infrastructure

Figure 9: Data/Metadata Lifecycle to support metadata quality

To get an overview of the metadata gathered within the CLARIN community, the resource discovery service *VLO* can be accessed. Figure 9 shows the facetted search of the VLO. To the right the facets in the facetted search are shown; the CLARIN centres are selected as data providers. Metadata are currently[8] harvested for 571620 resources with CMDI metadata, fron 64 repositories within the CLARIN infrastructure.

A detailed case study of the CLARIN-DK-UCPH repository can be found in appendix G. The CLARIN-DK-UCPH repository is chosen as an example of a national CLARIN repository. The repository exposes metadata for 7810 text resources[9]. The text resources use a specific CMDI metadata schema where the content is defined in the TEI standard but expressed in the CMDI format. All CMDI metadata components and elements refer to an ISOcat definition, which refers to the TEI standard. Centres have also been established at KNAW-DANS and at OEAW CLARIN, but both of these institutions are also involved in other research infrastructures and their descriptions in the appendix will therefore focus more on the other research infrastructures to ensure a wide

[8] VLO visited May 22, 2014: http://catalog.clarin.eu/vlo/search?4. Repository counts cover repositories that offer at the minimun of five resources.
[9] Harvested May 22, 2014.

perspective is provided.

5.4. Initiatives to ensure metadata quality in the infrastructure

Below a short descriptions on the CLARIN initiatives in the area of metadata quality are provided with a focus on the Hillmann criteria (Bruce and Hillmann, 2004)

Completeness
The use of the Component Registry to create, store and share the metadata schemas with links to ISOcat enables all repositories to exchange information about all their metadata information in the CMDI format.

As each repository has its own focus and priorities, the view of what is needed to be complete is decided by each repository. The VLO offers facetted search for 7 values that are taken from the metadata of the resources.

CLARIN has established a Metadata curation task force that the CLARIN centres can join to discuss and promote metadata curation. The work in the task force is currently in its initialisation phase.

Recently some researchers from of the CLARIN community have started work on automatic quality assessment of component metadata. In the paper Trippel et al. (2014) an automatic quality assessment of metadata files is suggested, which refers to the completeness criteria of Bruce and Hillman (2004). This measure described by the authors as the first approach to assessing the quality of highly variable metadata schemes and instances within the CMDI framework.

Accuracy
CLARIN promotes the use of ISOcat, so common and accessible documentation can be found for all metadata categories applied.

Provenance
As it is an assessment criterion for a CLARIN B centre to use registered and public schemas there is log of who created the metadata schemas used. The VLO keeps track of where the metadata is harvested from, but the extent to which the provenance of the metadata are documented and the information is accessible depends on the policy of each repository and can also be different from resource to resource in a repository.

Conformance to expectations
The CLARIN Metadata curation task force is a forum for discussions of conformance of metadata quality within the CLARIN centres community. In CLARIN a focus on communication with users is growing, but as the resources are very diverse the metadata are also very diverse, so CLARIN has currently no overview as to what extent the metadata conforms to expectations.

Logical consistency and coherence
CLARIN ERIC promotes and requires the use of PID's for resources, which gives a consistent access to the resources. Metadata is to a large extent also connected to a PID, by specifying a specific extension to the PID.
As each repository has its own focus and priorities, consistency and coherence can be seen as a problem for the heterogeneous, federated community of CLARIN. The use of ISOcat concept links in the CMDI profile eases the possibility for the user to get the definition of the different metadata element, and to some extent this addresses the challenge.

Timeliness: CURRENCY and LAG
By requiring the use of registered and public - and therefore unchangeable metadata schemas - from the Component Registry, CLARIN ensures that the metadata schemas are unchanged over time. Currently no common policy on changing metadata as such or on versioning metadata is planned by CLARIN, but on a higher level CLARIN has an assessment criteria that ensures the repositories are sustainable over a number of years and comply with the CLARIN criteria.

6. DARIAH's strategies for metadata

6.1. Organisation of DARIAH

DARIAH, the Digital Research Infrastructure for the Arts and Humanities, is a social and technical infrastructure which is composed of people, expertise, information, knowledge, content, methods, tools and technologies for investigating, exploring and supporting work across the broad spectrum of the digital arts and humanities. DARIAH aims to enhance and support digitally-enabled research and teaching across the humanities and arts. The current organizational infrastructure is depicted in Illustration 1.

DARIAH was started in January 2006 as an effort to provide digital services for the various research communities in the humanities under a single institutional umbrella. The idea was to move towards a consortium of institutions which would ensure the long-term sustainability of underlying infrastructures and a strong political voice towards the EU. After going through a preparatory phase of several years, DARIAH submitted an application to the European Commission to establish a European Research Infrastructure Consortium (ERIC) in 2013 and is expected to become an ERIC in summer 2014. This legal framework is meant to facilitate the long-term sustainability of DARIAH.

The Founding Members of the DARIAH-ERIC are Austria, Belgium, Croatia, Cyprus, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, The Netherlands, Serbia and Slovenia. France will be the host country of the DARIAH-ERIC.

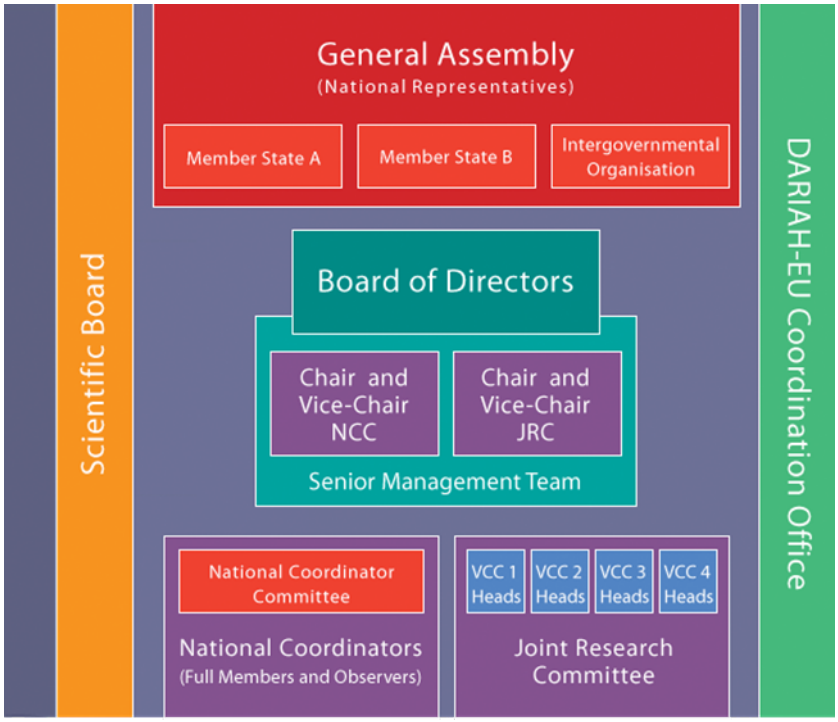


Figure 10: DARIAH organizational infrastructure

By contrast to other ESFRI projects, which build on particular institutions (often data centres), DARIAH organises its activities in a network of so-called *virtual competency centres*. There are four VCCs: (1) eInfrastructure, (2) Research and Education Liaison, (3) Scholarly Content Management and (4) Advocacy, Impact and Outreach. The issue of metadata falls broadly into the responsibility of VCC1 and VCC3, VCC1 being in charge mainly of implementation issues and development. The organization of the infrastructure regarding the data in repositories and archives is depicted in Figure 10.

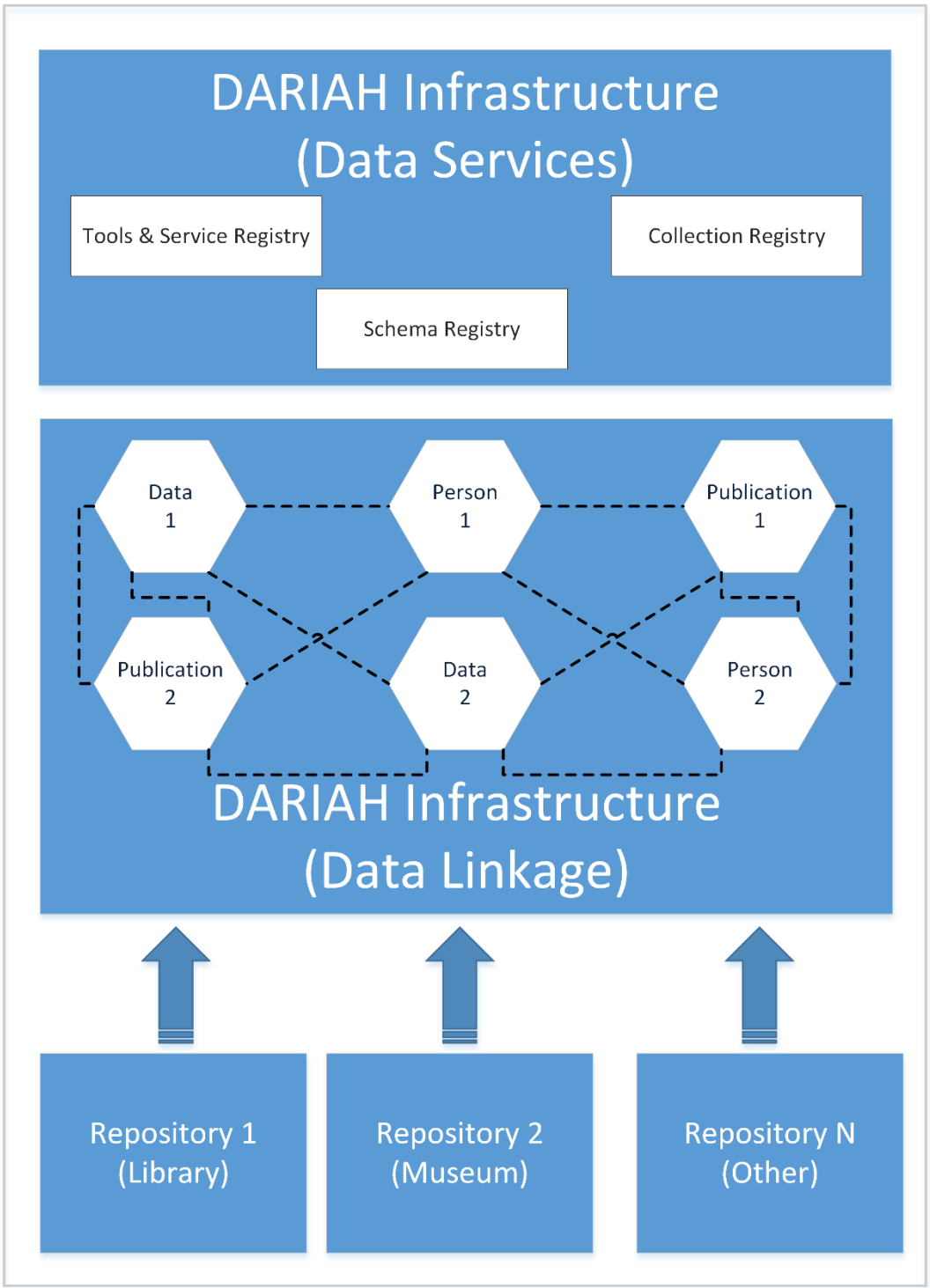


Figure 11: Overview of the DARIAH infrastructure

6.2. DARIAH standardisation strategies

With respect to standards, DARIAH has pursued very cautious strategies, which are motivated by the particular developmental stage many disciplines of the humanities are currently in:

Within a hazardous context in which the idea of going digital is not necessarily mainstream in the humanities, DARIAH has managed to move forward to a stage where it is about to become one of the most stable components in the eHumanities landscape.

In the same statement we read:

Standards are not regulations. There is no obligation to follow them except when one actually wants to produce results that can be compared with those of a wider community. This is why a standardisation policy for DARIAH should include recommendation as to which attitude the scholarly communities could or should adopt with regards to specific standards.

DARIAH attaches great importance to domain-specific best practises. DARIAH encourages the use of widely used standards. A central role is assigned to community oriented approaches. In general, the use of own local formats in projects is discouraged. Instead, projects should demonstrate needs arising from their particular research, needs not covered by the wide range of varieties of already existing initiatives in the digital humanities landscape.

DARIAH does not work on the definition of particular new standards. However, DARIAH members and partners are encouraged to take a pro-active role in helping communities to participate in standardisation activities where they exist and make use of existing approved standards.

6.3. DARIAH metadata strategies

A broad consensus on the usage of metadata and strong collaboration in the area of improvement and maintenance of high quality standards with regard to metadata is vital for the following key benefits, which DARIAH defines as crucial:

- Increased visibility of national research at the European level
- Increased international collaboration opportunities; enhancing exchange of knowledge, skills, expertise, training opportunities and good practice
- Increased potential for the sustainability of the outcomes digital research projects after the end of project funding, helping to ensure the sustainability of tools and services
- Increased access to research data, tools and services via the DARIAH infrastructure
- Increased influence at the European and international level and increased

- Opportunities for funding

In accordance with the above stated more general statements on standards, DARIAH does explicitly not prescribe the use of particular metadata schemes. They rather advocate the use of standards, de-facto standards and best practices.

An important aspect within DARIAH is its relationship with the "memory" institutions, such as libraries, museums, archives, which are seen as potential content providers for the Digital Humanities (DH) research. In many of these institutions Dublin Core plays an important role. These institutions however have had aggregated the information about their collection by various partly long lasting initiatives already for some time, e.g. WorldCat, DBIS, OBVSG (Federation of Austrian libraries), The European Library and Europeana (all memory institutions). Thus it seems worthwhile to consider direct cooperation with the aggregators, instead of duplicating the work of search and collecting individual repositories, while undertaking future efforts of standardization.

6.4. Particular Initiatives in the infrastructure

Inside DARIAH several past and ongoing activities with respect to metadata can be identified.

Collection and Schema Registry
One concrete such activity is DARIAH-DE's collection registry, offering a simple metadata search over resources on collection level using the Dublin Core application profiles format. The Collection Registry offers information on collections of humanities research data. In this context, the notion of collection represents a wide range of entities such as books, documents, texts, files, images or artefacts. Collection descriptions contain general information such as location or access points. The Collection Registry makes use of the OAI-PMH protocol.10

Another activity closely related to the Collection Registry is the Schema Registry which is conceptualised as a central component of the DARIAH federation architecture. It contains schemas that are required for the interpretation of research data contained in the collections listed in the Collection Registry. To facilitate the federation of collections and their respective research data, the Schema Registry further comprises associations between individual schemas—the so-called crosswalks.

Registry for national contributions
Another activity initiated by VCC3 (Scholarly Content Management) was what in internal jargon was called *Umbrella Theme no1*. This was an attempt to develop a straightforward workflow for creating and publishing metadata concerning the in-kind contributions of the various partners across Europe. To perform

[10] http://colreg.de.dariah.eu/colreg/colreg/main?execution=e2s2

harvesting and publishing of data the RDF-triple-store based DH knowledge portal Isidore[11] was used. Practically, the process of metadata creation is based on a light-weight approach to collecting information about partner's contributions by simply enriching HTML web-pages describing their resources via RDFa. Having been created by the partners, these HTML pages are harvested. Consequently, the conveyed information can be ingested into a common triple-store and exploited for various purposes. It is important to note that all of this is still in experimental stage. In the preparation of this workflow, a number of discussions about the metadata fields to be used were conducted in the community, in particular VCC3 and VCC1.

Reference Data and Controlled Vocabularies
In working with metadata, controlled vocabularies have come to play an increasingly important role. The importance of interaction with the communities is even more important, progress can only be achieved through a policy of small steps. One standard of relevance in a particular field is e.g. CIDOC Conceptual Reference Modell which provides an extensible ontology of cultural heritage concepts. The particular importance of controlled vocabularies in the humanities has led to the establishment of a specialised task force under the name *Reference Data and Controlled Vocabularies* (RDCV) at the 2nd VCC meeting in Vienna in November 2012.

The main goal of this expert group has since been to work on infrastructure components ultimately aiming at the establishment of a comprehensive infrastructure for harmonized provision and collaborative maintenance of controlled vocabularies and reference data for the digital humanities community. The task force has grown into a cross-VCC activity bringing together VCC1 (Task 5: Data federation and interoperability), VCC3 (Task3: Reference Data Registries) and also external partners). This work has also been linked VCC2 (Task2, Scholarly Methods Ontology.

Collaboration across Infrastructures
Finally, we should like to draw attention to a very recent initiative by DARIAH representatives and the CLARIN Standards Committee to start thinking about joint activities in identifying useful standards, in looking for communities of interest to contribute and to sharing expertise in the field. After first talks, this initiative will be taken further at the ISO/TC 37 and SCs meeting in Berlin (2014-06-22/27).

6.5. Metadata in the infrastructure

The current state of implementation of the strategies can be observed in the case studies of the two institutions DANS (Appendix E) and OEAW (Appendix F), which are both involved in the DARIAH infrastructure. These case studies were conducted by the authors of this report for the purpose of examining the current practical state of Metadata handling at individual institutions.

[11] http://rechercheisidore.fr/

6.6. Initiatives to ensure metadata quality in the infrastructure

With focus on the Hillmann criteria (Bruce and Hillmann, 204) a short description of the initiatives in DARIAH to ensure the metadata quality is given.

Completeness
The use of collection registries, schema registries and controlled vocabularies is an initiative towards the completeness of metadata. Due to the above mentioned reasons it is difficult to assign a single standard to guarantee completeness. Registries and controlled vocabularies are envisioned to distil and obtain standard metadata approaches, which then in turn can be checked for completion.

Accuracy
A single standard is not defined, as the approach is to obtain common best practices in a bottom up approach, rather than enforcing a single standard.

Provenance
The documentation of the provenance of the metadata depends on the repository. The wide range of resources throughout the repositories, and in some cases even within one repository makes tracking of provenance difficult.

Conformance to expectations
DARIAH's focus is the communication with archives and users. Nevertheless, the vast amount of different types of data makes an overview whether the expectations are met very difficult.

Logical consistency and coherence
The wide spectrum of information types dictates the management of logical consistency and coherence at the repository level. DARIAH functions as a connection between the institutions and promotes good practices, and mainly ensures a communication hub to ensure logical consistency and coherence.

Timeliness: CURRENCY and LAG
The bottom up approach of DARIAH regarding the metadata strategies proposes a very dynamic procedure. As an example the task force under the name *Reference Data and Controlled Vocabularies* (RDCV), at the 2nd VCC meeting in Vienna in November 2012, aims to work on infrastructure components ultimately addressing the establishment of a comprehensive infrastructure for harmonized provision and collaborative maintenance of controlled vocabularies and reference data for the digital humanities community, which will result in a well-established, wide-ranging infrastructure, that can be sustained over a long period of time.

7. Metadata strategies of CESSDA

7.1. Organisation of CESSDA

Since its establishment in 1976, CESSDA has served as an informal umbrella organisation for the European national data archives. As from June 2013, CESSDA is established as a permanent legal entity owned and financed by the individual member states' ministry of research or a delegated institution. Norway will host CESSDA, and the main office is located in Bergen.

13 European countries are member of the new CESSDA: Austria, Czech Republic, Denmark, Finland, France, Germany, Lithuania, Netherlands, Norway, Slovenia, Sweden, Switzerland, and United Kingdom.

The major objective for CESSDA is to provide seamless access to data across repositories, nations, languages and research purposes. CESSDA will encourage standardisation of data and metadata, data sharing and knowledge mobility across Europe. CESSDA aim to play an active part in the development of standards and, even more important, to encourage and facilitate the use of metadata standards for documenting and publishing the existing inventories of research data available from national as well as cross-national resources in Europe.

Each of the member states are represented by a national institution, a Service Provider, which will be responsible for providing the relevant services. The Service Providers will constitute the CESSDA main resource, and CESSDA will integrate the work of the Service Providers and by establishing a one-stop shop for data location, access, analysis and delivery. Software development will increase the quality of available data. Data from sources currently outside CESSDA will also become available. CESSDA will create a more dynamic knowledge management web and will contribute to metadata initiatives. The new CESSDA will also improve existing technical infrastructures and promote capacity building, support less developed and less well-resourced organisations, and work toward a widening of CESSDA.

The Service Providers for the different countries are: Austria - WISDOM, Netherlands – DANS, Czech Republic –CSDA, Norway – NSD, Denmark – DDA, Slovenia – ADP, Finland – FSD, Sweden – SND, France - Réseau Quetelet, Switzerland – FORS, Germany – GESIS, UK- UK Data Service, Lithuania – LiDA

For various reasons some of the nations represented in 'old' CESSDA (the Council) are unable (at present) to participate in 'new' CESSDA (the Consortium). The following countries have some national infrastructure for social science data sharing, but are not currently members of the consortium: Estonia, Greece, Hungary, Ireland, Italy, Luxembourg, Spain, Portugal, Romania and Slovakia. Slovakia has applied for observer status. More

information about CESSDA on the CESSDA website[12]

One of the first initiatives of the new CESSDA was the setup of five work groups to develop an integrated work plan 2014-2015. The in the following section described strategy of CESSDA regarding metadata is based on the ideas in this workplan. As the CESSDA workplan was approved by the General Assembly in June 2014 a proportion of the following sections remains conceptual and has yet to be realized.

7.2. CESSDA's metadata strategies

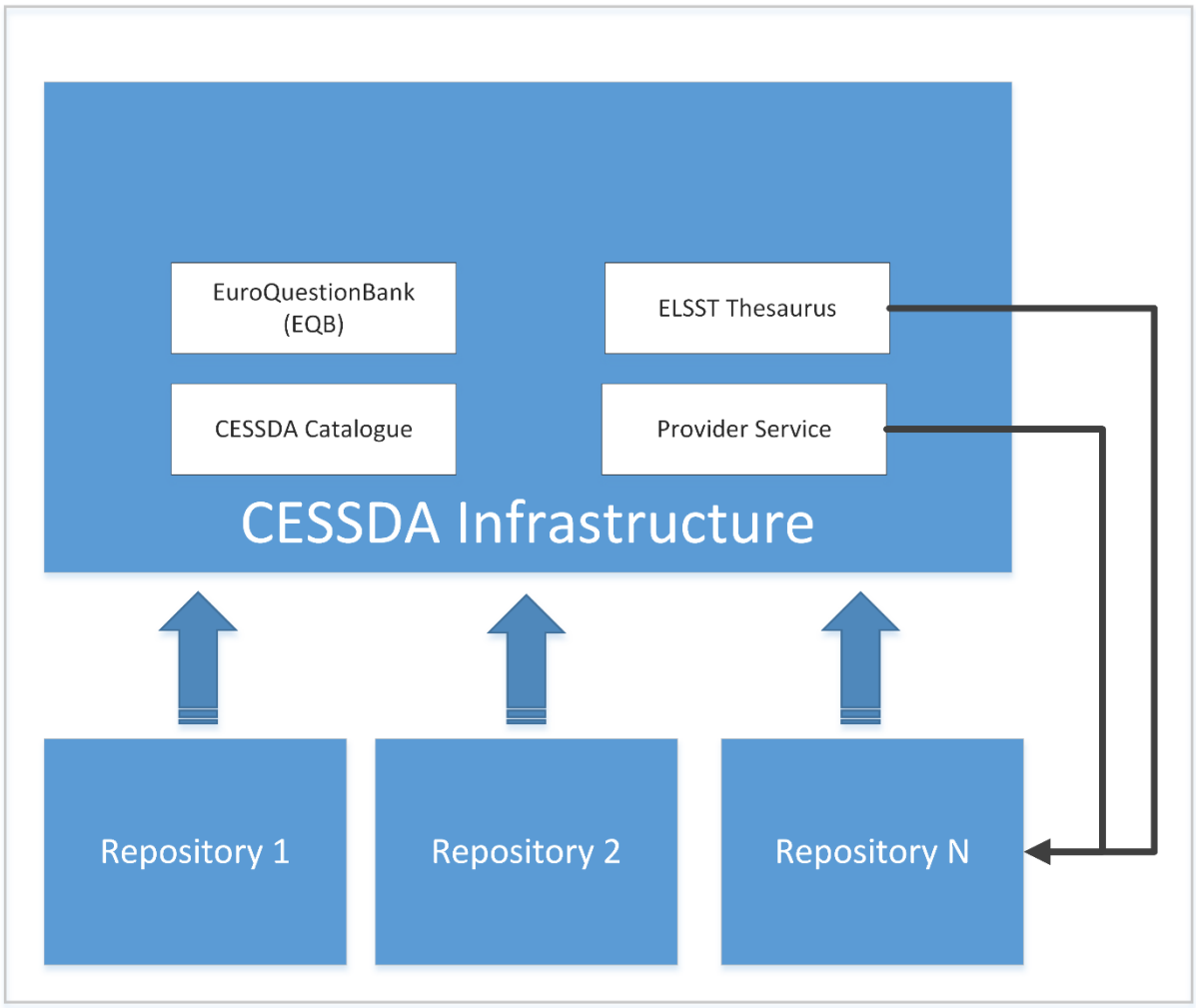


Figure 12: Overview of the CESSDA infrastructure

Metadata standard

The overarching vision for the consortium is to increase the impact of the activities of its Members by providing full scale, integrated and sustainable research infrastructures (Strategic Case for CESSDA-ERIC, April 2011) All activities, including the development of technical services should support this goal. CESSDA's network of services will be underpinned by the Data Documentation Initiative (DDI) metadata standard.

[12] http://www.cessda.net/

The DDI is an effort to create an international standard for describing data from the social, behavioural, and economic sciences. DDI has been branched into two separate development lines, DDI-Codebook (formerly DDI2) and DDI-Lifecycle (formerly DDI3). DDI-codebook can be used to document a single data collection, DDI-Lifecycle metadata supports the entire research data life cycle; it accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.

Work is under way on three RDF vocabularies, the DDI-RDF Discovery vocabulary for publishing metadata about datasets into the Web of Linked Data, PHDD, a vocabulary for describing existing data in rectangular format, and XKOS, an RDF vocabulary for describing statistical classifications, which is an extension of the popular SKOS vocabulary. The public review of all vocabularies is planned for 2014.

At the moment DDI is in particular survey data oriented. Two working groups are developing schemas for other social science research data. One working group is involved in the development of a DDI XML schema for qualitative data exchange, another group is working on recommendations to enhance the DDI specification for better documentation of experimental data, such as randomized controlled trials.

The development of the DDI specification is guided by the DDI Alliance, a self-sustaining membership organisation. Member institutes are research organisations, universities, statistical agencies and data archives all over the world. More about DDI on the DDI website.[13]

Other metadata standards

In the CESSDA workplan it has been recognised that it will be unrealistic to achieve a standardisation across all relevant data sources. "The metadata landscape is still very much in flux, and competing standards, versions, frameworks and initiatives exist in our domain. One example is the NSI community, which is an important provider of microdata in Europe. Currently, the NSI community develops the Generic Statistical Business Process Model (GSBPM) and the accompanying Generic Statistical Information Model (GSIM). These models/frameworks will likely guide metadata production and dissemination from the NSIs in the coming years, and although one likely implementation of GSIM is DDI, other competing standards may very well be formed. CESSDA can promote DDI, but focusing solely on DDI will likely not be sufficient for CESSDA medium to long term" (CESSDA: Proposed Work Plan 2014-2015, page 31).

Furthermore, the workplan recognise that CESSDA needs to monitor metadata development in domains adjacent to social science. To support interdisciplinary and transdisciplinary research, metadata from other domains and data types needs to be sufficiently well understood, and ideally mappings/interfaces to

[13] http://www.ddialliance.org/

other standards/research communities should be built.

Services
The main CESSDA service will be a central one-stop-shop for search/discovery of microdata sets relevant to social science research. The CESSDA Catalogue will be the main interaction point between CESSDA and researchers, and needs to support required functionality and workflows to fulfil this role. The CESSDA Catalogue should enable discovery of data regardless of their access conditions or location.

In addition to the portal is the aim to develop a CESSDA EuroQuestionBank (EQB); a central search facility across all Service Provider's survey holdings in a way that provides as much coverage of survey questions as possible. It should be an accessible single point of entry for question discovery or survey creation. EQB should be based on the DDI lifecycle metadata standard, with provision of a conversion tool from DDI codebook-based metadata.

To ensure increased metadata quality, consistency of discovery and to bridge some of the European language barriers for discovery, the multilingual European Language Social Science Thesaurus (ELSST) will be used within the catalogue. ELSST has been developed over the years by the CESSDA members. The objective is to manage and develop ELSST so that the Service Providers can freely use any or all language versions of the thesaurus, for documenting and as a search tool, also in their local systems. It is important to ensure a usable technical and legal (licensing) framework around ELSST development and management.

Aggregation
The CESSDA Catalogue will automatically, frequently and periodically collect (harvest) metadata from service providers and other sources, and use the metadata to power the Portal's functionality. Fully automated harvesting is crucial to ensure that Catalogue content is updated, and that it points to valid and correct end-points on the service provider-level.

For access to the data, the Portal interface will direct researchers to the source/service where the data may be obtained or analysed, be it data archives, NSI research data centres or to CESSDA-hosted services. Automated metadata collection from decentralized services should rely on common metadata standards and eventually common service APIs. Related to the portal development solutions for the data registration service/persistent identifiers (PID) are going to be developed and implemented. The CESSDA Catalogue will expose its content and services through APIs, so it can serve as endpoint for additional services in the CESSDA community and e.g. cluster services established under ESFRI.

Metadata improvement
The Catalogue should also stimulate content harmonization (both between and within service providers) through a "Provider Service" to the CESSDA Catalogue that enables metadata providers to inspect metadata collected from

them, and to run a battery of quality checks and improvements. This service can serve as a test-bed for metadata harmonisation and alignment tools that later can be commoditized and circulated to service providers that can add them to their local tools portfolio.

Additional to the CESSDA Catalogue the promoting of the adoption of DDI through a programme of training, best practice and up-take of tools to support the creation and exchange of DDI compliant metadata will be carried out.

Obligations of Service Providers
CESSDA have set obligations for Service Providers to become part of the infrastructure in the Annex 2 to the statutes, including annex 2.1:" be fully compliant with the elements of the DDI metadata standard that are required to enable the member to contribute fully to CESSDA-ERIC activities and which will be identified by the CESSDA-ERIC".
To lower the threshold for participation in metadata harvesting, canonical, common-denominator metadata profiles will be developed. These profiles are a subset of the full DDI specification and comprise a CESSDA profile for resource-discovery metadata, a CESSDA profile for EuroQuestionBank metadata, and a CESSDA profile for preservation metadata. Additional to these profiles, identification and development of accompanying controlled vocabularies (CVs) for relevant metadata fields are being accomplished, besides the development of best practices for usage of these respective CVs to ensure maximal harmonisation in usage across CESSDA Service Providers (SPs).

Related Training Activities
Best practice guides and recommendations related to the metadata profiles will be developed to ensure common understanding and harmonised collection and creation of metadata across all SPs. Furthermore, training on best practices regarding collection and creation of metadata and the use of controlled vocabularies will be developed to ensure high quality of metadata.

7.3. Metadata in the infrastructure

In the 'old' CESSDA there is already a common CESSDA catalogue. More than 9000 studies are available from 16 European social science data archives. The catalogue is available via the CESSDA website[14]. Mandatory and recommended elements for DDI codebook[15] and DDI lifecycle[16] were defined, but CESSDA was never able to enforce the use of these profiles because it did not have a legal status as a formal organisation.

Within the Data without Boundaries project (DwB) work package 12, a pilot

[14] http://www.cessda.net/catalogue/
[15] http://www.ddialliance.org/sites/default/files/cessda-rec.pdf
[16] http://www.ddialliance.org/sites/default/files/ddi3/cessda_core_instance.xml

project (DwB, 2013)[17] on metadata quality was carried out. A set of pilot reports generated based on metadata harvested from the Norwegian Social Science Data Archive (NSD) was given back to the NSD data management staff. The reports enabled them to view the collection in an integrated fashion and revealed them quality issues of their metadata. This pilot project shows that despite continuous efforts to keep metadata quality and consistency on a high level, quality improvements can still be made. The Provider Service mentioned above will be based on the experiences of this pilot project.

To illustrate the CESSDA infrastructure, two case studies are being performed. One case study describes the data management at the UK Data Services, the CESSDA service provider of the United Kingdom. The other describes the data management at DANS, the service provider of the Netherlands.

7.4. Initiatives to ensure metadata quality in the infrastructure

In respect to the Bruce and Hillmann criteria (see section 3) a short description of the initiatives within CESSDA to ensure the metadata quality is given.

Completeness
CESSDA's network of services will be driven by Data Documentation Initiative (DDI) metadata. DDI is a defacto standard for the social sciences, internationally recognised by a wide range of research organisations. The two branches of DDI, DDI-codebook and DDI-lifecycle suit the purpose of documenting respectively a single data collection, or the entire research data life cycle.

Although the Service Providers have freedom in the in the level of documentation created locally, CESSDA will develop mandatory metadata profiles which the Service Providers must meet.

Accuracy
Best practice guides and recommendations related to the metadata profiles will be developed to ensure common understanding and harmonised collection and creation of metadata across all Service Providers. To ensure metadata quality, consistency of discovery and to bridge some of the European language barriers for discovery, the multilingual European Language Social Science Thesaurus (ELSST) will be used for cataloguing and as search tool within the catalogue. Furthermore, training on best practices regarding collection and creation of metadata and the use of controlled vocabularies will be developed to ensure high quality of metadata.

The development of a "Provider Service" to the CESSDA Catalogue enables the Service Providers to inspect the metadata collected from them, and to run a

17 DwB Deliverable D12.2 (2013) Enrichment and concersion tool(s) for OS data wokpackage 12 Implementing Improved Resource Discovery for OS Data (internal report)

battery of quality checks and improvements.

Provenance
CESSDA will develop a profile for preservation metadata. Fields for information about how and by whom the metadata is created and how the various transformations have taken place should be part of this profile.

Conformance to expectations
The DDI metadata standard has been developed over the years by different experts from various research institutes and data archives within the social science community. The element set within DDI is the result of a thorough and prolonged development process. The same is the case for the content of the multilingual ELSST thesaurus.

The CESSDA mandatory subset of DDI will encompass all necessary elements.

Logical consistency and coherence
The mandatory CESSDA profiles, the development of accompanying controlled vocabularies and the use of the ELSST will ensure the logical consistency and coherence of the metadata within the CESSDA infrastructure.

Timeliness: CURRENCY and LAG
To become a Service Provider of CESSDA the repository should apply for a Data Seal of Approval[18], a basic certification standard for sustainable trusted digital repositories. In addition, although the DDI standard will evolve, the DDI alliance ensures that the branches are compatible with each other.

The currency of CESSDA portal data is ensured by regular harvesting but the issue of lag is dependent on the timely updating of collections by the Service Providers.

Accessibility
Related to the CESSDA portal development solutions for the data registration service/persistent identifiers (PID) are going to be developed and implemented.

[18] http://www.datasealofapproval.org/en/

8. Cross Fertilisation between CESSDA, CLARIN, and DARIAH

The Infrastructures CESSDA, DARIAH, and CLARIN are very different in organisation, research areas and focus of interest as can be seen in sections 5, 6, and 7 of this deliverable. The infrastructures differ so much in visions, strategies and initiatives; that at a first glance they might not seem to have a large overlap. Nevertheless, a number of elements of the intensions, plans, and initiatives touch upon the same issues and challenges for metadata and in the following we will sum up the elements that have an impact on metadata quality.

8.1. Sharing lifecycle models, descriptions, and diagrams of infrastructures

As shown in section three "Metadata lifecycle", a number of the models for data lifecycles originate from different communities and different situations. The metadata lifecycle we developed, is presented as a baseline which, alongside an understanding of metadata quality evaluation and metadata types, can be used to design and benchmark a local approach to describing, delivering and improving quality metadata.

The knowledge of the different lifecycle models for data, and the awareness of metadata having a distinct role in the lifecycle are important aspects to be aware of when planning to produce high quality metadata.

The ability to identify actors, agents, and roles of those actors and agents in the metadata lifecycle model is an important step forward to be able to describe the metadata creation process and to describe where quality issues can be discussed.

We hope that the proposed lifecycle model can be used when going into detail about identifying the challenges in creating high quality metadata in each repository.

We have also tried to describe the three infrastructures related to the common figure in section 4 of a super-infrastructure, to open up a common understanding on the propagation of metadata in the infrastructures; where metadata about a digital object propagates from a repository system to an infrastructure system, and potentially further up to a super-infrastructure. When the super-infrastructure receives metadata from multiple infrastructures then a simplified metadata schema for interoperability harvesting will often be used - also in cases where the original repository may offer richer metadata.

The Joint Metadata Repository, which will be described in section 9, currently harvests directly from a large number of repositories and is at the moment not

such a super-infrastructure, however one could easily imagine that the Joint Metadata Repository could evolve into a super-infrastructure.

A super-infrastructure will have a less challenging job during harvest, as only a few infrastructures will have to be addressed. However this architecture requires a two-phase mapping of metadata if the resulting metadata at the super-infrastructure level is to be usable. This reveals the challenge that metadata quality internal to a specific repository involves different aspects than those for an infrastructure or a super-infrastructure.

Furthermore, it is difficult for an infrastructure harvesting metadata to improve the metadata quality at the infrastructure level. If metadata quality has to be assured, this should be done on repository level, as low-quality metadata can propagate up to the infrastructure harvesting the metadata from each repository.

Collaboration and agreements between repositories are therefore needed, and in the next sections some aspects will be mentioned as important areas for metadata quality.

8.2. Mandatory or recommended metadata profiles

CESSDA, CLARIN and DARIAH all agree that it is unrealistic to achieve perfect standardisation across all relevant data sources and disciplines, and all mention that it is important to agree to promote collaboration on metadata profiles within each research community in the SSH area. They also agree that the most broadly used metadata standard, the Dublin Core standard is not sufficient for the different communities.

Nevertheless, the three infrastructures promote different initiatives for sharing and standardising metadata. CLARIN has developed the CMDI Component Registry where schemas and metadata components can be registered and shared. CESSDA has the DDI schemas to share, and DARIAH has a schema registry planned. Broadening the knowledge of metadata schema registries between the infrastructures would be beneficial and the exchange of information of these initiatives would be important.

Recommendable the three infrastructures could agree to define a common list of metadata elements that - crossing the different communities and standards – can be used as compatible between the different communities. Furthermore, easy-accessible definitions of these elements and mappings across the different metadata standards should be available.

8.3. Sharing of knowledge and linking of resources

Different initiatives in each of the infrastructures are active or planned, concerning discovering vocabularies, Simple Knowledge Organisation Systems (SKOS) and references to definitions of data categories.

The use of ISOcat definitions as reference for metadata elements within CLARIN could be shared with the other infrastructures. In the same way information about the DARIAH Controlled Vocabularies and the CESSDA DDI-RDF Discovery vocabulary for publishing metadata could be shared among the other infrastructures.

These systems do not have to outperform each other, but sharing the ideas and knowledge between the infrastructures are expected to be beneficial.

Furthermore, DARIAH has established an expert group with the main goal to work on infrastructure components aiming at the establishment of a comprehensive infrastructure for harmonized provision and collaborative maintenance of controlled vocabularies and reference data for the digital humanities community. CLARIN has an Open Vocabulary initiative based on OpenSKOS, which should enable to registry e.g. organisations and concepts with definitions that can be shared within the infrastructure. CESSDA has developed the multi-lingual concept thesaurus ELSST, and is developing other controlled vocabularies. In this area knowledge sharing between the infrastructures would also be beneficial.

8.4. Discussion on metadata quality aspects between and within infrastructures

Metadata quality must be discussed in relation to the activities for which they are used. We suggest that the infrastructures DARIAH and CLARIN prioritise future collaborate about standardisation efforts, which have already been initialised in dialogue between the CLARIN Standards Committee and the DARIAH representatives.

To improve metadata quality it is important to have a feedback mechanism about metadata issues, which could promote the discussion of metadata quality. Therefore, we suggest that each infrastructure should agree on how to disseminate information about missing metadata, questions on metadata content, or difficult mappings of metadata back to the repositories it receives metadata from. Such a feedback mechanism – manually or automatic – can form a basis for the metadata quality discussion in each infrastructure.

9. Using the DASISH Joint Metadata Repository Prototype to exemplify challenges on Metadata Quality

In this report the focus has until now mostly been to describe issues and challenges with metadata quality without looking into actual metadata elements and values. In this section we will inspect some challenges on Metadata Quality when looking at real metadata.

In the DASISH task 5.4, a DASISH Joint Metadata Repository[19] prototype (JMD repository), with a search interface is under development. The prototype includes metadata from repositories in the infrastructures CESSDA, CLARIN and DARIAH, where the metadata is harvested via the OAI-PMH protocol. Both facetted and free text search in the metadata can be performed.

The web interface of this repository is in this section used as an exemplification of some of the challenges on metadata quality. By using its web interface metadata harvested from the three infrastructures can be inspected grouped depending on which infrastructure the data are harvested from.

The prototype was put online on April 23 2014, and the status of the prototype is that metadata from the three infrastructures has been included with the following number of metadata records:

- CESSDA: > 35,000 metadata records
- DARIAH: > 400,000 metadata records
- CLARIN: >140,000 metadata records

Please note that this portal is still a prototype of the metadata catalogue repository. The metadata from the different communities is mapped onto a single and relatively simple format used by the repository; this causes some loss of information. From a metadata quality point of view, we will not focus on this loss of information, but instead some examples of metadata quality challenges of are given below.

The prototype of the JMD repository offers a facetted search for a few facets:

- Groups
- Creator
- Discipline
- Language
- Subject
- CreationDate

The facetted search can be supplemented by free text search in the metadata

[19] http://vmext24-215.gwdg.de/ckan/dataset

by adding text in a search field, to enable search for specific information across metadata fields, or enable search in the description field of the resources.

In this section we would not create a review of the JMD repository interface as such, but to focus on some of the issues that can be discussed under the headlines of the Hillmann criteria using the metadata elements 'CreationDate', 'Creator', 'Language', and 'Discipline' as examples.

9.1. CreationDate

CreationDate is a good example of a metadata element value where the mapping onto a common repository seems to have succeeded well. All files seem to have values for CreationDate in their harvested metadata and only a few items can be found with remarkable values[20] that can be difficult to understand. Those issues mostly seem to occur when the harvested metadata have more than one CreationDate value included, e.g. for the resource *Live de choeur*[21] the date is stated to be: *"1501/1600; 1701/1800"*, and the harvested metadata contains the information:

<dc:date>1501/1600</dc:date>
<dc:date>1701/1800</dc:date>

Hence, CreationDate seems to represent a metadata element, which fulfils *completeness, accuracy*, and *conformance to expectations*, and *accessibility* for the majority of the metadata in the JMD repository.

However, resources stating *"1501/1600; 1701/1800"* as creation date might give the user the impression that CreationDate is the date, when the original resource was created. But looking at another example - a resource called *"Jørgen Rischel Collection"*[22] - with CreationDate *"2014-01-24"* shows that the CreationDate is the date when the metadata was created in the repository as the harvested metadata contains the information:

"<MdCreationDate>2014-01-24</MdCreationDate>"[23].

So even if it, at first glance, seems obvious to the user what CreationDate contains, there is a number of interpretations. Collecting this information in the same data field blurs the information for the user, as the field currently contains both creation date of the source of the resource, and the creation date of the metadata.

A suggestion could be to use two separate fields for *"metadata creation date"* and

[20] http://vmext24-215.gwdg.de/ckan/dataset/756a6f10a2ee44267b2c11ff52900469c46dacaa1494f81fec078ea9f0ae15e5

[21] http://vmext24-215.gwdg.de/ckan/dataset/2695024ec9eb1dc69589d40ab5eeb8f3c5bb8b31454ee55be3b554e791b212c2

[22] http://vmext24-215.gwdg.de/ckan/dataset/d20c3a7aff5e9274d30e7fe916fc53c99728fb84461193008ae4210d119d7005

[23] The Jørgen Rischel Collection is digitised during 2013, with the goal to preserv the audio recordings and other material of Jørgen Rischel after he died in 2007. The material was uploaded to the RWAAI repository in January 2014.

the *"creation date of the resource"* to increase the criterion *conformance to expectations.*

9.2. Creator

For Creator most of the values are personal names, forming a very long listing of values see figure 13. Please notice that creators like *"Family"* are also available. The specification of "Family" as the metadata value for creator might not be wrong, but might not fulfil the criteria of *"conformance to expectation"* and *"accuracy"* for the metadata Creator. Other examples of values with the same challenge, which cannot be seen in the figure, are *"UZ"* and *"PP".* The most used value for creator is *"Not applicable",* which does not supply the user with any information, so the *completeness* criterion is not fulfilled for Creator.

For Creator one could suggest that a Researcher Identifier could be created, so it would be possible to track the data that a specific researcher produces without being depending on the name alone. This suggestion might go beyond the authority of the single repository, and we will not go into details with it.

Another facet to display to the user could be the information about the repository supplying the metadata. The repository will usually have more detailed information about the resources that would be more useful for a researcher/user than the value of the *creator*. The values of the repository supplying the metadata could be expected to be *accurate*, *complete* and also to state some *provenance* information of the resource.

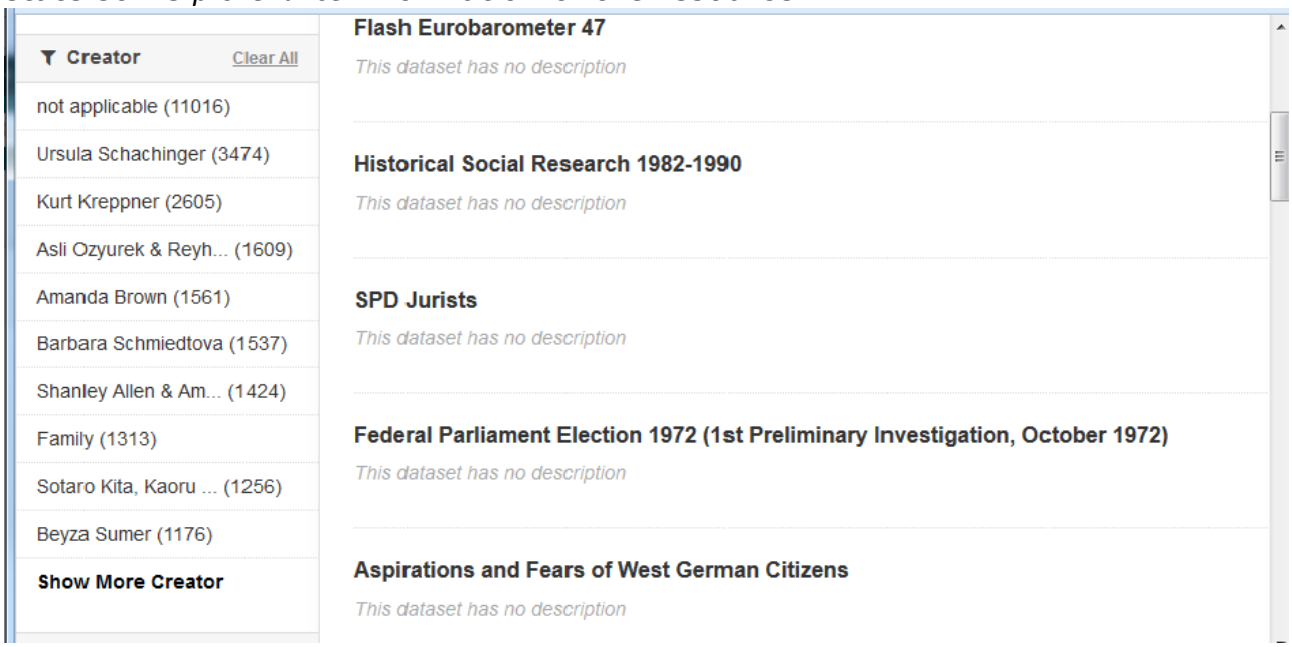


Figure 13: Top Creator values for the JMD Repository

9.3. Language

Language is an example of a metadata element, which at the first glance could seem to be straightforward to harmonise, but looking at figure 14 below, it reveals that this is currently not the case in the JMD repository.

For German at least the following values are used: *"de"* (ISO 639-1), *"ger"* (ISO 639-2), *"deu"* (ISO 639-2) and *"German"*. The first three values could be mapped to the same value, but for *"German"* a more detailed investigation need to be carried out to determine if it can just be mapped to the same value, as it does not clearly refers to a standard like the other three values do.

For English and Dutch and other languages the same issue occurs. Mapping of *"en"*, *"eng"* and *"English"* to the same value might give an easier search for the user, as well as mapping "nl", "Dutch" and "In het Nederlands" to the same value.

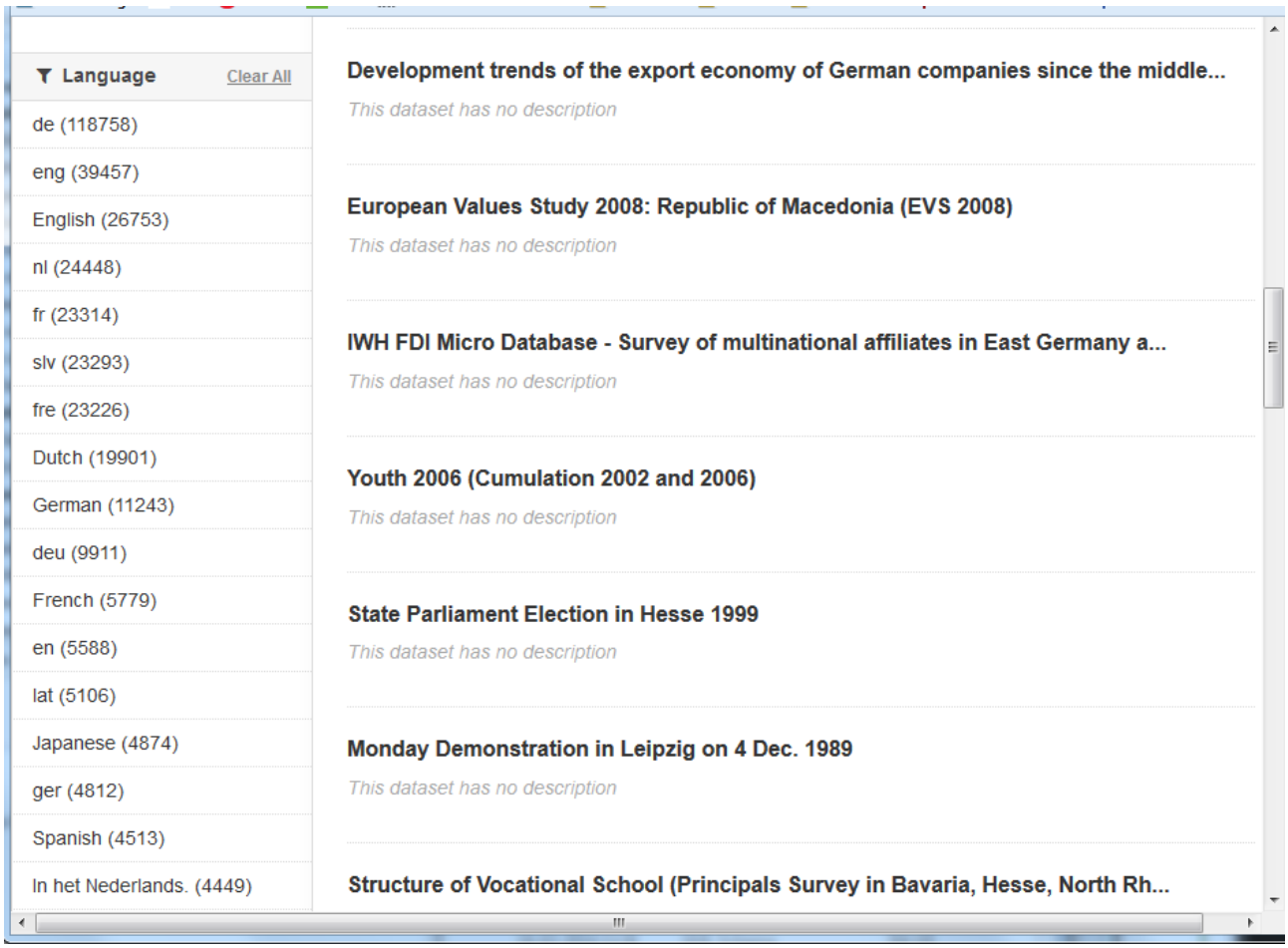


Figure 14: Top language values for the JMD Repository

9.4. Discipline

For the metadata *Discipline*, see figure 15, the values are on different conceptual level; as *discourse* is a generalisation of the concept of *conversation* within all modalities and contexts[24]. For a user it is not clear if *linguistics* is an even broader term than *discourse.* While other values as *Movie description* and *Picture description* might be easier to understand. For the user to benefit from metadata values gathered for a facet like *Discipline*, the challenge is to be able to explain

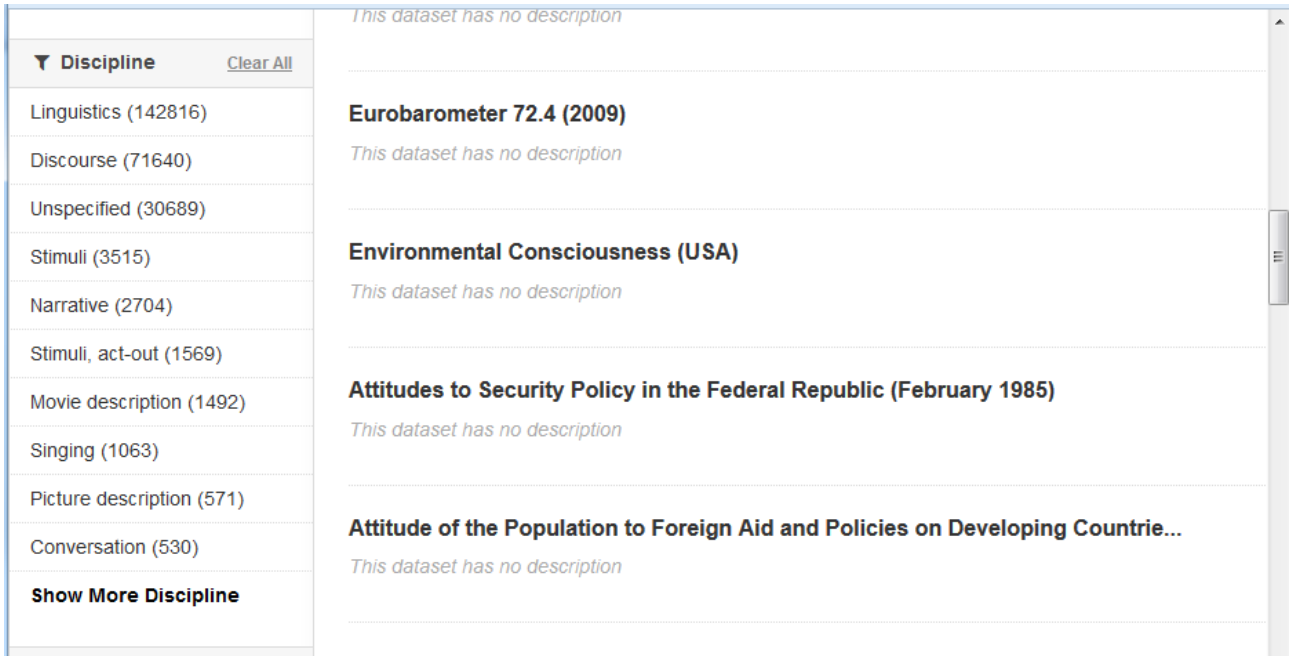


Figure 15: Top Discipline values for the JMD Catalogue Interface

the concepts used and their internal relations, as it might not be easy to understand. It might be very useful for the user, to have the mapping that is carried out for harvesting the metadata, accessible. Even better, if a thesaurus was used and available to the user.

9.5. Summing up

Evaluation of the values for these four aggregated metadata elements illustrate that real metadata brings more facets to the discussion of metadata quality. However, for some metadata elements it seems to be doable to agree on a mapping that can facilitate a joint metadata repository.

Gathering information from very different kind of sources in one search interface will need to generalise on some issues and leave many details out.

[24] http://en.wikipedia.org/wiki/Discourse

We suggest that the JMD repository adds the information from which repository the metadata is being harvested.

Looking at this portal makes it obvious that the metadata fields do not meet the Bruce and Hillmann criteria *"conformance to expectation"* and *"accuracy"* due to the fact that the various underlying repositories have different meanings for the same metadata element. A super-infrastructure will have a less challenging job during harvest, as only a few infrastructures will have to be addressed. However this architecture requires a two-phase mapping of metadata if the resulting metadata at the super-infrastructure level would be usable. A common list of metadata elements, easy-accessible definitions of these elements and mappings across the different metadata standards should improve the metadata quality of the aggregation. Further coordination is therefore needed.

10. Conclusion

This task undertook an examination of issues surrounding metadata quality from a full lifecycle perspective. Relevant metadata types and associated metadata aspects were defined from the data curator as well as the data creator perspective and this initial desk research (described in Appendix B) was used to provide a foundation for the analysis of metadata strategies across the three in-scope European infrastructures. The analysis identified a number of areas where cross-fertilisation of approaches to metadata may be of mutual benefit across CLARIN, DARIAH and CESSDA. This work forms the basis for derived training materials.

To better understand the structure and scope of issues surrounding metadata and metadata quality management, a number of research data lifecycles were examined. Each lifecycle reflects a particular business process model or research or curation focus. So a degree of variation is to be expected, but none of the lifecycles entirely addresses the curatorial metadata management perspective. It became clear during this investigation that most lifecycle illustrations have a logical and natural tendency to focus on the 'data' part of the digital object, which is the focus of research. There is an implicit assumption that the primary purpose of the lifecycle is to convey the journey of the pure, untouched, canonical output of research from producer to final consumer and while this is generally accurate it doesn't address the more dynamic nature of the surrounding metadata (both object metadata and administrative metadata) necessary to manage the full lifecycle. While from the researcher perspective the digital object and its metadata may feel 'complete', the design, redesign, creation and management of metadata remain current issues from those curating the remainder of the lifecycle. As repositories continue to update to new standards and re-enrich metadata to meet the changing needs of their target community we adapted the familiar lifecycle models to support a more dynamic view of metadata. This metadata lifecycle provides a structure for our report and may be used, alongside an understanding of metadata types and quality evaluation, to design and benchmark local approaches to the describing, delivering and improving metadata. The UK Data Service case study validated the lifecycle by employing it in its description of metadata management in practice.

A dual approach was taken to examine metadata management across the infrastructure. In the first instance a general analysis of policies and strategies was undertaken and each infrastructure was evaluated through the lens of the Bruce and Hillman metadata quality criteria. This was supported by a second strand of investigation, which undertook case studies of individual data repositories from within each infrastructure. The data and metadata management at four local institutes were examined: The UK Data Service at the UK Data Archive for CESSDA, the CLARIN-DK Repository at University of Copenhagen as a CLARIN B centre, the Austrian Center for Digital Humanities illustrates a both a CLARIN centre and a DARIAH repository, and the DANS

repository which encompasses, CLARIN, DARIAH and CESSDA.

The visions, strategies and initiatives around metadata vary across the CESSDA, CLARIN and DARIAH as well as across their member repositories, but there remain a number of common issues and challenges which opens the possibility for cross-fertilisation. One recommendation would be that the three infrastructures could agree to define a common list of metadata elements that could be deployed across the different communities. Furthermore, easily accessible definitions of these common elements and mappings across the different metadata standards should be available. Moreover, sharing of knowledge and linking resources would be beneficial for all the three infrastructures. Especially the sharing of knowledge about linked data initiatives concerning discovering vocabularies, Simple Knowledge Organisation Systems (SKOS) and references to definitions of data categories, currently active or planned in all three infrastructures, would leverage these developments.

Finally we evaluated the challenges of metadata quality by looking at the actual aggregation of metadata. In DASISH task 5.4 a joint CLARIN, DARIAH and CESSDA metadata portal is under development and the preliminary results of this aggregation prototype, which harvests from a number of repositories, were evaluated. An examination of the portal made it clear that some metadata fields do not meet the Bruce and Hillmann criteria *"conformance to expectation"* and *"accuracy"* due to the fact that the various underlying repositories have different meanings for the same metadata element. A super-infrastructure will have a less challenging job during harvest, as only a few infrastructures will have to be addressed. However this architecture requires a two-phase mapping of metadata if the resulting metadata at the super-infrastructure level are to be usable. As stated above, a common list of metadata elements, easy-accessible definitions of these elements and mappings across the different metadata standards should improve the metadata quality of the aggregation. Further coordination is therefore needed.

Metadata quality must be discussed in relation to the activities for which they are used. We suggest that the infrastructures DARIAH and CLARIN prioritise future collaboration about standardisation efforts, which have already been initialised in dialogue between the CLARIN Standards Committee and the DARIAH representatives. Similar initiatives could be established with CESSDA.

The effort available to task 5.3 was sufficient to touch upon several issues of metadata management and metadata quality. The development of real inducements of cross-fertilisation was beyond the scope of this task.

PART B: PORTAL PROGRESS REPORT

11. Introduction

Creating an interdisciplinary metadata catalogue, however limited in scope, is a task that transcends the usual available expertise available (if any) about metadata and metadata practices in a single discipline. This is due by the varying approaches taken by different disciplines with respect to the type of metadata resource descriptions, (see part A of this deliverable). It also by the need to decide on the technologies required to bring the targeted disciplines metadata together to create the DASISH Joint Metadata Domain (JMD).

The first challenge is to collect as much good quality metadata from the different disciplines as possible. The search for available metadata may be complicated if the communities themselves have no organized system for publishing metadata or if there is no systematic inventory available. Fortunately the predominant technology within the disciplines themselves, the Open Archives Initiative – Protocol for Metadata Harvesting[25] (OAI-PMH), is well suited for gathering the metadata cross-discipline.

The second challenge is to merge the metadata from the different disciplines in a metadata catalogue, so that users can find metadata from different disciplines aggregated in result sets when browsing or searching the collected metadata. To make this possible the semantic interoperability issues with respect to the metadata sets used by the different communities must be overcome.

In agreement with the terminology used by OAI-PMH, we refer to those centers, institutes or organizations that make metadata available to others as metadata providers. Centers that harvest (aggregate) the metadata from the metadata providers to for instance display the data in a metadata catalogue are called 'metadata service providers'.

Depending on the technology chosen for the metadata catalogue (faceted search & browsing or a complex query interface), there is a need for mapping the different metadata schema elements on a set of facets (faceted search) or finding crosswalks between different metadata schemas.

The DoW describes the purpose and work of task 5.4 in some detail but it was not always applicable under the circumstances. We found that information gathering about the organized use of metadata by the communities required more intensive effort when investigating the metadata availability and practices, especially of the DARIAH infrastructure

With respect to the task of concept registration for all the encountered SSH metadata schemas, we decided focusing on registering only the concepts used

[25] http://www.openarchives.org/pmh/

in the DASISH metadata catalogue.

In the DoW, a relationship is described between the reporting of tasks 5.3 and 5.4, which in practice could only be partly realized. For instance the necessary parallel execution of 5.3 and 5.4 prevented us from using detailed input and results of task 5.3 for 5.4. Vice-versa, results of 5.4 were only partially available for the 5.3 reporting. However useful feedback was available to task 5.4 (chapter 9 of the 5.2A report) in the end.

12. The Use of Interdisciplinary Metadata Catalogues

The purpose of the DASISH JMD (or DASISH metadata catalogue) is similar from other interdisciplinary metadata efforts like DataCite[26], the EUDAT project with B2FIND[27] and OpenAIR[28]. Big metadata catalogues enhance the visibility of and provide easier access to research data. Researchers are able to search and easily find data that are useful for their research purposes from the whole SSH. Projects like the DASISH JMD also promote a culture of data sharing, data reuse, verification and citation and help to increase the "acceptance of research data as legitimate, citable contributions to the scholarly record"[29] by making data visible and accessible for the broader research community.

We can summarize further reasons for the DASISH JMD:

- An interdisciplinary catalogue for the SSH facilitates cross-disciplinary research interests of projects within these fields of research. Within the SSH, the cross-disciplinary research interest would be higher than between unrelated disciplines. This is of course also one of the motivations that underlie the DASISH project.
- The DASISH JMD should offer access to the whole[30] SSH published metadata domain via a single application, facilitating cross-SSH domain resource discovery.
- It shows specific metadata records in the context of other SSH metadata.
- It offers a metadata discovery tool, a tool not available to some research infrastructures e.g. DARIAH, (even though there are some (national) initiatives within DARIAH for establishing a resource catalogue (see below) there is no overall agreed upon approach for metadata aggregation and resource discovery).

Metadata catalogues are available at different levels of granularity and aggregation. A metadata catalogue can contain the metadata from a specific organization or institute, from a specific community of multiple institutes, or it can contain multi-disciplinary level data that comes from different communities. There are also subject-oriented metadata catalogues that concentrate on specific subjects. The Registry of Research Data Repositories counts 244 repositories for the SSH[31]. As examples of community-specific metadata catalogues, we present:

- The Virtual Language Observatory (VLO)[32] for CLARIN
- The CESSDA Data Catalogue[33] for the Social Sciences

[26] DataCite catalogue: http://search.datacite.org/ui

[27] EUDAT B2FIND catalogue: http://www.eudat.eu/b2find

[28] OpenAIR catalogue: https://www.openaire.eu/search

[29] http://www.datacite.org/whatisdatacite

[30] In practice the amount of published metadata visible through the JMD is smaller due to compute resources and time constraints.

[31] http://service.re3data.org/search

[32] http://www.clarin.eu/VLO

[33] http://www.cessda.net/catalogue/

- The collection registry[34] or the knowledge portal Isodore[35] for DARIAH, (see also section 6.4 in the Deliverable D5.2A)
- Catalogues in the archaeological community:
  - The search service of the UK archaeology data service[36]
  - The central object database ARACHNE[37] run by the Archaeological Institute University of Cologne

  Within the RI for archaeology ARIADNE, a common Metadata Registry is under construction.

[34] http://dev3.dariah.eu/search/
[35] http://rechercheisidore.fr
[36] http://archaeologydataservice.ac.uk/archsearch/
[37] http://arachne.uni-koeln.de/

13. Implementation

The DoW for task 5.4 specifies a joint metadata domain implemented by harvesting available metadata from metadata providers from all the participating infrastructures and presenting this metadata after suitable mapping transforms in a joint metadata catalogue.

In line with the DASISH DoW the metadata was to be visualized as a faceted browser. Faceted browsing (or faceted search) is a way to find records in a collection based on a system where each record is classified along multiple explicit dimensions (the facets) that correspond to the properties of the records. Users can navigate the different dimensions independent of other facets. In a faceted browser for metadata, the metadata elements are mapped to selection lists (the facets) in the User Interface (UI) that a user can use to filter the metadata.

Following the DoW we structured the work in the following subtasks:

- Collecting lists of OAI end-points (or if needed using other transfer methods) from the different participating communities.
- Choosing a suitable catalogue software and where necessary adapting it for our purposes. In our case, we used and adapted CKAN as our suitable catalogue software.
- Decide on the set of facets shown in the metadata catalogue
- Creating suitable mappings between the schemas used by the harvested metadata and a common set of facets shown in the DASISH catalogue.
- Harvesting and in collaboration with the partners, refine the set of shown facets and mappings.
- Use the output of DASISH task 5.3 (Deliverable 5.2A) to draw some conclusions about usability and form of a common SSH catalogue.

13.1. The SSH Metadata Providers

For the creation of interdisciplinary metadata catalogues or portals there exist basically two different architectures that can be applied also mixed. The first approach is directly harvesting metadata from the original metadata provider centers. The second method relies on other metadata service providers (aggregators) that republish the metadata they harvest for harvesting by other metadata service providers[38]. This is referred to as "harvesting the harvesters". Within the SSH, however, it is difficult to apply the second strategy since, at the time of writing, none of the participating infrastructures has such harvestable harvesters.

Therefore as a first step, lists with metadata-providing centers (preferably OAI-PMH) were needed. Input was requested from the partners in 5.4 from the different participating infrastructures. For CLARIN, such list existed in the

[38] We mentioned the OpenAIR[28] and EUDAT[27] metadata catalogues as providing such a service

form of the CLARIN center registry[39] with only minor updates necessary. For CESSDA and DARIAH, no such comprehensive lists existed. It was decided to start contacting the infrastructures and their institutions trying to reach as many institutions as possible.

After initially not getting many results, all the DASISH task partners were urged to be more persistent in contacting institutions from their respective infrastructures and request information about used meta-data standards and harvesting possibilities. After another round of consultations some more end-points were discovered, but the method of contacting the institutions directly was not very efficient. Also, the knowledge about such end-points within the infrastructures seems difficult to locate. Especially within DARIAH our attempts produced no results. Instead, a wide search using web-available information was performed and some new information sources were found. This included lists of institutions providing metadata harvestable over OAI-PMH, including also DARIAH, CESSDA and CLARIN centers.

Using CESSDA's director's e-mail distribution list in January 2014, CESSDA institutions were contacted. After few results the missing institutes were contacted directly via e-mail or phone call. The response was not complete before September 2014. One important CESSDA metadata provider, the "CESSDA catalogue" was a candidate for the harvesting of the harvesters strategy mentioned earlier. The CESSDA Catalogue[40] harvests metadata from the Nesstar servers[41] maintained by the CESSDA member data archives by using a proprietary protocol without the possibility to harvest via OAI-PMH. However, there is an early release of an additional component that provides OAI-PMH capabilities for Nesstar servers, which some CESSDA members have installed. A few CESSDA members are also members of DataCite[42], an organisation whose portal provides access to metadata via OAI-PMH. CESSDA plans to rebuild their catalogue during 2015.

"Digital heritage" institutions (libraries, museums, archives) are seen as potential content providers for Digital Humanities (DH) research. Given the relationship of the DARIAH community with these institutions it is worth considering tapping these providers' content. These institutions already mandate aggregating information about their collections due to long-lasting initiatives, such as WorldCat, DBIS, OBVSG (Federation of Austrian libraries), The European Library and Europeana. Thus it seems worthwhile for the operators of interdisciplinary catalogues to consider direct cooperation with these aggregators instead of duplicating the tasks of searching and collecting individual repositories and following the "harvesting the harvesters" strategy. However, in the process of collecting OAI endpoints we were confronted with some institutions that provide metadata for Europeana - being ostensibly keen to support DH research - appear reluctant to make their OAI-PMH endpoints

[39] http://centerregistry-clarin.esc.rzg.mpg.de/
[40] http://www.cessda.net/catalogue
[41] http://nesstar.com
[42] http://www.datacite.org/

public. The concern seems less about access/licensing restrictions for the metadata, but rather an attempt to limit the administrative effort and the load on the technical infrastructure.

One source of information was also the collection registry of DARIAH-DE that harvests metadata from a number of data providers in Germany and lists the collected endpoints[43]. Another suggestion was contacting high-level representatives of the DARIAH infrastructure and requesting their assistance with the discovery of OAI-PMH end-points. We contacted the DARIAH-EU Coordination Office at the Göttingen Centre for Digital Humanities (GCDH) and obtained a list with DARIAH institutes that could possibly be data providers. All these centers were requested to send us the required information.

A complete overview of all metadata providers we encountered can be found in Appendix G: List of SSH Metadata Providers. However the providers included in DASISH catalogue are a only subset of this list for reasons of limited human and computational resources.

13.2. SSH Metadata Frameworks and Schemas

Our inventory of metadata providers, metadata portals and schemas in the SSH resulted in this list and became the basis for later mapping and normalization work.

DDI (DDI Codebook, DDI Lifecycle)
The Data Documentation Initiative[44] (DDI) is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.

DC
Dublin Core[45] (DC) defines a base set of metadata elements and is commonly used as a baseline in other standards. When harvesting via OAI-PMH this format is the most common one and according to the OAI-PMH rules, it should be mandatory and be provided alongside other formats, although this rule is sometimes violated.

CMDI
Component Metadata Infrastructure[46] (CMDI) is the metadata infrastructure of the CLARIN research infrastructure. All CLARIN centers have to provide metadata in this format for the resources they contribute to the CLARIN infrastructure. The metadata is harvested and aggregated in the Virtual

[43] http://dev3.dariah.eu/search/collections
[44] http://www.ddialliance.org/
[45] http://dublincore.org/
[46] http://www.clarin.eu/content/component-metadata

Language Observatory (VLO) that serves as the CLARIN metadata portal. CMDI is not a single metadata schema but rather specifies a way how different metadata schemas can be constructed using reusable schema components and made interoperable using partly a mandatory syntactic structure and proscribing the use of a semantic registry for describing the schema elements for semantic interoperability.

DataCite
DataCite[47] is are a not-for-profit organization with members from several data centers around the world. DataCite provides a registry with persistent identifiers for datasets. DataCite manages basic metadata related to the datasets. Their metadata format[48] has strong relations to DC terms[49] and DDI.

TEI – Text Encoding Initiative
A widely used format in the (text-oriented branch of the) DARIAH community is the Text Encoding Initiative[50] (TEI), a "standard for digital representation of texts". For the purpose of encoding metadata the standard provides the structured element teiHeader that describes various aspects of both the source material as well as the digitized text, such as an extended bibliographic record (author, title, publication date and place, publisher, etc.), encoding decisions or administrative record of changes.

Note though, that TEI does not prescribe one fixed schema, but rather a complex set of elements that individual projects could use to create their own profile best suiting the project's needs. While this is valuable for the researchers and certainly one reason for the widespread use of TEI, it complicates the task of an aggregator, as it is not possible to reliably identify individual elements by means of fixed XPaths across different providers. This teiHeader is a similar approach to the CLARIN CMDI set of schemas and a similar solution to handling the semantic interoperability issues would be possible if the semantics of teiHeader would be explicit in a machine readable way.

DCLAP
The Collection Registry developed within DARIAH-DE uses a Dublin Core application profile[51]; an extension of the basic Dublin Core set of elements to better support the description of collections (providing additional fields as itemType, itemEncodingScheme, or accumulationDateRange).

EDM – Europeana Data Model
Among the digital heritage institutions, the Europeana Data Model[52] (EDM) is probably the first candidate given that it has been used for describing more than 30 million objects collected through the extensive aggregation network of

[47] http://www.datacite.org/
[48] http://schema.datacite.org/
[49] http://dublincore.org/documents/dcmi-terms/
[50] http://www.tei-c.org/
[51] http://schema.dariah.eu/colreg/dclap/dclap.xsd
[52] http://pro.europeana.eu/edm-documentation

Europeana. EDM is a successor for the Europeana Semantic Elements (ESE), a schema which was first used in the Europeana network and was basically an extension of Dublin Core. As opposed to ESE, EDM is strongly rooted in the Semantic Web.

CIDOC-CRM
The main reference point in the archaeological community is the Conceptual Reference Model developed by the International Committee for Documentation (CIDOC), the CIDOC-CRM that is used as conceptual grounding for a number of derived profiles or schemas[53], ensuring their interoperability.

Other formats used in archaeological context
In the archaeological community, that is one of the more advanced Humanities areas with respect to metadata, a number of formats have been encountered in the harvested data, e.g. variants of the schema suite Geographic Metadata Schema (GMD)[54], the TEI-based schema ENRICH[55].
Within the research infrastructure project ARIADNE, the partners mentioned – besides CIDOC-CRM – also using DDI, DataCite, MARC/UNIMARC, TriDAS, INSPIRE[56], ISO 11915, CARARE[57] and LIDO. The basis for the planned Metadata Registry and the ARIADNE Catalog Data Model (ACDM) is the Data Catalog Vocabulary standard (DCAT), together with the standard ISO/IEC 11179 on Metadata Registries. See also the deliverable of the project ARIADNE on metadata standards relevant in the discipline: "D3.1 Initial report on standards and on the project registry"[58]

13.3. The Metadata Catalogue Software and Workflow

As a choice for the metadata catalogue software, the MPI-PL partner has implemented a metadata catalogue based on the open source CKAN[59] software from the Open Knowledge Foundation. Although CKAN also permits the storage of data resources, the DASISH JMD uses it only to aggregate and present metadata from different metadata providers from the SSH infrastructures. In line with the DASISH DoW the metadata catalog functions as a faceted browser.

The rationale for this choice was based on the need for open software with a broad user base requiring limited configuration and adaptation. Some participating communities already have their own metadata catalogues (CESSDA Data Portal[60] and the CLARIN VLO[61]). An option was to collaborate and use their software stack. However the need to cater for metadata with

[53] http://www.cidoc-crm.org/
[54] http://www.isotc211.org/schemas/2005/gmd/
[55] http://projects.oucs.ox.ac.uk/ENRICH/Deliverables/referenceManual_en.html
[56] http://inspire.ec.europa.eu/
[57] http://www.carare.eu
[58] http://www.ariadne-infrastructure.eu/Media/Files/D3.1-Initial-Report-on-the-project-registry
[59] http://ckan.org/
[60] The CESSDA Data Portal, http://www.cessda.net/catalogue/, http://www.cessda.net/catalogue/
[61] The CLARIN Virtual Language Observatory, http://www.clarin.eu/content/virtual-language-observatory

very diverse metadata schemas, especially from the DARIAH community, pointed us in the direction of the EU EUDAT project, a project that also needed to create an interdisciplinary metadata catalogue: B2FIND. The latter includes a semantic mapping module for mapping the different schemas on a set of shown facets. This mapping module was open and made available through the MPI-PL partner. All the extensions of CKAN and the mapping and normalization module software are available via GitHub, which is a public domain software code repository. The metadata processing workflow comprises four stages as illustrated in Figure 16: The Workflow for filling the Metadata Catalogue. First fetching of original metadata records with OAI- PMH harvesting from the OAI Metadata Providers. Subsequently performing semantic mappings into our internal representation of facets, and following this a series of substitutions on the element values takes place for normalization purposes. The last step is ingestion of metadata into CKAN. The relevant scripts can be found at the DASISH github[62] repository.

The CKAN extensions and workflow software developed for the DASISH metadata catalogue are:

1. UI Modifications
2. Metadata Mapping Module
3. Automatic CMDI Mappings Generation
4. Addressing some CKAN performance issues, see Appendix J: CKAN Performance testing

[62] https://github.com/DASISH/jmd-scripts

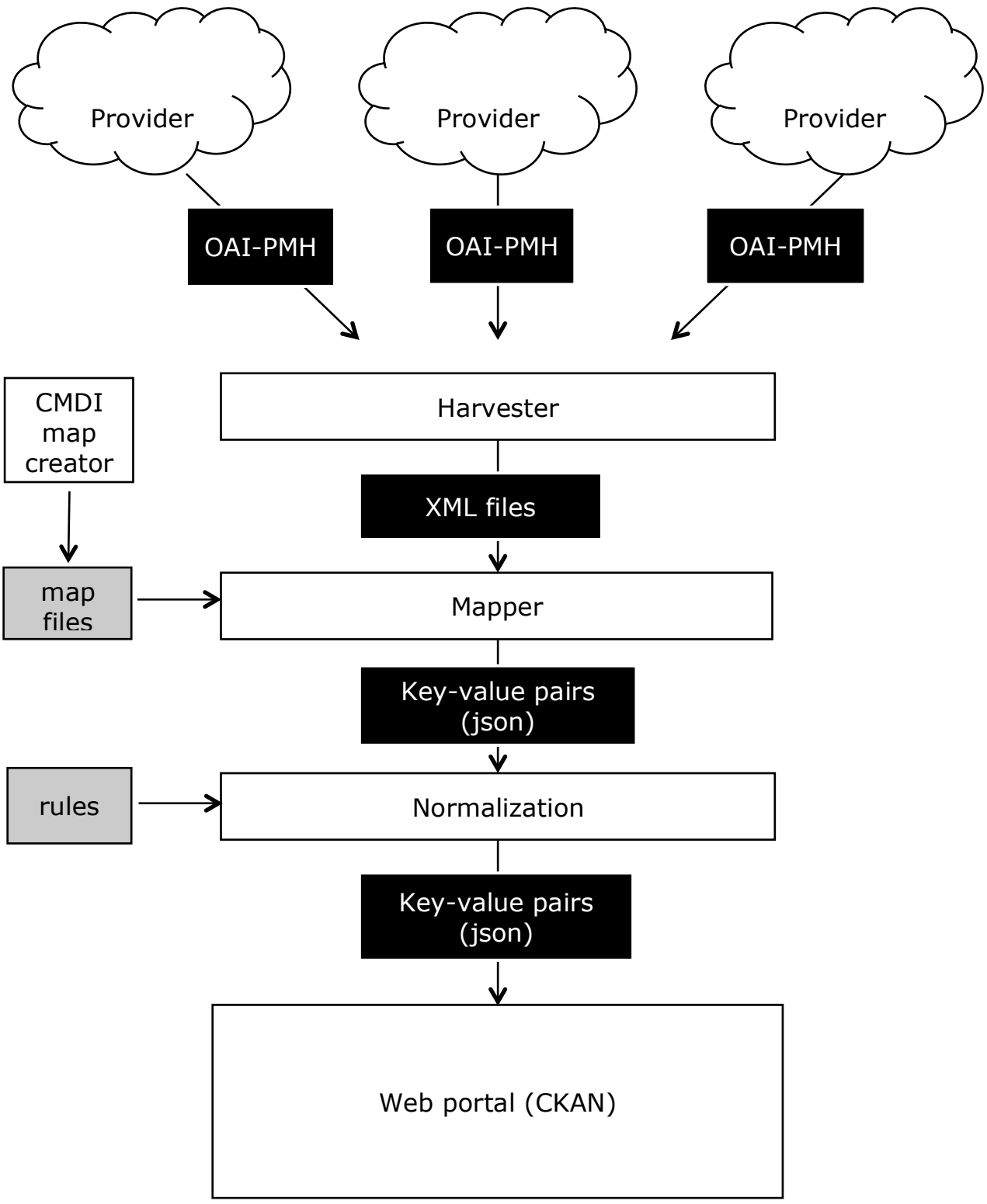


Figure 16: The Workflow for filling the Metadata Catalogue

13.3.1 UI Modifications

With respect to the user interface (UI), we modified CKAN for two reasons: a) to enable faceted search, and b) to enable at the homepage a slideshow of the communities at the start page.

At the top of every page, the default user interface has links for Datasets, Groups and Organizations.

Datasets - in CKAN, data are organized in units called "datasets". A dataset is a parcel of data. It has several attributes (name, title, author, origin, language, URL, and so on). When users search for data, the search results they see are individual datasets. A dataset may only have metadata: for example, the title and publisher, date, formats the data is available in, what license it is released under, etc. A dataset may also have resources (data itself). CKAN can store the resource internally, or store it simply as a link.

Groups - in CKAN, groups are ways to put together datasets under a community (for example, CLARIN) or topic (e.g. Linguistics, Health, Environment) to make it easier for users to browse datasets by theme. Datasets may belong to more than one group.

Organizations - in CKAN, organizations act like publishing departments for datasets (for example, the Department of Health). Within organizations, administrators can assign roles and authorization to its members to publish or delete datasets.

Modifications:

1. The feature of "Organizations" in CKAN is not used in our DASISH metadata catalogue. It is assumed that there is only one organization that can publish and delete datasets (i.e. the administrator of DASISH metadata catalogue).

2. We wrote a CKAN extension to display the facets on the left sidebar so that the user can perform faceted search. The user can click on one of the facet values and CKAN will list all datasets that have a selected facet with that value. The extensions are written in Python and can be found at the DASISH github[63].

3. On the catalogue home page, we display all the communities in a cyclic slideshow with the name, logo and description of the group.

13.3.2 Metadata Mapping Module

The metadata-mapping module (also called the mapper) is external to CKAN.

[63] https://github.com/DASISH/ckanext-dasish

It is the software tool developed for the purpose of mapping xml files to key-value pairs (JSON format) suitable for importing into CKAN. The keys are the facets (e.g. creator, title) or internal data attributes (e.g. full text and the mapper version used). The software is written in JAVA and can be found at the DASISH github[64].

The mapper has two components: the map files and the mapping engine. The mappings are specified as XPath expressions. The XPath expressions define semantic mappings to convert from community-specific metadata schemas into the internal schema of the DASISH metadata catalogue. Saxon is used as the XPath engine, but only via standard APIs.

The mapping engine is the software that actually performs the semantic mapping between the harvested XML and the internal schema of the catalogue. The mapping engine takes as input the XML metadata stored as files in a directory and the mapping file (one per community) and outputs a JSON file (key-value pairs). This is performed in a streaming fashion.

13.3.3 CMDI Mapping Generator

A special module was developed for processing CMDI metadata harvested from the CLARIN infrastructure. CMDI is not a single schema but rather an open set of schema with provisions for semantic interoperability. Every element in a CMDI compatible schema is adorned with an attribute whose value refers to a concept in the CLARIN concept registry (ISOcat). These references are then used to create specific mappings between metadata elements and facets. To address the large variety of CMDI schemas we developed an additional software module that allows automatic generation of mapping specifications on the basis of harvested CMDI metadata[65]. (See 13.3.3 CMDI Mapping Generator).

13.3.4 CKAN Performance Issues

How much data can CKAN handle? This was the question we tried to answer before using CKAN. With our tests, we found that CKAN in its default configuration performs adequately with ten thousands of datasets (records) but that with millions of records, it becomes too slow. However after some performance tuning measures, CKAN can handle several millions of datasets (we managed to import about 2 million datasets in less than 2 weeks). The performance tuning measures concern:

- changing CKAN configuration file
- changing designs in the PostgreSQL database tables and
- changing a few PostgreSQL (postgresql.conf) configurations to take advantage of available memory and CPU.

[64] [code] https://github.com/TheLanguageArchive/md-mapper, [mapfiles] https://github.com/DASISH/md-mapping
[65] https://github.com/DASISH/jmd-scripts/tree/master/util-scripts

Details of the performance tuning measures and their impacts on performance are given in Appendix J: CKAN Performance tuning.

13.4. Facets for the DASISH Catalogue

The advantage of a single specific set of facets of the DASISH JMD was that it should make it possible for all SSH infrastructures users to explore the available metadata using understandable terminology relevant to their domain. From the beginning of the facet discussions, its design could only be the least common denominator. Since there have not been enough resources in the task to collect data and opinions from every DASISH institute on such a set of facets, a pragmatic development solution was to have a small expert group put together, with more people than the task members, to ensure that all communities (CLARIN, DARIAH, CESSDA) could play a part in the discussion and share their expertise. The DASISH JMD Set of Facets was created during this collaboration. It is based on the status quo of well-established metadata policies used in the communities and is not so much a new development. During the work on the mappings and the metadata quality improvements, the set of facets was slightly modified. The definitions of the facets and of overlapping metadata fields can be found in Appendix H: List of facets with Definitions.

13.5. Mapping Metadata to Facets and Fields

We distinguish between facets and fields when mapping metadata information from the harvested metadata records into the catalogue. Facets provide the browsable dimensions; they are classifiers whose values are shared by many records such as 'Country', 'Creator', 'Language' etc. While fields are metadata elements whose values are often unique such as 'Name' and 'Title'.

The mappings and presentation of the metadata is not meant to result in an absolutely truthful representation of the original metadata, but rather to increase the usefulness of the DASISH metadata catalogue. Countering the sparseness of the visualized metadata is an important issue here. For instance metadata records are usually identified by the value of the title element. If however, there is no title available for a specific metadata record, the value of another 'identifier data type' can be presented to the user as an identification of the record.

For example, when mapping the 'title' facet from DDI 3.1, the first XPath expression tries to find if there is any 'title' marked as English available: *[s:StudyUnit/r:Citation/r:Title[@xml:lang='en'].* If this does not give any result, the mapper will continue with the next XPath expression without the language filter: *[s:StudyUnit/r:Citation/r:Title]*

This type of test in a prioritized order is performed for several facet values and

can be reviewed in detail in the mapping file for each format[66]. For the complete set of mappings used for the DASISH catalogue see Appendix I: List of Mappings.

In the DoW a task was defined registering the concepts from the SSH metadata schemas in a semantic registry as the ISOcat[67] that is used by CLARIN. Due to the large number and not always well documented schemas we decided to limit this effort to registering only the concepts used in the DASISH metadata catalogue itself which should be sufficient for the purpose of documenting the DASISH catalogue itself.

13.6. Normalization

Two tasks are performed by the normalization module implemented as a post processing script. Firstly, it changes various date formats to UTC format (YYYY-MM-DDThh:mmTZD). Secondly, it can substitute fixed input strings to other fixed output strings and can thus be used for instance to replace language code by language names.

Examples of replacement rules are:

Facet	**Input**	**Ouput**
Language	nl	Dutch; Flemish
	nld	Dutch; Flemish
Subject		
	HISTORIA	History
	Historia	History
	Sammlung, Münzsammlung, Numismatik	collection, coin collection, numismatic
Country	nl	Netherlands
	nld	Netherlands

See for further information Appendix L: Normalization.

[66] https://github.com/DASISH/md-mapping/tree/master/mapfiles
[67] http://www.clarin.eu/faq-page/266

14. Metadata Quality Improvement

In the DASISH task 5.3, a report on metadata quality improvement was created which included an evaluation of the preliminary results in DASISH the task 5.4 based on the criteria used in the report. An early DASISH JMD prototype published on April 23 2014 was the basis of this evaluation[68]. As examples for improving metadata, some facets of the prototype catalogue were analyzed: "CreationDate", "Creator", "Language" and "Discipline". The suggestions are summarized in chapter 5.1. The implementation is described in chapter 5.2.

14.1. Suggestions on Metadata Improvement from Task 5.3

In the deliverable of the DASISH Task 5.3, problems with some facets were pointed out. These problems should be regarded as examples of a major issue encountered across a number of facets and occurring with many metadata catalogues – the high variation in values that actually "mean the same" but are represented by different strings, either due to spelling variants or through the use of synonyms. This problem is encountered in all facets or metadata fields that expect the values to come from a controlled list of possible values, where however this list cannot be exhaustively defined (open controlled vocabularies). Typical examples of such facets are group/organization, resource type, subject/genre, rights/licensing, etc.

The following fields were specifically mentioned as being problematic:

CreationDate
Since the prototype mapped the metadata onto a simple format no differentiation between different creation date values was conducted. The report describes on the basis of examples that collecting different date values like the creation date of the original resource or the creation date of the metadata in the same metadata field would blur the information for the user. To increase the conformance to user expectations it was suggested to use separate date fields for different date values.

Creator
The values of the catalogues' 'Creator' facet do not always correlate with real researchers. "Not applicable" was the most frequent value. Therefore the suggestion was to create a facet 'Researcher Identifier' in the future. Alternatively, instead of using a 'Creator' facet, use the metadata harvesting origin as a facet. This information could be a sub element of the group facet.

Language
The report states that language at first seems to be easy to normalize. But the given examples show the diversity of used values. The used values for German are at least "de" (ISO 639-1), "ger" (ISO 639-2), "deu" (ISO 639-2) and

[68] Latest version can be found at: http://ckan.dasish.eu/ckan

"German". Something similar applies to English and Dutch.

Discipline
The facet 'Discipline' as well as the facet 'Subject' were highly diverse and on varying conceptual levels. The benefit of the facets in such a status was considered dubious for the user. The suggestion was to use a thesaurus or classification schemes then either make them available for the user or to remove the facets from the catalogue.

14.2. Improving the Catalogue

There is no easy remedy to the problems stated above, but there are approaches to mitigate them. The cornerstone is the extensive use of controlled vocabularies collaboratively built and shared within the community or better shared across communities. These ideally can (and should) be applied during the metadata authoring step, but can also be used for normalization in a curation step after harvesting at the side of the aggregator.

Data sparseness (e.g. as can appear often with the 'Name' or 'Creator' of the resource), can be addressed by making "if-then-else" logical constructions. This tests the availability of that information in a suitable metadata element and if not finding it, test the availability in the next best element. For instance if 'Creator' has no value filled in, next best elements can be 'Project' or 'Organization'. Although not the best semantic practice, it does give a better user experience as explained also in Facets for the DASISH Catalogue.

The use of the Normalization module can be used to overcome differences in controlled vocabularies. Where one community uses Language Code (ISO 369 or other) to fill a 'Language' metadata element, others may use language names. The Normalization module allows normalizing this. This can be used for 'Country' name values, disciplines etc. When broadly accepted vocabularies are missing, the metadata catalogue should provide its own vocabulary but this may lead to loss of generality.

We followed the suggestion to provide a 'data provider' facet. That omission in the early prototype was also noted when using the catalogue for validating the contributions from the different OAI-PMH end-points. Furthermore we introduced a further date facet. In addition to "Temporal Coverage" and "Creation Date" we added a "Publication Date" facet.

We feel that, however inadequate, the catalogue needs a 'Discipline' facet since the DASISH catalogue is an interdisciplinary catalogue and needs this information. A normalized discipline vocabulary would have been useful, and is being considered for the EUDAT B2FIND catalogue. The same reasoning holds true for the 'Subject' facet, which we consider important to identify data for specific research interests.

15. Findings

We consider the experience of finding and documenting the OAI endpoints, the inventory of metadata schema and the developed mapping and normalization rules as an important outcome of the project. This includes the work on improving the metadata quality. The catalogue software itself does have its problems with respect to performance, but should be replaceable by some other software that is also based on key/value pairs of facet information.

We encountered a large disparity in the handling of metadata across the research infrastructures. While CLARIN implements a highly integrated system periodically harvesting the metadata from all content providers, in DARIAH there is no strict policy on collecting metadata and the landscape shows the use of a plethora of metadata formats even within individual subcommunities (like archaeology). Thus, the collection of OAI endpoints identified within this task together with a preliminary screening of the content of the corresponding repositories may be the most comprehensive information on metadata in DARIAH available and would certainly be of interest to be offered to the DARIAH community and its decision makers.

With respect to the CKAN software used to implement the DASISH catalogue, we can state that when dealing with millions of records, CKAN appears to be too slow with importing the JSON formatted key/value pairs, resulting in several days of processing time. We think this performance problem can be traced back to CKAN using a database for storing some of the record's information instead of relying completely on indexes as some other (better performing) catalogue implementation do.

With respect to the actual collected metadata we observed a number of substantial problems that we already touched upon earlier:

- sparseness of data, i.e. missing values and including non-informative place-holders like "Unspecified"
- the enormous variation in spelling and formats of the values.

A partial remedy is the normalization effort in the workflow on the side of the catalogue. This allows, for example, to normalize the date formats, or map selected high-frequency values to a common label. However this approach works only that far. For semantically more challenging fields, like 'Subject' or 'Discipline' it is next to impossible for the maintainer of the catalogue to identify synonyms (or even hype- and hyponyms) in the vast extraordinarily disparate lists of categories. The resolution here can only be the collaborative work by the communities on shared reference resources (controlled vocabularies and taxonomies) that can be used as normative sources for values when authoring metadata.

Another challenge is the semantic interpretation of certain metadata elements, especially in the case of too-coarsely grained concepts like DC's 'Date'. Without further information it cannot be determined, if the creation or the publication

date is meant, or perhaps the temporal coverage of the described resource.

Our final conclusion is that this task was a worthwhile exercise covering all aspects of gathering and interpreting metadata from CESSDA, CLARIN and DARIAH. It was especially illuminating to be able to compare the different approaches and levels of knowledge and integration that exist in the three infrastructures with respect to metadata use and exchange. Fine tuning and polishing the mappings and normalizations used in the DASISH JMD is, we think, a matter of sustained work and we hope the suggested follow-up (See Future of the DASISH Catalogue) can profit from our work.

16. Future of the DASISH Catalogue

In addition to its relevance for the common SSH to learn and exchange information about the different metadata practices in the SSH, it would be advantageous to keep the DASISH metadata catalogue available after the DASISH project ends. This requires a project or organization that has an interest in such a service and is able to perform the necessary configuration and maintenance.

At the moment it is not yet clear whether there will be a follow-up SSH cluster project, and if so, whether it would cover the same communities as is the case for DASISH. For this reason and also because of the shared technology approach, we decided to approach the EUDAT project that runs a broad interdisciplinary metadata catalogue B2FIND. Although we cannot expect that a broad interdisciplinary catalogue can be created as optimal as for a specific discipline or cluster, the scale advantages are clear.

References

Atlas of Living Australia (ATLAS). (2011). *Guide to data quality* (Version 1.2). Nicholls, M. Retrieved from http://www.ala.org.au/about-the-atlas/how-we-integrate-data/data-quality-assurance/

Balkan, L., Miller, M., Austin, B., Etheridge, A., Bernabe, M. G., & Miller, P. (2002). ELSST: A broad-based multilingual thesaurus for the social sciences. *LREC 2002 Third International Conference on Language Resources and Evalutation,* Las Palmas. 1873-1877. Retrieved from http://www.lrec-conf.org/proceedings/lrec2002/

Bargmeyer, B., & Gillman, D. (2000). *Metadata standards and metadata registries: An overview.* Retrieved from http://stats.bls.gov/ore/pdf/st000010.pdf

Barker, E., & Ryan, B. (2003). *Case studies in implementing educational metadata standards: The higher level skills for industry repository*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.4993&rep=rep1&type=pdf

Barton, J., Currier, S., & Hey, J. M. N. (2003). Building quality assurance into metadata creation: An analysis based on the learning objects and EPrints communities of practice. Paper presented at the *Proceedings 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications,* Seattle. Retrieved from http://eprints.erpanet.org/83/01/paper60.pdf

Beckett, R. C., & Hyland, P. (2011). Communication and learning structures that facilitate transfer of knowledge at innovation transition points. *JCOM: Journal of Science Communication, 10*(4). Retrieved from http://jcom.sissa.it/archive/10/04/Jcom1004(2011)A03/

Berners-Lee, T., Bizer, C., & Heath, T. (2009). Linked data--the story so far. *International Journal on Semantic Web and Information Systems, 5*(3). DOI: 10.4018/jswis.2009081901

Broeder, D., Declerck, T., Romary, L., Uneson, M., Stromqvist, S. & Wittenburg, P. (2006). *A large metadata domain of language resources*. Retrieved from http://gandalf.aksis.uib.no/non/lrec2004/pdf/478.pdf

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P. & Wittenburg, C. (2010). *A data category registry- and component-based metadata framework.* Retrieved from www.windhouwer.nl/menzo/professional/papers/metaData.pdf

Bruce, T.R., & Hillmann, D.I. (2004). The Continuum of metadata quality: defining, expressing, exploiting. *Metadata in Practice,* American Library Association, Chicago. Retrieved from http://www.ecommons.cornell.edu/handle/1813/7895

Campbell, L. (2007). Learning object metadata. In S. Ross & M. Day (Eds.),

DCC: Digital curation manual. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/learning-object-metadata/learning-object-metadata.pdf

Caplan, P. (2006). Preservation metadata. In S. Ross & M. Day (Eds.), *DCC: Digital curation manual.* Retrieved from http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata

Carpenter, L. (2003). *OAI for beginners: The open archives forum online tutorial*. Retrieved from http://www.oaforum.org/tutorial/english/intro.htm

CLARIN-ERIC. (2013). *Component metadata*. Retrieved from http://www.clarin.eu/node/3219

Currier, S., Barton, J., O'Beirne, R., & Ryan, B. (2004). Quality assurance for digital learning object repositories: Issues for the metadata creation process. *ALT-J, Research in Learning Technology, 12*(1), 5-20. doi: 10.1080/0968776042000211494

Data Documentation Initiative (DDI). (2009). *Best practices across the data life cycle.* Retrieved from http://www.ddialliance.org/resources/publications/working/BestPractices/DataLifeCycle

Data Without Boundaries (DwB). (2013). *Final report proposing portal resource discovery functionality for a search/browse portal interface, Improving Resource Discovery for OS Data.* (Deliverable D8.4, Work Package 8). NSD, CIS, Destatis, MT, UEssex, KNAW-DANS, CNPS-INS.

Day, M. (2005). Metadata. In S. Ross & M. Day (Eds.), *DCC: Digital curation manual.* Retrieved from http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata

DCMI. (2013). *Dublin core metadata initiative*. Retrieved from http://dublincore.org

DDI Alliance. (2009). *DDI alliance.* Retrieved from http://www.ddialliance.org/alliance

Digital Library Federation (DLF). (2007). *Best practices for shareable metadata.* Retrieved from http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/ShareableMetadataPublic#Best_Practices_for_Shareable_Metadata

Digital Library Federation (DLF). (2010). *<METS> metadata encoding and transmission standard: Primer and reference manual: version 1.6 revised*. Retrieved from http://www.loc.gov/standards/mets/METSPrimerRevised.pdf

Digital Preservation Coalition (DPC). (2004). *The open archival information system reference model: Introductory guide* (DPC Technology Watch Series Report 04-01). London: Lavoie, B.F. Retrieved from http://www.dpconline.org/docs/lavoie_OAIS.pdf

Doctorow, C. (2001). *Metacrap: Putting the torch to seven straw-men of meta-utopia*. Retrieved from http://www.well.com/~doctorow/metacrap.htm

Dunlap, R., Mark, L., Rugaber, S., Balaji, V., Chastang, J., Cinquini, L., Murphy, S. (2008). Earth system curator: Metadata infrastructure for climate modeling. *Earth Science Informatics, 1*(3-4), 131-149. doi: 10.1007/s12145-008-0016-1

Economic and Social Research Council. (2013). *Research data policy*. Retrieved from http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx

EDINA and Data Library, University of Edinburgh. (n.d.). *Research data MANTRA* [online course]. Retrieved from http://datalib.edina.ac.uk/mantra

Edwards, P., Mayernik, M., Batcheller, A., Bowker, G., & Borgman, C. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science, 41*(5), 667-690. doi: 10.1177/0306312711413314

Eppler, M. J. (2008). Knowledge communication. In M.E. Jennex (Ed.), *Knowledge management: Concepts, methodologies, tools, and applications* (pp. 324-335). IGI Global. doi: 10.4018/978-1-59904-933-5.ch030

GBIF. (2011). *Darwin core archive assistant user guide: Version 1.1.* Retrieved from http://tools.gbif.org/dwca-assistant/

Granda, P., Kramer, S., Linnerud, J., Marker, H. J., Miller, K. & Vardigan, M. (2009). *DDI working paper series: Best practices, no. 5.* K. Miller (Ed.). Retrieved from http://www.ddialliance.org/sites/default/files/bp/DDIBestPractices_ControlledVocabularies.doc.pdf

Grootveld, M., van Egmond, J., & Sørensen, B. (Eds.). (2011). *Data reviews, peer-reviewed research data*. DANS studies in Digital Archiving 5. The Hague: Data Archiving and Networked Services (DANS).

Guy, M., Powell, A., & Day, M. (2004). Improving the quality of metadata in EPrint archives. *Ariadne,* (38). Retrieved from http://www.ariadne.ac.uk/issue38/guy

Halevy, A., Franklin, M., & Maier, D. (2006). Principles of dataspace systems. *Proceedings of the Twenty-Fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,* Chicago. http://dx.doi.org.myaccess.library.utoronto.ca/10.1145/1142351.1142352

Hillmann, D., Dushay, N., & Phipps, J. (2004). Improving Metadata Quality: Augmentation and Recombination. *DC-2004—ShanghaiProceedings.* Retrieved from http://dcpapers.dublincore.org/pubs/article/view/770

Hillmann, D.I, Metadata Quality: From Evaluation to Augmentation (2007) Retrived from http://www.ecommons.cornell.edu/bitstream/1813/7899/1/Metadata_Qu

ality_rev.pdf

Holdsworth, D. (2007). Preservation strategies for digital libraries. In S. Ross & M. Day (Eds.), *DCC digital curation manual.* Retrieved from http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/preservation-strategies

Humphrey, C. (2006). *E-science and the life cycle of research.* Retrieved from http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

International Federation of Library Associations and Institutions (IFLA) (1997) *Functional requirements for bibliographic records Final report.* Retrieved from http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

International Organization for Standardization (ISO). (2009). *ISO 15836:2009 information and documentation--the dublin core metadata element set*. Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52142

International Organization for Standardization (ISO). (2010). *International standards and "private standards".* Geneva: ISO.

International Organization for Standardization (ISO). (2013). *ISO 3103:1980 tea--preparation of liquor for use in sensory tests*. Retrieved from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=8250

JHOVE: JSTOR/Harvard object validation environment. (2009). Retrieved from http://jhove.sourceforge.net/index.html

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *IEEE Internet Computing, 5*(5), 59-68. doi: 10.1109/4236.957896

Jones, S. (2011). *How to develop a data management and sharing plan.* Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How%20to%20Develop.pdf

Library of Congress. (2008). *MIX- NISO metadata for images in XML schema*. Retrieved from http://www.loc.gov/standards/mix/

Library of Congress. (2013). *TextMD: Technical metadata for text*. Retrieved from http://www.loc.gov/standards/textMD/

Liyanage, C., Elhag, T., Ballal, T., & Li, Q. (2009). Knowledge communication and translation – a knowledge transfer model. *Journal of Knowledge Management, 13*(3), 118-131. doi: 10.1108/13673270910962914

Miller, K., & Vardigan, M. (2005). How initiative benefits the research community - the data documentation initiative. *First International Conference on e-Social Science,* Manchester. Retrieved from http://www.ddialliance.org/sites/default/files/miller.pdf

Mize, J., & Robertson, C. F. (2009). A solution to metadata: Using XML transformations to automate metadata. *OCEANS 2009, MTS/IEEE Biloxi - Marine Technology for our Future: Global and Local Challenges,* Biloxi, MS. Retrieved from http://ieeexplore.ieee.org.myaccess.library.utoronto.ca/stamp/stamp.jsp?tp=&arnumber=5422136&isnumber=5422059

Mohler, P. P., Hansen, S. E., Pennell, B., Thomas, W., Wackerow, J., & Hubbard, F. (2010). A survey process quality perspective on documentation. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts,* (pp. 299-314). Retrieved from Wiley Online Library. doi: 10.1002/9780470609927.ch16

National Information Standards Organization (NISO). (2004). *Understanding metadata.* Bethesda, MD. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf

National Information Standards Organization (NISO). (2007). *A framework of guidance for building good digital collections* (3rd ed.). Baltimore, MD: NISO Framework Working Group. Retrieved from http://www.niso.org/publications/rp/framework3.pdf

National Science Foundation (NSF). (2012). Proposal preparation instructions. In *Proposal and award policies and procedures guide: Part I- grant proposal guide,* (pp. II1-II40). Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpgprint.pdf

Nelson, B. (2009). Empty archives. *Nature, 461*(10), 160-163. Retrieved from http://www.nature.com.myaccess.library.utoronto.ca/news/2009/090909/pdf/461160a.pdf

Open Archives Initiative. (n.d. a). *Open archives initiative*. Retrieved from http://www.openarchives.org

Palavitsinis, N., Manouselis, N., Sanchez-Alonso, S (2014). Metadata Quality in Digital Repositories: Emperical Results from the Cross-Domain Transfer of a Quality Assurance Process. *Journal of the association for information scinece and Technology*

Park, J. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging and Classification Quarterly, 47*(3-4), 213-228. doi: 10.1080/01639370902737240

Preservation Metadata: Implementation Strategies (PREMIS). (2012). *PREMIS data dictionary for preservation metadata* (Version 2.2). PREMIS Editorial Committee. Retrieved from http://www.loc.gov.myaccess.library.utoronto.ca/standards/premis/v2/premis-2-2.pdf

Research Data Management Team. (2012). *Create and manage data: Training resources*. Retrieved from http://data-archive.ac.uk/create-manage/training-resources

Research Information Network. (RIN). (2008). *Stewardship of research data: Principles and guidelines: Responsibilities of research institutions and funders, data managers, learned societies and publishers*. Retrieved from http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines

Riley, J. & Becker, D. (2010). *Seeing standards: A visualization of the metadata universe*

Schweitzer, P. (2012). *Frequently asked questions on FGDC metadata.* Retrieved from http://geology.usgs.gov/tools/metadata/tools/doc/faq.html - motivation

Shankaranarayanan, G., & Even, A. (2006). The metadata enigma. *Communications of the ACM, 49*(2), 88-94. Retrieved from http://search.ebscohost.com.myaccess.library.utoronto.ca/login.aspx?direct=true&db=buh&AN=19568343&site=ehost-live

Shreeves, S.L., Knutson, E.M., Stvillia, B., Palmer, C.L., Twidale, M.B., & Cole, T.W. (2005), *Is "Quelaity" Metadata "Shareable" Metadata? The implications of Local Metadata Practices for Federated Collections*. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/145/shreeves05.pdf?sequence=2

Stvilia, B., Gasser, L., Twidale, M. B., Shreeves, S. L. & Cole, T. W. (2004). *Metadata quality for federated collections*. Retrieved from http://mailer.fsu.edu/~bstvilia/papers/iciq_144_final_v1.pdf

TC46 SC 11 Interest Group. (2011). *Where to start: Advice on creating a metadata schema or application profile.* (No. N800R1).NISO. Retrieved from http://www.niso.org/apps/group_public/document.php?document_id=7272&wg_abbrev=tc46sc11interest

TEI Consortium. (n.d.). *TEI: Text encoding initiative*. Retrieved from http://www.tei-c.org/index.xml

Trippel, T., Broeder, D., Durco, M.,& Ohren, G. (2014). Towards automatic quality assessment of component metadata*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14,* Iceland. 3852-3856 Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/1011_Paper.pdf

UK Data Archive. (2011). *Managing and sharing data* (3rd ed.). Essex: Van den Eynden, V., Corti, L., & Horton, L. Retrieved from http://data-archive.ac.uk/media/2894/managingsharing.pdf

UK Data Service. (2013). *Support for depositors*. Retrieved from http://ukdataservice.ac.uk/deposit-data/support.aspx

UKOLN. (2007). *Dealing with data: Roles, rights, responsibilities and relationships consultancy report* (V1.0). Lyon, L. Retrieved from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

University Library, University of Illinois at Urbana-Champaign. (2010*). Best practices for creating*

digital collections. Retrieved from http://www.library.illinois.edu/dcc/bestpractices/contents.html

University of Virginia Library Data Management Consulting Group. (2013). *Steps in the research life cycle*. Retrieved from http://dmconsult.library.virginia.edu/lifecycle/

Van Uytvanck, D. (2010). *CLARIN Short Guide: Virtual collections*. Retrieved from http://www.clarin.eu/sites/default/files/virtual_collections-CLARIN-ShortGuide.pdf

Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation, 3*(1), 107-113. http://dx.doi.org/10.2218/ijdc.v3i1.45

Wayne, L. (2005). *Institutionalize metadata before it institutionalizes you.* Retrieved from http://www.fgdc.gov/metadata/metadata-publications-list

Wilson, A. J. (2007). Toward releasing the metadata bottleneck: A baseline evaluation of contributor-supplied metadata. *Library Resources & Technical Services, 51*(1), 16-28. Retrieved from http://alcts.metapress.com/content/q162387hu6237078/fulltext.pdf

World Wide Web Consortium (W3C). (2012). *SKOS Simple Knowledge Organization System—Home Page.* Retrieved from http://www.w3.org/2004/02/skos/

Yale University Library. (2008). *Best practices for structural metadata*. Retrieved from http://www.library.yale.edu/dpip/bestpractices/BestPracticesForStructuralMetadata.pdf

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries, 1*(2), 5-16. doi: 10.1007/s00799-007-0015-8

Glossary

Actor - An individual or organisation, which is responsible for fulfilling a role. The term actor is commonly used within business process analysis, often presented in Unified Modelling Language (UML) or BPMN (Business Process Modelling Notation) diagrams.

Agent - An alternate term for Actor, within the PREMIS preservation metadata schema an 'Agent' may refer to either a human or software agent. Here the term 'software agent' will be used to describe situations where the creation or validation of metadata is undertaken by machine or where a role is machine actionable in some other way.

Administrative Metadata - Provides information that helps with the management of data. Some examples of administrative metadata are descriptions of the intellectual property rights of an object, how to access an object, or any changes that have been made to data.

Archive System - LTDP (Long Term Digital Preservation) Archives are a special case of the Curation System with some additional responsibilities (a 'mission') necessary to ensure the continued usability of data/metadata over a long period of time, but certainly beyond the next round of technological change. While an LTDP Archive may take specific 'preservation actions' or have additional measures for 'Archival Storage' the majority of activities including maintaining a record of actions undertaken, file format migration and integrity measures may also be applied in environments without LTDP responsibility.

CKAN - an open-source data management system and a data portal platform. It allows for simple publishing, sharing, finding and using data. CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available. http://ckan.org/

CMDI - Component Metadata Infrastructure
A flexible framework for defining metadata schemas and a distributed system for creation and publication of metadata records developed within CLARIN. http://clarin.eu/cmdi

Collection - set of resources grouped together according to some principle. Sometimes recursive collections are encountered (similar to directories in the file system).

Content Provider, Data Provider - institution offering access to resources. Usually this involves an institutional content repository, where the resources are stored together with corresponding metadata records. At least the metadata has to be publicly available (ideally harvestable) to enable resource discovery. The access to the actual resources can be restricted, however the licensing terms and active contact to responsible persons needs to be indicated in the metadata.

Contextual Metadata - Provides information about data that helps a researcher to interpret and use data correctly. This information might include what instruments were used to collect data, what questions were asked on a survey, or when data was collected.

Crosswalk -The process of mapping the content of one metadata scheme to another in order to facilitate interoperability between data repositories.

Curation System - Any system, which is used to capture, store, manage or disseminate data/metadata before the Archival phase of the lifecycle. A curation system could be anything from a researcher's hard drive to a full institutional repository or government department data/metadata management system.

Data Documentation Initiative (DDI) - An international standard expressed in XML used to describe social science data sets.

Data Management Plan - A written plan that a researcher makes in the first stages of a project. It outlines the steps the researcher will take to organize data. Using quality metadata is an integral part of the data management plan.

DCLib AP - DC-Library Application Profiles. Metadata schema, described in detail at http://wiki.dublincore.org/index.php/DCLib_AP.

Depositor - An individual or organisation responsible for liaising between a curation system and a LTDP archive to arrange the delivery of data/metadata. This may be a Producer or may be undertaken by another party, such as a funder or other rights holder.

Descriptive Metadata - Describes the characteristics of data that will help with resource discovery. Some common descriptive metadata elements include title, creator, subject, and keywords.

Dublin Core - A standard metadata schema, made up of a general set of elements, compatible with many other schemas and can be used to describe a variety of objects.

End-User - The target of resource discovery and access systems in an Archive. The individual or system, which will use/re-use the digital resource and the associated metadata. This does not refer to an Infrastructure.

Europeana Metadata Schema - The Europeana Data Model (EDM) is the current reference schema for metadata in Europeana. Its namespace is http://www.europeana.eu/schemas/edm/. Documentation relating to EDM can be found at http://pro.europeana.eu/edm-documentation/.

Faceted Search - faceted search is a way to find records in a collection based on a system where each record is classified along multiple explicit dimensions (the facets) that correspond to the properties of the records. Users can navigate the different dimensions independent of other facets

Funders - The research institution, university, or foundation that funds a research project.

Infrastructure - The conceptual body, which sits above the repository in terms of information flow. Within this deliverable this refers to the CESSDA, CLARIN and DARIAH. The Infrastructure may require data and/or metadata are made available by member repositories in order to offer aggregation services of some kind.

ISOcat - A central registry for all concepts relevant in linguistics and the domain of language resources, including metadata categories etc. See http://www.isocat.org

Metadata Catalogue - A system providing search over metadata. It is either an institutional system, exposing only the metadata for "own" resources or an "aggregator" offering metadata collected from a number of content providers (usually with a specific community in focus, see e.g. CLARIN's VLO). It usually offers a full-text search and faceted search over selected categories (date, author, etc.), sometimes it is also accompanied by more complex representations, like geo-spatial visualizations or similar. It mostly offers a full view of the metadata record and some way of accessing the underlying resource.

Metadata Harvesting - Extracting metadata from many different repositories in order to collect it in one central catalogue. This provides researchers with easier access to a wider selection of data.

Metadata Provider – From the OAI-PMH architecture description, an organization that offers metadata for harvesting by Metadata Service Providers. Also used outside the OAI-PMH model for a party publishing metadata.

Long Term Digital Preservation (LTDP) - The guarantee that digital information and its retrieval methods are sustained over a long period of time.

Metadata Schema - A set of metadata elements (fields) and the guidelines for using them to accurately and completely describe data. For XML-based formats mostly expressed as XML Schema (XSD).

Metadata Service Provider – From the OAI-PMH architecture description, an organization that harvests metadata from Metadata Providers to provide a service such as presenting the data in a catalogue.

Normalization - The process of reducing data to a canonical form. In the metadata context it means transforming the data to a 'preferred' scheme or format e.g. normalizing date formats or normalizing to a preferred spelling.

Open Archives Initiative (OAI) - Initiative that works to create interoperability between standards in order to promote open access of data. See http://www.openarchives.org/

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) - A set of standards for metadata harvesting created by the OAI. It is based on the HTTP and XML standards, and uses unqualified Dublin Core to provide basic interoperability, although any metadata format may be used in addition to Dublin Core.

Open Archival Information System (OAIS) (Reference Model) – is a ISO standard ISO 14721:2012 defining "an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community".

Preservation Metadata - This is sometimes considered a form of administrative or technical metadata. Preservation metadata contains information that will help preserve data, such as the original format of a file.

Preservation Strategy - A course of action which is taken to guarantee the accessibility of digital data over an extended period of time (several decades), resulting in Long Term Digital Preservation (LTDP). The two major aspects are: The avoidance of loss of data (backup strategies),

and the preservation of tools and programs to access and manipulate the data.

Producer - The standard OAIS 'Producer' role responsible for pre-archival actions from conception to collection/creation of data/metadata. Often assumed to be a 'researcher' There may be several Producers and the pre-Archival phase may encompass several Curation Systems.

Repository - In cases where the presence or absence of LTDP responsibility is not of relevance, especially when describing the interaction with Infrastructures, the term repository is used over either Curation System or Archive System.

Repository - a system to store and publish resources (research data or results). A repository needs to have an organizational backing ensuring the long-term stability of the system. Individual resources are usually maintained as (potentially complex) digital objects with associated metadata and a versioning mechanism. In a repository the presence of LTDP responsibility is not of relevance.

Researcher - The person who initially creates data. This person may also be responsible for creating accompanying metadata.

Role - A defined purpose or function related to the data/metadata lifecycle.

Service Provider - The organization that harvests and provides access to metadata from different repositories. Various software to manipulate, summarize, and display the data is often included.

Structural Metadata - Describes the structure and relationships of data so that it can be interpreted correctly and viewed in the intended order. It can describe the physical or logical structure of data. An example would be the individual page numbers of a digitized book.

Super Infrastructure - A conceptual body which sits above the Infrastructures in terms of information flow. A theoretical Super Infrastructure may take data and/or metadata from Infrastructures (either via the Infrastructures or direct from the Repositories) in order to offer aggregation services of some kind. The Super Infrastructure is used as an example within the Research Infrastructure Model to indicate the layers of complexity implied by multiple layers of aggregation.

Technical Metadata - Describes the technical information of digital objects. An example of technical metadata includes the format of a digital file (i.e. pdf, jpeg, etc.) This could be seen as a subset of Administrative Metadata.

Virtual Language Observatory (VLO) – joint metadata catalog for the CLARIN community, offering round 500.000 records collected from 60 providers. http://catalog.clarin.eu/ds/vlo/

PART A APPENDICES

Appendix A: Background information about metadata

Appendix B provides background information about metadata. These sections will be used for training material.

Metadata Standards and Schemas

When talking about metadata, it is important to make clear the difference between standards and schemas, because these two terms are often used ambiguously in metadata literature. However, they have different meanings, and both are important for describing metadata.

A standard is a regulation or guideline that is created and maintained by a standards organization. We use standards to design things like railroad tracks and emergency exit signs. There is even an international standard for how to correctly brew tea (International Organization for Standardization [ISO], 2013).

Any person or organization can develop a standard, but standards created or endorsed by national and international standards organizations are usually better maintained and more widely recognized than others. As the ISO puts it, "…not all standards are created equal," (ISO, 2010, p.8). Metadata standards are metadata formats that are endorsed and maintained by a standards organization such as the ISO. An example of a metadata standard is the Dublin Core Metadata Element Set, also known as ISO 15836:2009 (ISO, 2009).

A schema is a set of individual metadata elements used to describe data (NISO, 2004). Schemas may or may not be endorsed by a standards organization. In fact, most schemas are developed and endorsed by invested community organizations rather than standards organizations. These types of schemas are usually reliable, too. An example of a community-developed schema is the Data Documentation Initiative (DDI), which is maintained by the DDI Alliance, "…a self-sustaining membership organization whose members have a voice in the development of the DDI specification," (DDI Alliance, 2009).

Choosing a Metadata Schema

It is important to utilize a widely used and well-maintained metadata schema. It may be tempting to create a new schema that is tailored specifically to the metadata needs of a certain discipline or organization, but ultimately this will hinder the shareability and reliability of your metadata. One NISO report states, "In general, the fewer metadata schema, the better. We use standards

to improve interoperability and to reduce unnecessary variation. It is better and easier to adopt something that already exists, is well modeled, and comprehensively supported,"(TC46 SC 11 Interest Group, 2011, 1.1 Introduction, para. 2). The report also points out that if you choose to create your own schema, it will be your responsibility to maintain it. Many tried and true schemas already exist. Most schemas allow you to add extensions or extra elements, enabling you to make adaptations if you find that the schema you are working with does not fit all of your needs. (Schweitzer, 2012).

Metadata Schemas

Riley and Becker's poster Seeing Standards: A Visualization of the Metadata Universe (2010)[69] is an illustrative resource showing the myriad of metadata schemas and their uses. There is no single schema that can adequately support all types of data, and this has resulted in the development of so many schemas (Broeder et al., 2010). The relationships between types of metadata, their functions, and different schemas are very complex, and the figures 17 and 18 Functions and Schemas for Different Types of Metadata illustrate one way of representing these relationships.

Metadata Standard	Type of Metadata					
	Descriptive	Contextual	Technical	Preservation	Administrative	Structural
Dublin Core	x		x		x	
DDI	x	x	x		x	
CMDI	x	x	x			x
SKOS	x					x
TEI	x	x	x	x		x
Text MD			x			
MIX			x	x		x
OAIS				x		
PREMIS			x	x	x	x
METS					x	x
OAI-ORE	x					x

Figure 17: Standards and Type of metadata

Funct	Type of Metadata					
	Descriptive	Contextual	Technical	Preservation	Administrative	Structural
Search and Discovery	x					
Describe Data	x	x				
Identify Data Creators	x				x	
Describe Data Context		x				

[69] http://www.dlib.indiana.edu/~jenlrile/metadatamap/

	Authenticate Data			x	x	x	
	Record Data Format and Size			x	x		
	Data Management			x	x	x	
	Preserve Data			x	x	x	
	Record Data Provenance			x	x	x	
	Provide Rights Management Information					x	
	Describe Data Structure				x		x

Figure 18: Functions and types of metadata

Metadata Interoperability

NISO defines interoperability as, "…the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality," (2004, p.2). Interoperability between different systems is something we experience every day. Just think about how you have to bring an electrical adaptor with you every time you travel between North America and Europe. This is because the electrical sockets in these two regions are not interoperable with each other. Now imagine that electrical sockets were not only incompatible between regions, but from one town to another. This frustrating scenario can be compared to the challenges we face when trying to create and maintain interoperability between the many different metadata schemas currently in use. In order to be able to aggregate and use metadata from different sources, it needs to be interoperable. Using the same metadata schema or creating a bridge between schemas can accomplish this; similar to how an electrical adaptor can be used as a bridge, or "crosswalk", between different electrical sockets (Digital Library Federation [DLF], 2007).

Creating and using interoperable metadata vocabularies and standards is one of the major challenges in ensuring quality metadata. Now that service providers use metadata harvesting to bring together metadata from different sources, it is even more important for metadata to be interoperable with a variety of interfaces. (Barton, Currier, & Hey, 2003; DLF, 2007; Dunlap et al., 2008; Jones, Berkley, Bojilova, & Schildhauer, 2001; Vardigan et al., 2008;). In addition to this, search processes are becoming increasingly automated (UKOLN, 2007). Automatic search processes rely heavily on accurate metadata vocabularies and structure. Before the use of automated searching, poorly documented metadata could still be useable if a human was able to interpret it. However, a machine will not be able to understand inconsistent metadata and will look over these objects, rendering the metadata virtually non-existent (Miller & Vardigan, 2005).

Structural Interoperability

In order for metadata to be structurally interoperable with other metadata, it needs to be stored in a standard format or schema. Different schemas are used for different disciplines, but the lines between disciplines are becoming blurred as researchers perform more cross-disciplinary research. Ideally, researchers should be able to search a variety of networked collections for information. In order to facilitate this, data providers need to use metadata schemas that are general enough to be interoperable with a large variety of metadata, while maintaining the detail they need to accurately describe specific information to their designated communities (Jones et al., 2001).

One method of making metadata from different sources that use different schemas interoperable is through crosswalking. Crosswalking is the process of converting metadata from one format to another by creating a bridge between formats. Data providers may want to crosswalk their metadata to a more interoperable schema in order to expose it to metadata harvesters, enabling their metadata to be accessed from different locations. When crosswalking between metadata schemas it is best to start with a schema that has rich descriptions and many elements, and then crosswalk down to a simpler format. If you are crosswalking to a much simpler format, it may be a good idea to use a number of steps to prevent loss of information (DLF, 2007).

One major benefit of creating metadata in an interoperable format is that the metadata will be available for metadata harvesting. This allows metadata to be collected and gathered into different collections to promote accessibility (Carpenter, 2003). The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is an initiative that supports sharing metadata in combined data stores, and it sets out guidelines that metadata providers can follow if they want their metadata to be exposed for harvesting. The OAI states in its mission statement that, "The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content," (Open Archives Initiative, n.d., Standards for Web Content Interoperability section).

Controlled Vocabularies

In addition to using compatible structural standards in metadata, it is also a good idea to use controlled vocabularies and to identify those vocabularies. Controlled vocabularies support interoperability between metadata records from different sources through the use of standardized terminology. Controlled vocabularies also promote precision, consistency, efficiency, and harmonization in creating metadata and the resource discovery process. If using a controlled vocabulary in a repository, make sure to provide access to a thesaurus so that researchers can view and understand the terminology. It is also important to keep the controlled vocabulary up to date so that it reflects changes in the repository's target discipline or community (DLF, 2007; Granda et al., 2009).

In order to develop and maintain an effective controlled vocabulary, a repository must identify its primary users. This can be difficult because information in a repository is not defined by its primary users, but rather the scope of a repository's primary users defines the information that it holds (Digital Preservation Coalition [DPC], 2004). A vocabulary needs to be specific enough to define concepts accurately for experts in a field while remaining general enough to be understandable to other users who are not as familiar with the terminology (Broeder et al., 2010). Most designated communities are growing wider with the rise of open access publishing. Because more information and data is now freely available, a repository's user base might be enlarged to include the public in general (DPC, 2004).

European Language Social Science Thesaurus (ELSST)
A thesaurus is a controlled vocabulary that is arranged hierarchically. It can be implemented to help users find data by allowing them to broaden or narrow their search using suggested search terms. ELSST is a multilingual thesaurus for social science that was created to enable cross-European research that is independent of language. ELSST is based on the Humanities and Social Science Thesaurus (HASSET) and was started by the Language Independent Metadata Browsing of European Resources (LIMBER) project in collaboration with the Council of European Social Science Data Archives (CESSDA). It has been extended considerably since. It uses RDF, which allows for interoperability between different metadata schemas and thesauri. One of the greatest benefits of a multilingual thesaurus like ELSST is that users can search in their own language, but still discover resources in other languages (Balkan et al., 2002).

Metadata Schema Registries

If we see metadata as a tool that is used to describe and organize data, then we can look at a metadata schema registry as a tool used to describe metadata. Metadata schema registries document information about metadata schemas such as element sets, metadata models, and thesauri for controlled vocabularies. Metadata registries help organizations develop their own consistent and interoperable metadata schemas and controlled vocabularies. Because metadata registries define standards and definitions for metadata, metadata from diverse organizations or across different organizations that use the same registry can easily be combined in a single repository (Bargmeyer & Gillmann, 2000; NISO, 2004).

ISO/IEC 11179

A helpful tool for developing a metadata registry for statistical metadata is the ISO/IEC 11179 standard. This standard provides guidelines for organizing information about individual data elements. Metadata registries based on the ISO/IEC 11179 standard describe data elements using three concepts. These are object class, property, and value domain. Object class represents the thing or idea that is being studied (e.g. "women in Canada"), property

represents a characteristic of all of the members in the object class (e.g. "income"), and the representation defines how something is typified (e.g. "non-negative integers"). Combined, these concepts create an individual data element (e.g. "Women in Canada with an income of $20,000"). Using these descriptions to describe a data element enables the transfer of data in the form of individual element sets, regardless of database structure (Bargmeyer & Gillmann, 2000).

ISO/IEC 11179 also sets out guidelines for defining and naming data elements, which creates compatibility for all of the data element sets. Another feature of this standard is that similar concepts can be harmonized, or combined. Repositories can benefit from using the ISO/IEC 11179 standard to create a metadata registry because it allows for compatibility between data from different organizations. The Australian National Health Information Knowledgebase uses ISO/IEC 11179 to link data from different health topics, definitions, standards, work programs, organizations, etc., to create unified access to all health information (Bargmeyer & Gillmann, 2000).

Ideally, metadata would be completely compatible across disciplines. This would provide the opportunity to link data from various subjects, which could lead to new interpretations of data. The UKOLN report "Dealing with Data" gives the example of being able to link crystal structures from the eCrystals institutional data repository with related protein structures in the Protein Data Bank (2007, p.52). The possibility of linking data across disciplines could lead to incredible innovations in research. The concept of linked data, which will be discussed later, is one way of realizing this capability.

However, it is difficult to establish compatible standards across disciplines. The ISO/IEC 11179 standard is one way of trying to accomplish this task, but Cory Doctorow, a blogger who is a supporter of open access, said in 2001 that, "A world of exhaustive, reliable metadata would be a utopia. It's also a pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities," (Introduction, para. 3). Much has improved in metadata quality and compatibility since then, but this statement reminds us of the continuous effort that needs to go into the standardization and compatibility of ever-changing metadata.

Types of metadata

The following section outlines the different types of metadata. Again, most types of metadata serve functions that overlap with other types of metadata, and it is very difficult to separate them into concrete categories. The descriptions below are merely one interpretation of the different types of metadata, and the distinctions between them can be redrawn in many different ways. There are many metadata schemas available to researchers and repositories for organizing and sharing data, which will be illustrated in relation to the types of metadata and their functions later.

Descriptive Metadata

Descriptive metadata is used to locate relevant data during the resource discovery stage of the research lifecycle (NISO, 2004). At this stage, the researcher is looking at previous research data in repositories to use in formulating a research question and developing a proposal. Because it is essential to first *find* relevant data in order to *use* data in new research, descriptive metadata is very important and consequently given a lot of attention in literature. The following section will focus on the benefits of quality descriptive metadata, give a brief outline of some common standards used in descriptive metadata, discuss how to create quality descriptive metadata, and provide some resources for researchers creating their own descriptive metadata.

Why Do We Need Quality Descriptive Metadata?
The primary reason a researcher relies on descriptive metadata, is because it acts as a gateway to data. Researchers can use descriptive metadata to search for and bring together data from many different locations. This helps researchers avoid bias in their studies by providing them with a large collection of data upon which to base their research (Zimmerman, 2007). Perhaps more importantly, when researchers are able to assemble a unique collection of data, they can study or compare old data in new ways and find links among previously unrelated sets of data, much like the example of linking crystal and protein structures from the UKOLN report mentioned above (Nelson 2009; UKOLN, 2007; van Uytvanck, 2010). UKOLN also points out that researchers can make new discoveries by bringing together collections of data that have "…a unique position in time and place" (2007, p. 18). This kind of data is extremely valuable because it cannot be recreated, and quality descriptive metadata prevents it from getting lost or becoming "essentially invisible" in a portal (Barton et al., 2003, p. 1). Researchers cannot conduct research if they do not have access to data, and quality descriptive metadata provides them with the tools they need to access and collect data.

The ability to find and re-use existing data using quality descriptive metadata saves time for researchers because they are less likely to unknowingly replicate research data that already exists. Saved time and resources for researchers means saved money for their funders, which makes quality descriptive metadata beneficial for them, too. However, as mentioned above, this is a two-fold process; funding organizations must also consider the extra time and work that goes into creating quality metadata when funding a research project (RIN, 2008).

In terms of the researcher who created the original data, she will benefit from quality descriptive metadata because her data will be easier for other researchers to find. This creates a higher potential for data re-use in new research, which translates into more citations for the original researcher (Nelson, 2009). Conversely, if data is accompanied by poor descriptive metadata, it will be harder for other researchers to find, meaning less visibility for the original researcher. This is an issue for repositories, because if people

have recurring difficulties in locating relevant research data within a particular repository, they may be deterred from using that repository in the future (Barton et al., 2003). Descriptive metadata is a crucial part of the research lifecycle, and the following section will provide some basic tips and guidelines for creating quality descriptive metadata.

How Do You Create Quality Descriptive Metadata?
Now that we know *why* it is important to create quality descriptive metadata, we need to know *how* to create quality descriptive metadata. The answer is through communication between actors. It is a repository's responsibility to ensure that it provides a high enough quality of metadata to make the data it describes useful to researchers (UKOLN, 2007). Descriptive metadata may be created by a repository or by the researcher who originally created the data. The most important thing to remember about researcher created metadata is that *most researchers are not information specialists* (Campbell, 2007; Hillman et al., 2004). Ultimately, the repository and the researcher must come to a compromise between creating extremely high quality metadata and avoiding standards that are above the skill level of the people who will be creating the metadata (Barker & Ryan, 2003).

Keeping this in mind, there are many tools such as guidelines and easy-to-use interfaces that can guide any researcher through the descriptive metadata information process. Utilizing these will result in a higher quality of descriptive metadata being deposited into a repository, resulting in end-user satisfaction and less time spent fixing poor metadata. Some ideas are as follows:

- **Create an online interface that helps researchers to create quality metadata.** This may be in the form of an online template with features such as dropdown menus, spellcheck tools, and functions for browsing and searching authority lists (Currier, Barton, O'Beirne, & Ryan, 2004; Mohler et al. 2010; Wilson, 2007). The Darwin Core Archive Assistant uses a mouse hover function on their metadata creation interface that provides terms and explanations for each of the different elements (GBIF, 2011), and the UK Data Service provides a helpful online deposit form that can be found on their website (2013).
- **Produce a manual or guideline that outlines how to create quality metadata.** Providing a simple metadata manual to researchers could help them understand the basic fundamentals of metadata creation. It could describe the benefits of documenting high quality metadata, and include information about the metadata schema, how to fill in different element fields, and the terminology or controlled vocabulary that is being used. A manual can also be used in conjunction with a survey or template, and give directions for using these tools. (Barton et al., 2003; Wayne 2005).
- **Integrate a means of checking metadata quality directly into the metadata creation process**. Providing an automatic correction system into a metadata creation tool will help researchers identify and fix mistakes before submitting metadata. However, if implementing a system that automatically detects mistakes, it is important to also include a solu-

tion for fixing the mistake. Otherwise, researchers may not correct their errors because they do not know how, or, in a worst-case scenario, they will simply stop submitting metadata (Barton et al., 2003; Broeder et al., 2006; Park 2009).

- **Test the Interpretability and Discoverability of metadata before making it public.** A good question to ask is, "Would someone unfamiliar with this resource be able to identify it from looking at the metadata?" You can even ask someone else to look at your metadata and attempt to answer this question (University Library, University of Illinois at Urbana-Champaign, 2010).
- **Provide a means for peer-review of data.** Data Archiving and Networked Services (DANS) conducted a pilot project in which data-users were surveyed about the quality of the data they accessed on EASY, their online archiving system. The project was successful, and the data-users provided feedback that could help metadata creators improve their metadata. DANS has plans to implement data reviewing as a permanent feature of its archive (Grootveld, Egmond, & Sørensen, 2011).

Contextual Metadata

Contextual metadata is the metadata that describes details about the background of research data. Researchers usually create this type of data, and other researchers use it during the "Gather Resources" and "Analyse and Experiment" stages of the research lifecycle. It places data within a context by describing elements such as location, time, provenance, how data was collected, what tools were used to collect data, sources of data, methodology of a study, and what questions were asked in interviews. Contextual metadata is especially important when sharing numerical data, because this kind of data would have no meaning without these descriptive elements attached to it. Researchers need contextual metadata in order to understand and use other researchers' data correctly (EDINA and Data Library, University of Edinburgh, n.d.).

Why Do We Need Quality Contextual Metadata?
Contextual metadata is important because it shows researchers *how* to use data. It describes data so that researchers can determine if it is suitable for their needs. It answers questions such as, "Does this data come from a large enough sample to use in my research?" or "Is this data recent enough to use in my research?" This kind of information also proves the integrity of the original data creators by providing documentation showing how their data was collected (Dunlap et al., 2008; DPC, 2004; EDINA and Data Library, University of Edinburgh, n.d.; Vardigan et al., 2008; Zimmerman, 2007).

Complete contextual data serves as a substitute for direct communication between researchers. Whereas a researcher trying to understand a dataset without contextual metadata would have to contact the original creator of the data in order to understand its meaning and context, complete contextual metadata replaces this communication by allowing researchers to fully

understand the meaning of a set of data (Edwards et al., 2011; DPC, 2004; Dunlap et al., 2008; UK Data Archive, 2011). This opens up a world of information to researchers, since contacting the original creator of data is usually not a possibility. Contextual metadata is crucial for creating new knowledge, because data does not become knowledge until the context, background, and basic assumptions that accompany it are also conveyed (Eppler, 2008).

Researchers will not be able to use someone else's data if it is not accompanied by complete contextual metadata, because data sets have very little meaning without a context. A lack of contextual information could force researchers to narrow their database, which would hinder research advancements (Zimmerman, 2007).

How Do You Create Quality Contextual Metadata?
Contextual metadata is usually created by resource creators; therefore, it is important to keep the same lines of communication and instruction open between researchers and repositories that you would for descriptive metadata. One important thing to remember when creating contextual metadata is that metadata and the way in which we use it is continuously changing (Edwards et al., 2011). We cannot foresee how someone else will use data, and we do not know how people will use data in fifty years. It is impossible to anticipate every single purpose that data might serve, but it is best practice to record as much contextual metadata as possible, and not only what *you* think will be useful (Day, 2005; Nelson, 2009).

Researchers have the most contextual knowledge over their own data. One study compared metadata created by resource creators with metadata created by information specialists. The results showed that while information specialists had a better understanding of metadata schemas than resource creators, the resource creators had a *much* better understanding of their own data's context. The study concluded that the best metadata is produced when resource creators and information specialists combine their knowledge to create metadata (O'Beirne as cited in Barton et al., 2003, p. 4-5). It may not be possible for a repository to provide this resource to researchers, but O'Beirne's study demonstrates that we should not underestimate the abilities of researchers to produce good metadata. If they are given the tools to learn how to create metadata, they will ultimately be able to create high quality metadata because of their exclusive knowledge of their own research data.

Technical Metadata

Technical metadata is closely linked with administrative metadata and preservation metadata. Preservation and technical metadata are sometimes considered types of administrative metadata, because these types of metadata are primarily used for data management. They are sometimes considered "back end" metadata because they are used for data processing and storage at

the repository rather than the interpretation of data content (Shankaranarayanan & Even, 2006).

Technical metadata records aspects of a digital file such as file type, size, date of creation, and the digital capture process (if the object was not born digital). Technical metadata is typically captured automatically. This may be done by a software framework such as JHOVE, which "…provides functions to perform format-specific identification, validation, and characterization of digital objects," (JHOVE, 2009).

Why Do We Need Quality Technical Metadata?
Technical metadata is important because it provides information for viewing digital data, such as font size, dimensions, and bit depth (University Library, University of Illinois at Urbana-Champaign, 2010). Another purpose of technical metadata is to provide information about how a digital file was captured. This is especially important for visual objects, such as photos, so that a researcher or other end-user can ascertain the quality and accuracy of a digital representation of an object (NISO, 2004). Technical metadata is used to manage data after it has been submitted to a repository; its presence is not as apparent to end-users as descriptive or contextual metadata, but without technical metadata, digital data would be unusable.

How Do You Create Quality Technical Metadata?
Because technical metadata is usually captured automatically, the repository's main function in terms of managing quality technical metadata is to ensure that all necessary metadata is being created and kept with its data. Additionally, technical metadata should be easily accessible by end-users should they want access to this information.

Preservation Metadata

Preservation metadata records information that maintains the longevity of a digital data object for future use. Large amounts of useful digital data from the past are no longer accessible because they exist in outdated formats. Therefore, more steps are being taken now to preserve newly created digital data far into the future. Preservation metadata includes the information that enables us to store digital data in sustainable formats. When researchers deposit their data into a repository, they trust that repository to preserve their data for a long time, and it is the repository's responsibility to researchers to fulfill this task (UKOLN, 2007). Preservation metadata is part of the "Store and Archive" stage of the research lifecycle.

Why Do We Need Quality Preservation Metadata?
Digital files degrade over time and need the proper care to remain useful, much like a physical object such as a book will fall apart if it is not stored correctly. In fact, digital data is much more fragile than paper files because of the fast moving nature of technology. In David Holdsworth's article on

preservation strategies (2007), he notes that data written on the earliest forms of software, which are only sixty years old, has already been lost because the software is now obsolete. Software developers in that time did not foresee that their data would be so valuable one day, nor were they thinking about technological obsolescence or digital preservation. However, we now realize the importance of preserving data, and we use preservation metadata to do this.

Preservation metadata also proves data's integrity to researchers. They can be assured that there has been no loss, degradation, or alteration of data by looking at metadata that contains information about data's fixity. This information is usually obtained by periodically running a checksum on data sets. Preservation metadata contains the hashing algorithm used on data and results that were produced from earlier tests. The new test results can be compared with the older results in order to identify any changes in the data (Caplan, 2006; Day, 2005).

Quality preservation metadata provides a payoff just like other types of metadata. Although there is an initial cost for a repository to implement these standards and procedures, it will ultimately save money by preserving irreplaceable data, as well as preventing the need for researchers to recreate lost data (RIN, 2008).

How Do You Create Quality Preservation Metadata?
Preserving digital data requires an entirely different mindset than that required for preserving analogue data. Whereas we preserve the *medium* of physical data (like the paper it is recorded on and the ink it is written in), we preserve the *information* contained in digital data. In fact, the digital data contained on a medium such as a CD will usually become unreadable through obsoletion before the physical CD deteriorates (Holdsworth, 2007).

Keeping this in mind, repositories and researchers need to have some sort of communication about what parts of digital data are most important to preserve, and record this information in the metadata. These crucial parts of data are called "significant properties" (Caplan, 2006, p. 13). Typically significant properties include data such as the text of a memo or the numbers in a data set, but there is debate over the importance of other properties such as markup and font. Significant properties need to be recorded in the preservation metadata to tell the repository what parts of data need to be preserved throughout future migrations and emulations (Caplan, 2006). Holdsworth believes that repositories should preserve as much data as possible rather than choosing the most important elements. He points out that the cost of storing digital files is inexpensive and that you never know what information will be of interest to someone in the future. Just consider how archaeologists excavate the trash heaps of ancient cultures to discover information about them (2007, p. 22).

The following are additional pieces of information that repositories need to include in preservation metadata to make sure that digital data remains

usable:

- **Migrations and other preservation activities** Information about these activities such as what actions were performed on data, dates that these actions took place, who performed them, and why they were performed should be recorded (Caplan, 2006).
- **Reliability and quality of data** A review of the data will help future researchers decide for themselves if data is trustworthy (Caplan, 2006). DANS is planning on implementing an online review process for users of data sets in EASY, so that end-users can peer-review and see other reviews of data (Grootveld et al., 2011).
- **Environment for use** This is the hardware, software, and ancillary files needed to use a digital object, and it could include information such as the database model used to read data tables in a database. Research data is useless if a researcher does not have the tools needed to access and use it (Caplan, 2006).

Administrative Metadata

Administrative metadata records information that is used to manage and document the life of data. This is a part of the "store and archive" stage of the data lifecycle, and it is usually the repository's responsibility to maintain administrative metadata. This type of metadata is closely related to preservation and technical metadata because it records what actions have been done to data, who performed those actions, and when those actions took place. These functions may also be categorized under preservation metadata, and sometimes preservation metadata is considered a form of administrative metadata. These elements provide information about the data's history and any changes that have been made to it, which helps researchers to judge its integrity. Administrative metadata also includes rights management information, which tells how data can be used by other people (Day, 2005).

Why Do We Need Administrative Metadata?
Administrative metadata not only helps repositories manage data, it also prevents researchers from misinterpreting data by recording any changes that may have occurred to data. This information is called provenance (University Library, University of Illinois at Urbana-Champaign, 2010). Ideally, when metadata is updated to reflect changes that have been made to data, the older version of the metadata will also be saved. Allowing access to this kind of provenance information helps prove the integrity of data by providing a record of all changes that have happened to data *and* metadata (DDI, 2009). Not only do researchers need the information found in administrative metadata to judge if data is trustworthy, but recording information like provenance reflects well on a repository's reliability (Day, 2005).

How Do You Create Quality Administrative Metadata?
Quality administrative metadata will include provenance information about

data, including information about events that occurred before the repository acquired the data. A repository should maintain transparency about all actions and changes that occur to data while it is at that repository. The following are elements that should be included in administrative metadata:

- **Persistent identifiers** These identify the location of data. It is important that these references are kept up to date so that the location of the data does not get lost. Persistent identifiers should also be included in descriptive metadata because they provide access to data. (University Library, University of Illinois at Urbana-Champaign, 2010).
- **Rights management information** This includes information about copyright, access, use, and licensing. Rights management information is important so that data isn't used in a way that the owner or original creator of the data did not intend (University Library, University of Illinois at Urbana-Champaign, 2010).
- **Provenance** This records information about the lifecycle of data, such as who has owned data and changes that have been made to it (Day, 2005; University Library, University of Illinois at Urbana-Champaign, 2010).
- **Proof of authenticity** This helps prove the integrity of data. An example of proof of authenticity is providing documentation of peer review (RIN, 2008).

Structural Metadata

Structural metadata tells how a data file is organized. For example, it could record the order of the pages of a digitized book (NISO, 2004). A compound data object, also called an aggregation, might contain text, audio, and visual media, such as a PowerPoint presentation with an audio file and accompanying lecture notes. Structural metadata explains the structure and relation between these individual objects, and helps end-users to use an object in the way the creator intended. Structural metadata should also be used when the individual parts of a digital object are not only useful as a whole, but also separately (University Library, University of Illinois at Urbana-Champaign, 2010; Yale University Library, 2008).

Why Do We Need Quality Structural Metadata?
Structural metadata makes aggregate objects easier to use for the end-user. In the case of a digital object with many different parts, it may be impossible to make sense of the object without knowing the logical order defined in the structural metadata. On another level, structural metadata can even create an aggregate object out of digital objects from a variety of sources by setting out a way of organizing them in a meaningful and logical way. In some cases structural metadata is not necessary; for example, if a digital object is comprised of only one part, such as a single picture that is not part of a collection (Yale University Library, 2008).

How Do You Create Quality Structural Metadata?
When creating structural metadata, it is important to set standards for how each part of a complex object will be represented, so that each component is identified using the same structure (University Library, University of Illinois at Urbana-Champaign, 2010). There may be a different level of granularity for each complex object, so the different levels at which you create individual structural metadata may vary for different objects. One preliminary study conducted at Yale showed that users were more satisfied when structural metadata was represented in a user interface with access points that allowed people to easily navigate around a complex object (Yale University Library, 2008). However, such an interface would not be sustainable, and it would be interesting to know if Yale has a method for preserving the user interface along with the metadata.

Saving Time and Money with Quality Metadata

There exists a definite tension between the time and effort it takes to create high quality metadata and the costs associated with it. Many scholars believe that creating quality metadata will ultimately save money. Although extra time may be spent on creating quality metadata, the payback later comes in the form of saved time in managing and searching for data (Currier, et al., 2004; EDINA and Data Library, University of Edinburgh, n.d.; Lyon, 2007; Mohler, et al., 2010; NISO, 2007; RIN, 2008). Barton et al. (2003) are of the opinion that we should strive to create high quality metadata "…within the inevitable limitations of time and cost," (p.4). In other words, the tradeoff between extremely high quality metadata and the time it takes to create it is not worth the potential time and money that could be saved later. No study has been able to show if the value of time saved from using quality metadata outweighs the monetary cost of creating high quality metadata, and this is an area that would greatly benefit from further research.

Some Tips for Creating Quality Metadata

Keeping in line with the theme of maintaining good communication throughout the research process, it is generally believed that advocacy and education can greatly increase the quality of metadata (UKOLN, 2007). All actors should be involved in these activities. Funders and repositories can play a big role in advocacy for good data management, because they represent the "…standard bearers for metadata best practices," (DDI Alliance, 2009, Introduction). Funders especially can advocate for quality metadata by creating metadata policies and requiring the researchers or institutions they fund to follow those policies (UKOLN, 2007).

Advocacy for quality metadata can begin at the undergraduate level. Universities and research institutions would benefit greatly from investing in educating young researchers in metadata, and students would have the foundations needed to create high quality metadata for their future research projects (UKOLN, 2007). However, metadata and technology change rapidly,

and it is important for repositories and institutions to implement ongoing training programs to keep key actors up to date on innovations in metadata standards and best practices. They should also provide support services throughout each stage of the research lifecycle to ensure quality and consistency (Currier et al., 2004; Park, 2009; RIN, 2008).

When creating metadata schemas and vocabularies, good communication is key. Many schemas have been created by member participation, and are continuously updated using member feedback (Broeder et al., 2010; DPC, 2004; Dunlap et al., 2008; Vardigan et al., 2008). Additionally, providing a forum where users can express their suggestions and needs can be a great way for a repository to identify and correct issues with its metadata (Stvilia et al., 2004).

Another important measure that repositories should take to ensure metadata quality is implementing a system of checks and audits before making metadata public (UKOLN, 2007). This could include regularly assessing random samples of metadata from your repository to get an idea of its strengths and weaknesses. There are tools available for automated evaluation of metadata, and The National Science Digital Library claims that this has greatly improved the efficiency of its evaluation (Guy et al., 2004, Implement Appropriate Quality Control Processes section, para. 2). Repositories can also ask a service provider to conduct a test harvest on their sites. This helps identify technical issues before making metadata public (DLF, 2007).

Being aware of metadata quality is important. Metadata creators should consider how people at other stages of the research lifecycle or from other disciplines may use their metadata (UKOLN, 2007). Currently, there is still much work to be done in raising the standard of metadata quality. Just to give an example, in Stvilia et al.'s study of 150 metadata records harvested under the OAI protocols, they found that ninety-four percent of the records contained duplicate information and twenty-four percent had broken identifier links (2004, section 3, para. 2). Employing just some of the ideas listed here can greatly improve a repository's metadata.

Resources for Creating Quality Metadata

Here are some examples of online learning resources that researchers can use to learn more about creating quality metadata:

UK Data Archive, Create & Manage Data: Training Resources A series of PowerPoint presentations covering a variety of topics about data management, including "Formatting Your Data" and "Sharing Your Data" (Research Data Management Team, 2012).

Other resources of UK Data Archive: Managing sharing[70] and Documenting Your Data[71]

[70] http://data-archive.ac.uk/media/2894/managingsharing.pdf

[71] http://www.data-archive.ac.uk/create-manage/document

How to Develop a Data Management and Sharing Plan - This guide provides an outline for researchers explaining how to manage and prepare data for effective sharing (Jones, 2011).

Mantra Research Data Management Training - A series of online interactive lessons that cover topics such as "Organising Data" and "Documentation and Metadata" (EDINA and Data Library, University of Edinburgh, n.d.).

Functions and Schemas for different Types of Metadata

Data Documentation Initiative (DDI)
A metadata standard expressed in XML that is often used in the Social Sciences because it allows for description of numerical data sets, (Vardigan et al., 2008).

Simple Knowledge Organization System (SKOS)
The SKOS website describes this system as, "…an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web," (World Wide Web Consortium [W3C], 2012, Introduction to SKOS). SKOS uses Resource Description Framework (RDF), which makes it interoperable with many other standards.

TextMD
An XML schema typically used as an extension schema for Metadata Encoding and Transmission Standard (METS). It is well suited for recording technical metadata for text-based digital objects (Library of Congress, 2013).

NISO Metadata for Images in XML (MIX)
An XML schema used for recording technical metadata for digital images. It provides a format for storing the metadata elements defined in the "Data Dictionary- Technical Metadata for Digital Still Images (ANSI/NISO Z39.87-2006)" (Library of Congress, 2008).

Open Archival Information System (OAIS)
A reference model that provides a set of criteria intended to define the role of repositories in preserving and providing access to information in the long term (DPC, 2004).

Preservation Metadata: Implementation Strategies (PREMIS)
A Data Dictionary that defines preservation metadata, and a set of XML schema that support the implementation of the Data Dictionary (Preservation Metadata: Implementation Strategies [PREMIS], 2012).

Metadata Encoding and Transmission Standard (METS)
An XML standard that was designed to enable repositories to easily exchange digital objects. It has an administrative metadata section that includes

elements for intellectual property rights and provenance (DLF, 2010).

Open Archives Initiative Object Reuse and Exchange (OAI-ORE)
A standard that is designed for describing aggregations of Web resources in order to highlight the potential of the rich content of these complex objects (Open Archives Initiative, n.d.).

Component Metadata Infrastructure (CMDI)
CMDI is unique because it provides a framework to re-use existing sets of metadata elements. It stores these groups of "components" in a component registry, which allows for the sharing of metadata from different communities (CLARIN-ERIC, 2013).

Dublin Core (DC)
A set of core descriptive metadata elements that is used in many metadataschemas and the OAI-PMH. Dublin Core can be used in conjunction with more specific vocabularies and schemas to tailor to a community's needs (DCMI, 2013).

Text Encoding Initiative (TEI)
Largely used for text objects in the humanities, social sciences, and linguistics, the TEI is ideal for search and discovery and preservation of digital text objects (TEI Consortium, n.d.).

Appendix B: Data Lifecycle models

B1: OAIS Reference Model

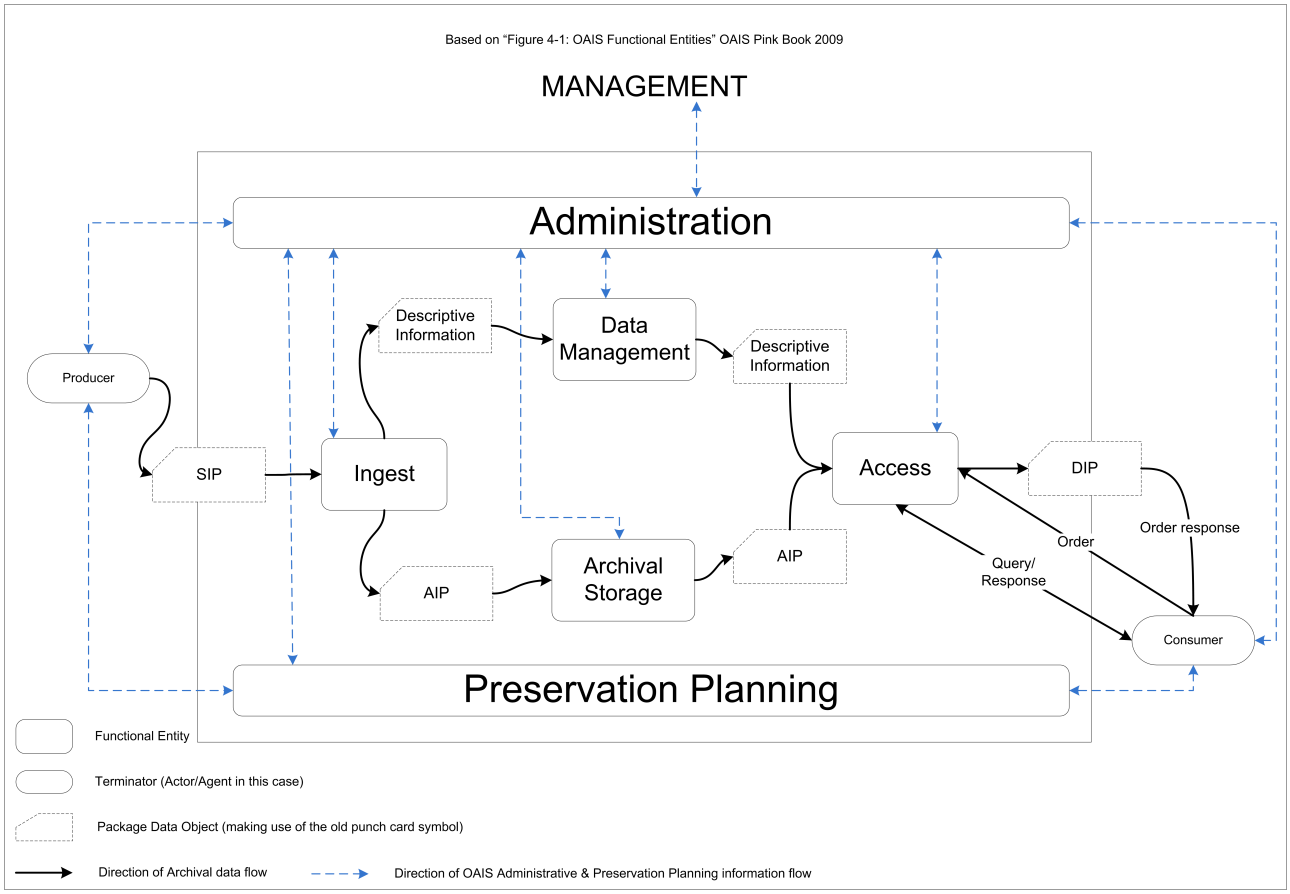


Figure 19: OAIS Reference Model

B2: DCC Curation Lifecycle Model

All of the DCC Lifecycle Items are mapped into the main body of this text.

http://www.dcc.ac.uk/resources/curation-lifecycle-model

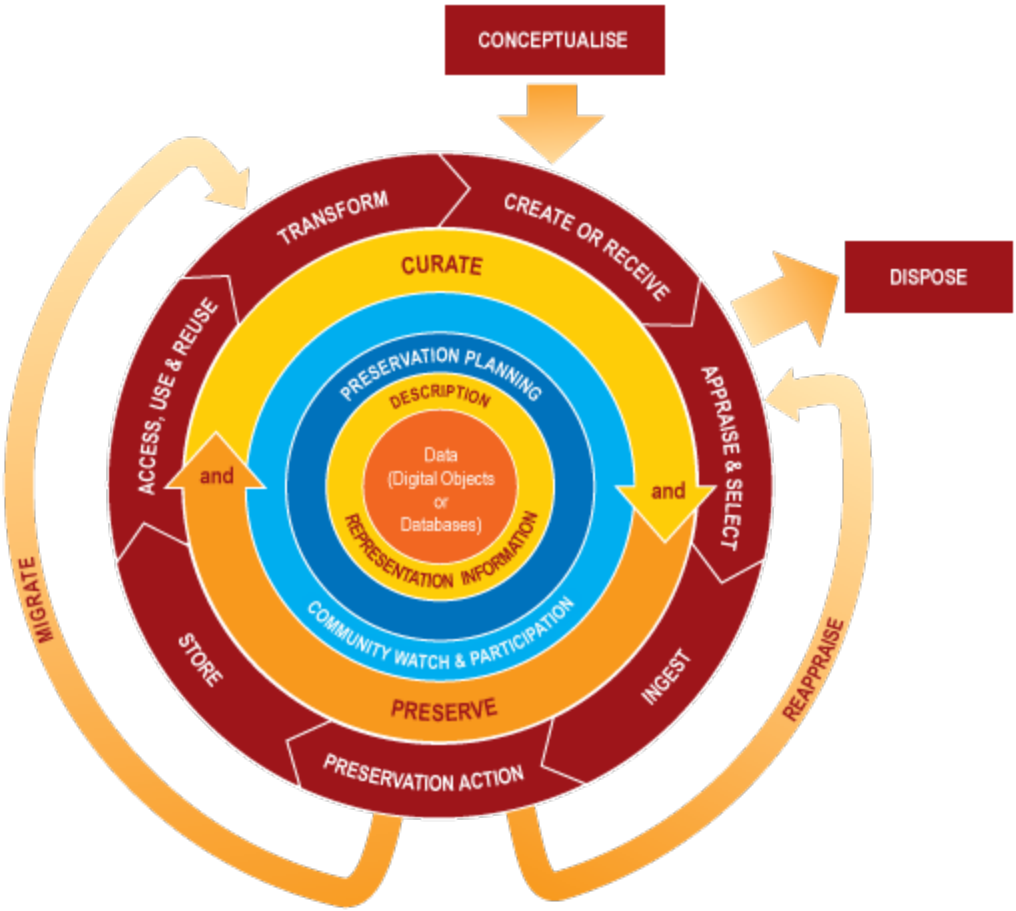


Figure 20: DCC Lifecycle

DATA

Data, any information in binary digital form, is at the centre of the Curation Lifecycle. This includes:

Digital Objects: simple digital objects (discrete digital items such as text files, image files or sound files, along with their related identifiers and metadata) or complex digital objects (discrete digital objects made by combining a number of other digital objects, such as websites).

Databases: structured collections of records or data stored in a computer system.

FULL LIFECYCLE ACTIONS

Description and Representation Information
Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata.

Preservation Planning
Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.

Community Watch and Participation
Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.

Curate and Preserve
Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

SEQUENTIAL ACTIONS

Conceptualise
Conceive and plan the creation of data, including capture method and storage options.
Link to Checklist

Create or Receive
Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.
Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.
Link to Checklist

Appraise and Select
Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.
Link to Checklist

Ingest
Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.
Link to Checklist

Preservation Action
Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.
Link to Checklist

Store
Store the data in a secure manner adhering to relevant standards.
Link to Checklist

Access, Use and Reuse
Ensure that data is accessible to both designated users and reusers, on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.
Link to Checklist

Transform
Create new data from the original, for example:

by migration into a different format, or
by creating a subset, by selection or query, to create newly derived results, perhaps for publication.

OCCASIONAL ACTIONS

Dispose
Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements.

Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction.

Reappraise
Return data that fails validation procedures for further appraisal and re-selection.

Migrate
Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.[72]

B3: DDI-L: Combined Lifecycle Model

All of the DDI-L Lifecycle Items are mapped into the main body of this text but there are some partial mappings. Data Archiving is mapped to Archival Storage but applies to most points from Pre-Ingest to Access; Data Processing occurs during both Create/Capture and in a more limited sense during Ingest.

[72] See more at: http://www.dcc.ac.uk/resources/curation-lifecycle-model#sthash.JqqcRsZL.dpuf

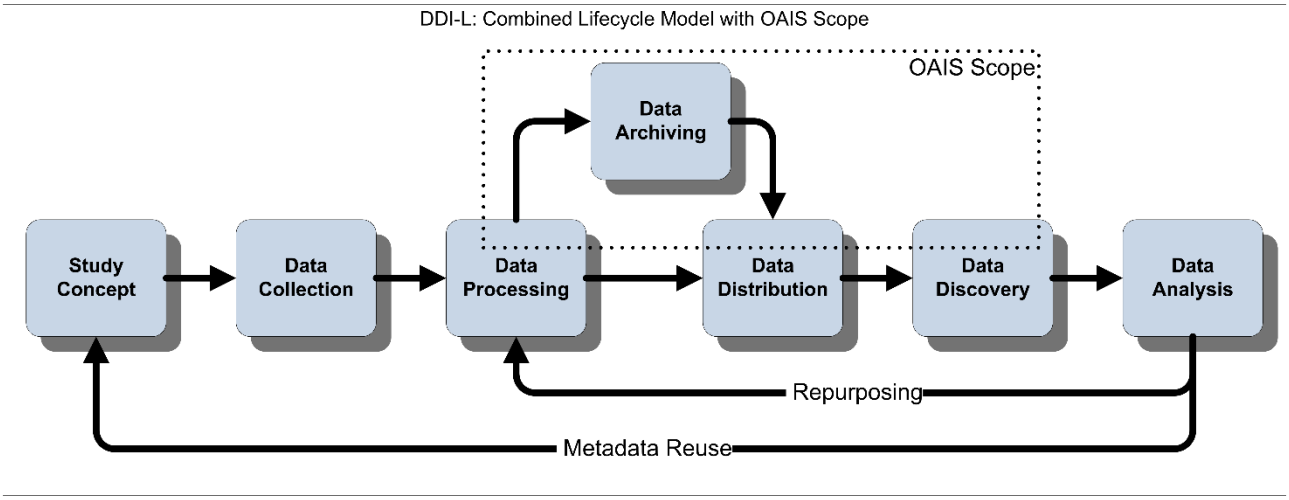


Figure 21: DDI-L Combined Lifecycle

Study Concept
Data Collection
Data Processing
Data Archiving
Data Distribution
Data Discovery
Data Analysis

B4: Generic Longitudinal Business Process Model (GLBPM)

All of the GLBPM primary headings (1-9) are mapped into the main body of this text.

http://www.ddialliance.org/system/files/GenericLongitudinalBusinessProcessModel.pdf

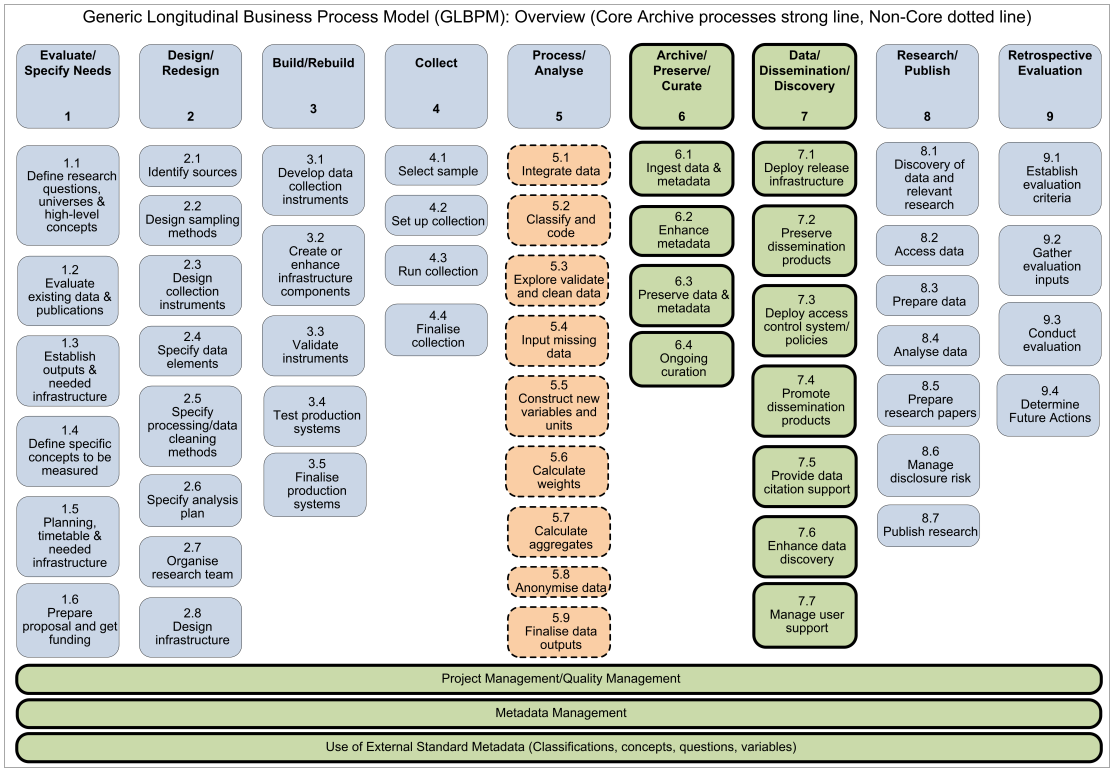


Figure 22: GLBPM Generic Longitudinal Business Process Model

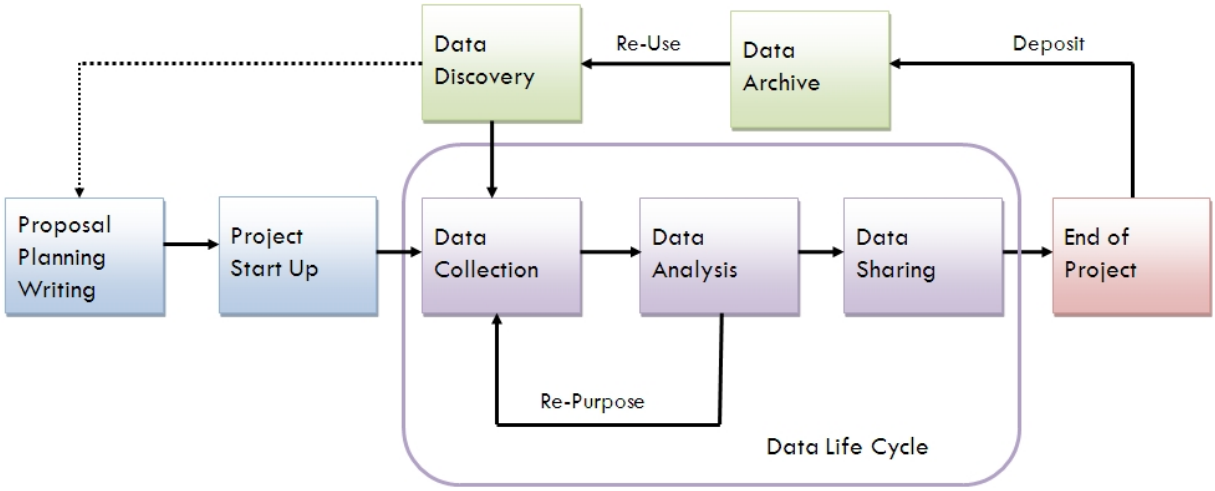
Evaluate/Specify Needs
Design/Redesign
Build/Rebuild
Collect
Process Analyse
Archive/Preserve/Curate
Data Dissemination/Discovery
Research/Publish
Retrospective Evaluation

B5: The Research Lifecycle: Traditional Model (DWB)

Figure 23: Research Data Lifecycle Diagram from Data without Boundaries (DWB)

Collaborate and Communicate (Centre)
Develop Proposal
Gather Resources
Analyse and Experiment
Publish and Disseminate
Store and Archive
Search and Discovery

B6: Steps in the Research Life Cycle (DMConsult)



http://dmconsult.library.virginia.edu/

Figure 24: Data Management Consulting Group (DMConsult) Research Lifecycle

Note that the reference headings and the diagram (provided for reference under DASISH T5.3) don't match exactly.

Document Headings
Proposal Planning & Writing

- Conduct a review of existing data sets
- Determine if project will produce a new dataset (or combing existing)
- Investigate archiving challenges, consent and confidentiality
- Identify potential users of your data
- Determine costs related to archiving
- Contact Archives for advice (Look for archives)
- Project Start Up

Create a data management plan

- Make decisions about documentation form and content
- Conduct pretest & tests of materials and methods
- Data Collection

Follow best practice

- Organize files, backups & storage, QA for data collection
- Think about access control and security

Data Analysis

- Manage file versions
- Document analysis and file manipulations
- Data Sharing

Determine file formats

- Contact Archive for advice
- Document (more) and clean up data
- End of Project

Write Paper

- Submit Report Findings
- Deposit Data in Data Archive (Repository)

Diagram Headings

Proposal Planning Writing

Project Start UP

Data Collection

Data Analysis

Data Sharing

End of Project

Data Archive

B7: Authenticity Protocol Information from APARSEN WP24

In the WP24 work on APARSEN overall stages (identified as a critical minimum for apply an authenticity protocol) are defined as

Pre-Ingest/Keeping Phase

- CAPTURE: the DR is delivered by its author to a keeping system;
- INTEGRATE: new information is added or associated to a DR already stored in the keeping system;
- AGGREGATE: several DR, already stored in the keeping system, are aggregated to form a new DR;
- DELETE: a DR, stored in the keeping system is deleted, after is preservation time has expired, according to a stated policy;
- MIGRATE: one or several components of the DR are converted to a new format;
- TRANSFER: a DR stored in a keeping system is transferred to another keeping system;
- SUBMIT: a DR stored in a keeping system is delivered to a LTDP system.

Long Term Digital Preservation (LTDP) Phase

An AIC below is defined as an object composed of several AIP.

- LTDP-INGEST: a DR delivered from a producer is ingested by the LTDP system and stored as an AIP;
- LTDP-AGGREGATE: one or several DRs stored in different AIPs, are aggregated in a single AIC;
- LTDP-EXTRACT: one or several DRs which are extracted from an AIC to form an individual AIPs;
- LTDP-MIGRATE: one or several components of a DR are converted to a new format;
- LTDP-DELETE: one or several DR, preserved in the LTDP system and stored as part of an AIP are deleted, after their stated preservation time has expired;
- LTDP-TRANSFER: a DR stored in a LTDP system is transferred to another LTDP system.

Appendix C: Case study UK Data Archive

C.1 Introduction

This case study describes the metadata workflows within the UK Data Archive (the 'Archive'), which forms part of the CESSDA Infrastructure, with a view to supporting future guidance on increasing metadata quality as part of DASISH T5.3. The case study describes the kinds of metadata in use, the workflows and procedures in which metadata plays a role, the different roles and responsibilities of the stakeholders and the procedures to ensure metadata quality.

With over 40 years of use and promotion of metadata standards for collection description and data description in the social science domain, the UK Data Archive is an active promoter and guardian of standards.

We are active players in the development and maintenance of standards for the full range of data types that we support. This includes:

- working to develop robust metadata standards for economic and social science data through the Data Documentation Initiative (DDI) - our staff sit on the DDI Technical Implementation Group, the DDI Controlled Vocabularies Group and the DDI Qualitative metadata working group
- leading work on developing an easy-to-use standard for complex qualitative data collections, through the use of QuDEx and TEI metadata standards
- actively supporting the Discovery Open Metadata Principles
- collaborating in the identification of a single Organisational Identifier model for the UK though our membership on the Jisc-CASRAI OrgID Working Group

C.2 Background

History

The UK Data Archive acquires, curates and provides access to the largest collection of digital data in the social sciences and humanities in the United Kingdom. With several thousand datasets relating to society, both historical and contemporary, covering surveys questionnaires and interview since its establishment as the Social Science Research Council (SSRC) Data Bank (The SSRC was the original name of the Economic and Social Research Council ESRC) in 1967.

A detailed record of the 40 year history of the UK Data Archive is available at http://data-archive.ac.uk/about/archive/decades

Organizational Context & Infrastructure
The UK Data Archive is part of the CESSDA umbrella organization for social science data archives across Europe; currently transitioning to the status of a European Research Infrastructure Consortium (CESSDA ERIC).

Our organization and activities are largely funded by the ESRC and as a department of the University of Essex.

The UK Data Archive works closely with its funders, the ESRC and JISC, the Office for National Statistics and other key government data providers. It also has close links with the National Centre for e-Social Science and the National Centre for Research Methods. It has been designated a Place of Deposit for public records for The National Archives.

The UK Data Archive is the UK national member institution of Inter-university Consortium for Political and Social Research (ICPSR) in the USA as well as the International Federation of Data Organizations (IFDO). It also contributes to the development of the Data Documentation Initiative (DDI).

Since 2005 the Archive has been designated a Place of Deposit by the National Archives allowing it to curate public records. High quality data are acquired from the academic, public, and commercial sectors, providing continuous access to these data while the Archive also supports existing and emerging communities of data users.

The Archive manages the UK Data Service which is the UK's flagship portal for research resources, where key national and international survey data collections, international databanks, census data and qualitative data are hosted. The UK Data Service also provides access to disclosive and more sensitive data through the Archive's secure data services. The Archive is engaged in a number of data management and preservation initiatives, supported by the ESRC, MRC (Medical Research Council), Jisc and the EU as well as providing data curation for other organisations.

The UK Data Service replaces a number of previous services and the UK Data Archive and UK Data Service replace or maintain previous collections and services including:

- The Economic and Social Data Service (ESDS)
- The History Data Service incorporating Online Historical Population Reports at www.Histpop.org, the Enclosure Maps database at hds.essex.ac.uk/em/index.html and the Contemporary and Historical Census Collections (CHCC) at http://hds.essex.ac.uk/history/data/chcc.asp.
- Census.ac.uk
- Rural Economy and Land Use Programme Data Support Service
- Secure Data Service

- Survey Resources Network
- The Survey Resources Network (SRN)
- UKDA StatServe
- ESRC Data Store (formerly UKDA-store)

ESRC Data Store is a self-archiving system hosted by the UK Data Archive aligned with but distinct from the primary curated processed and the central Archival Storage system. Its focus is the storage and sharing of primary research data from the social and behavioural sciences.

Mission
The mission of the UK Data Archive is to support high quality research, teaching and learning in the social sciences and humanities by acquiring, developing and managing data and related digital resources, and by promoting and disseminating these resources as widely and effectively as possible.

Source: http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

Main activities
The UK Data Archive provides access to over 5,000 social science data sets including both quantitative data and qualitative data from a wide range of disciplines. Access to most resources currently requires registration but access to the data catalogue, including online documentation such as questionnaires, does not require registration. The Archive is committed to increases in Open Data availability and the provision of data at multiple levels of disclosure risk with appropriate controls.

The UK Data Archive ensures that the data will be available not only to current researchers but also to future researchers through digital preservation and migration to new storage media as technology evolves.

To promote the use and re-use of its data, the UK Data Archive provides technical support and advice to users on how to access and use the data, and on data management issues. The UK Data Archive also works closely with national and international partners on data-related projects and initiatives. We provide user support in the form of an email help desk and telephone help line.

Services provided on our websites aim to be well laid out and enable users to identify available resources, directing them to the resources they need or to the people who can provide them with access to such resources.

Materials
The materials handled at the UK Data Archive range from Standard office documents (text documents, spreadsheets, presentations) to databases, images, audio-visual multimedia, scientific and statistical data formats, raw data, plain text and structured text. The represent 'data of relevance to research' (a wider catchment than 'data generated by research') across the humanities, social sciences and life sciences including government data, cultural data, biomedical data and currently moving towards administrative

data.

A large part of the UK Data Archive's data collection consists of publicly funded data, especially large-scale statistical surveys such as the General Household Survey and Labour Force Survey. Another important source of data is the academic community, sponsored by the ESRC and other funding bodies; in this category we hold studies such as the British Household Panel Survey and the Millennium Cohort Study. The UK Data Archive also provides access to important international macrodata series (aggregate data) such as those held by the Organisation for Economic Co-operation and Development (OECD), International Monetary Fund (IMF) and World Bank via its partnership with Mimas.

C.3 Metadata production Overview

Creating comprehensive data documentation is easiest when begun at the onset of a project and continued throughout the research. It should be considered as part of best practice in creating, organising and managing data.

As for most archives it's very difficult to separate our workflows and approach between data and metadata as they are so co-dependant.

Types of Metadata in Play
A key challenge across the data lifecycle is the need to define and discuss overlapping concepts as though they are completely discrete; this support a structured approach but shouldn't blind us to the complexities of the situation on the ground.

The UK Data Archive general refers to three types of material in any data collection we hold:

Data: the microdata and/or aggregated data that was the original subject of collection

Data documentation: explains how data were created or digitised, what data mean, what their content and structure are, and any manipulations that may have taken place. It ensures that data can be understood during research projects, that researchers continue to understand data in the longer term and that re-users of data are able to interpret the data. Good documentation is also vital for successful data preservation. Good documentation for research data contains both study-level information about the research and data creation, as well as descriptions and annotations at the variable, data item or data file level.

Metadata: A subset of core data documentation, which provides standardised structured information explaining the purpose, origin, time references, geographic location, creator, access conditions and terms of use of a data collection. Metadata are typically used for resource discovery, providing

searchable information that helps users to find existing data, as a bibliographic record for citation, or for online data browsing.

But of course modern standards including the DDI-L (see 'Metadata Standards) can contain data and documentation and one researchers' metadata may be another's critical data for analysis, increasingly so with greater interest in and understanding of social network and administrative data/metadata as targets for analysis.

Similarly much of the wider material defined as 'documentation' would actual be amenable to increased standardisation and structure to align it with the more common understanding of metadata.

The UK Data Archive sees all the following classic 'types' of metadata as critical:

Descriptive metadata: It can include elements such as identifier title, abstract, author, and keywords. The Archive treats 'resource discovery' metadata as a partial subset of Descriptive metadata in general though all metadata 'types' are used for communications purposes, we're clear that one metadata element may fulfil several purposes, for instance identifiers are critical resource discovery and administrative metadata.

Our DDI records contain mandatory and optional metadata elements on:

- study description - information about the context of the data collection such as bibliographic citation of the study and data, scope of the study (topics, geography, time), methodology of data collection, sampling and processing, data access information, and information on accompanying materials
- data file description - information on data format, file type, file structure, missing data, weighting variables and software
- variable descriptions

Structural metadata: indicates how compound objects are put together, for example, how pages are ordered to form chapters.

Administrative metadata: provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of administrative data; three that are sometimes listed as separate metadata types are:

- **Rights management metadata**: deals with intellectual property rights,
- **Preservation metadata**: contains information needed to archive and preserve a resource.
- **Technical Metadata:** may refer to the technical processes used to produce, or required to use a digital object or to the file/format specific metadata that might be extracted from files which make

Again we have enormous potential for overlap here with key structural

metadata supporting granular resource discovery but these broad divisions seem to have stood the test of time.

A critical consideration in metadata design, creation, validation and quality insurance is whether the metadata is

- Manually generated
- Automatically generated
- Machine actionable.

Metadata Standards, Context and Purpose

Of the UK Data Archive is committed to the application of metadata standards wherever possible as increased standardisation supports metadata quality through greater automation, machine-readability, validation and re-use.

Standards

Data Documentation Initiative (DDI)

The Data Documentation Initiative (DDI) is a specification for capturing metadata about social science data.

The DDI is the standard metadata specification for social science data, and has a large and active international community of users and developers. It was:

- originally created to capture the information found in survey codebooks, which remains the focus of earlier versions
- used for basic study-level catalogue metadata and rich variable description for survey files
- maintained by the Data Documentation Initiative Alliance, a membership-driven consortium including universities, data archives, and international organisations.

We make use of the DDI 2.5 for our collection-level records in our metadata catalogue used by our *Discover* resource discovery system and in our Nesstar online data browsing system.

DDI-C

Both DDI-L (below) and DDI-C are Data Documentation Initiative products of the DDI Alliance http://www.ddialliance.org/Specification/

DDI-C (Codebook) is used at version 1.2 to align with the NESSTAR online data browsing product at http://nesstar.esds.ac.uk/webview/index.jsp. The core Archive database uses DDI 2.5 as its core reference. DDI2.5 is used throughout the deposit form and bespoke data ingest forms in use at the UK Data Archive.

DDI-L

The Archive is transitioning to supporting aspects of DDI-L which is designed to support the full longitudinal data lifecycle and has greater support for metadata below study level and re-use of metadata (i.e. less repetition). A

subset of DDI-L is currently used in UK Data Archive question bank work.

The DDI is an international XML-based descriptive metadata standard for social science data used by most social science data archives in the world.

At the Archive we use DDI to structure our catalogue records. The use of standardised records in eXtensible Mark-up Language (XML) brings key data documentation together into a single document, creating rich and structured content about the data.

Our staff sit on the DDI Technical Implementation Group, the DDI Controlled Vocabularies Group and the DDI Qualitative metadata working group.

SDMX
The Statistical Data and Metadata Exchange (SDMX) technical specification comes out of the world of official statistics (http://sdmx.org/). It aims to foster harmonisation and standards for the exchange of statistical information and has cooperating international organisations include the IMF, Eurostat, World Bank and OECD. We use this standard for the UK Data Service's aggregate databanks.

Dublin Core (DC)
DDI-C is mapped to DC at study/collection level and released alongside DDI-C via OAI-PMH

Text Encoding Initiative (TEI)
The Text Encoding Initiative (TEI) is a widely used metadata standard for describing textual documents. It is maintained by a consortium which collectively develops and maintains a standard for the representation of texts in digital form. It has many profiles but UK Data Service uses it for:

- structural mark-up of textual qualitative data
- three mandatory TEI header elements
- body elements: turn takers, paragraphs, headers
- inline tags: corrections, errors

TEI Header and TEI markup (http://www.tei-c.org/index.xml) is used in a subset of our qualitative datasets, primary for interview material and therefore forms a specialist part of the Ingest process.

QUDEX

Qualitative data standard developed by the UK Data Archive and Metadata Technologies. Currently being developed to support object and sub-object level metadata using the DDI, and using the Text Encoding Initiative (TEI) for encoding textual data. http://www.data-archive.ac.uk/media/387603/qudex_v03_01.xsd.

QuDEx enables discovery, locating, retrieving and citing complex qualitative data collections in context. The schema is complementary to

DDI and enables:

- highly structured and consistently marked-up data
- rich descriptive metadata for files e.g. interview characteristics, interview setting, type of object
- logical links between data objects: text to related audio, images, and other research outputs
- preserves references to annotations performed on data
- common metadata elements that enable federated catalogues across providers and borders

The standard is maintained by the UK Data Archive, University of Essex on the QuDEx site (http://www.data-archive.ac.uk/create-manage/projects/qudex). The UK Data service QualiBank (http://discover.ukdataservice.ac.uk/QualiBank) uses the schema, along with TEI and DDI 2.5

In 2013 the DDI Working Group on Qualitative Data group produced a detailed and complex model to accommodate the widest range of possibilities and use cases for qualitative data. It has not yet been implemented. QuDEx represents a very simple subset of this larger model

DataCite metadata Schema

We generate the basic DataCite metadata (http://schema.datacite.org/) alongside each update to our persistent identifiers (DOI). DOI are applied to all data collections made available via the UK Data Archive catalogue.

Controlled Vocabularies

The UK Data Archive strictly defines and control metadata through controlled vocabularies and wherever possible applies recognised controlled vocabularies and related standards.

The UK Data Archive uses the Humanities and Social Science Electronic Thesaurus (HASSET) and of its multi-lingual sister, the European Language Social Science Thesaurus (ELSST).

In summary:

- ELSST has been translated into nine languages, with three more on the way
- HASSET (http://www.data-archive.ac.uk/find/hasset-thesaurus/hasset-browser) has been in use and developed by the UK Data Archive over more than 30 years and is used for indexing data, at study level and variable level, and allows retrieval of data and related documentation using hierarchies of keywords
- in 2012 HASSET was converted to SKOS (http://www.data-archive.ac.uk/find/hasset-thesaurus/skos-hasset) , using the Pubby tool, containing 101,808 triples (at the relationship level)
- between them, HASSET and ELSST represent 7,695 concepts and 4,032

synonyms
- both thesauri follow ISO 25964: Thesauri and interoperability with other vocabularies (http://www.niso.org/schemas/iso25964/) as far as possible

Metadata Granularity
Descriptive metadata at the Study/Collection level is uniform throughout the core repository and the Self-Archive product. Metadata to variable level is available for collections included in the Nesstar product.

Metadata Roles and Methods
As a member of the CESSDA European Infrastructure we align a defined subset of DDI metadata fields with sister social science archives.

We collect the initial catalogue record information from our data collection deposit form (http://www.ukdataservice.ac.uk/deposit-data/how-to/regular/regular-depositors.aspx) , which is completed by the data depositor. We then enhance information from accompanying documentation to create a conformant metadata record. Where researchers can provide detailed and meaningful data collection titles, descriptions, keywords, contextual and methodological information in the deposit form, it helps us create rich resource-discovery metadata for their deposited collections. We assign key words from our own HASSET thesaurus (http://www.data-archive.ac.uk/find/hasset-thesaurus).

Depositors are encouraged to provide information about original and subsequent reports and publications or presentations based on our data collections so these references can be added as further documentation.

We prepare a standard bibliographic citation for each data collection so that users can correctly cite the data sources in research outputs. We believe that a well-documented high quality dataset deserves equivalent recognition and acknowledgement as published research outputs.

Different data collections are subject to different levels of curation and of quality assurance.

Metadata is critical to our Pre-Ingest team (for assessment/selection), Ingest teams (full spectrum metadata) and Access team (with special regard to resource discovery and access criteria). Our *Application Development and Maintenance* team undertake primary Data Management and our *Digital Preservation Systems and Security* team undertake primary responsibility for Archival Storage.

Pre-Ingest deposit metadata is controlled via an automatically validated deposit form completed by the depositor with further quality assurance of items not amenable to machine-validation. Some pre-Ingest metadata for grant-funded projects is automatically harvested from third party (ESRC) systems.

Ingest metadata is controlled via bespoke in-house applications and scripts which is either automatically validated or manually quality assured for items not amenable to machine-validation.

There are extensive procedures to support the Pre-Ingest and Ingest teams' creation of accurate metadata.

The Archive is currently planning a migration to a schema-neutral (but DDI-L compliant) database system with data entry (manual and automated) via a single, bespoke in-house system which will replace a number of in-house products. After this product is deployed we will make further changes to workflows and procedures to manage new forms of data, including administrative data.

C.4 Mapped to Data Lifecycle

Comments about modelling standard lifecycle elements have been extracted to the separate issue document *DASISH-T5-3-Issue-MetadataLifecycle-v*.

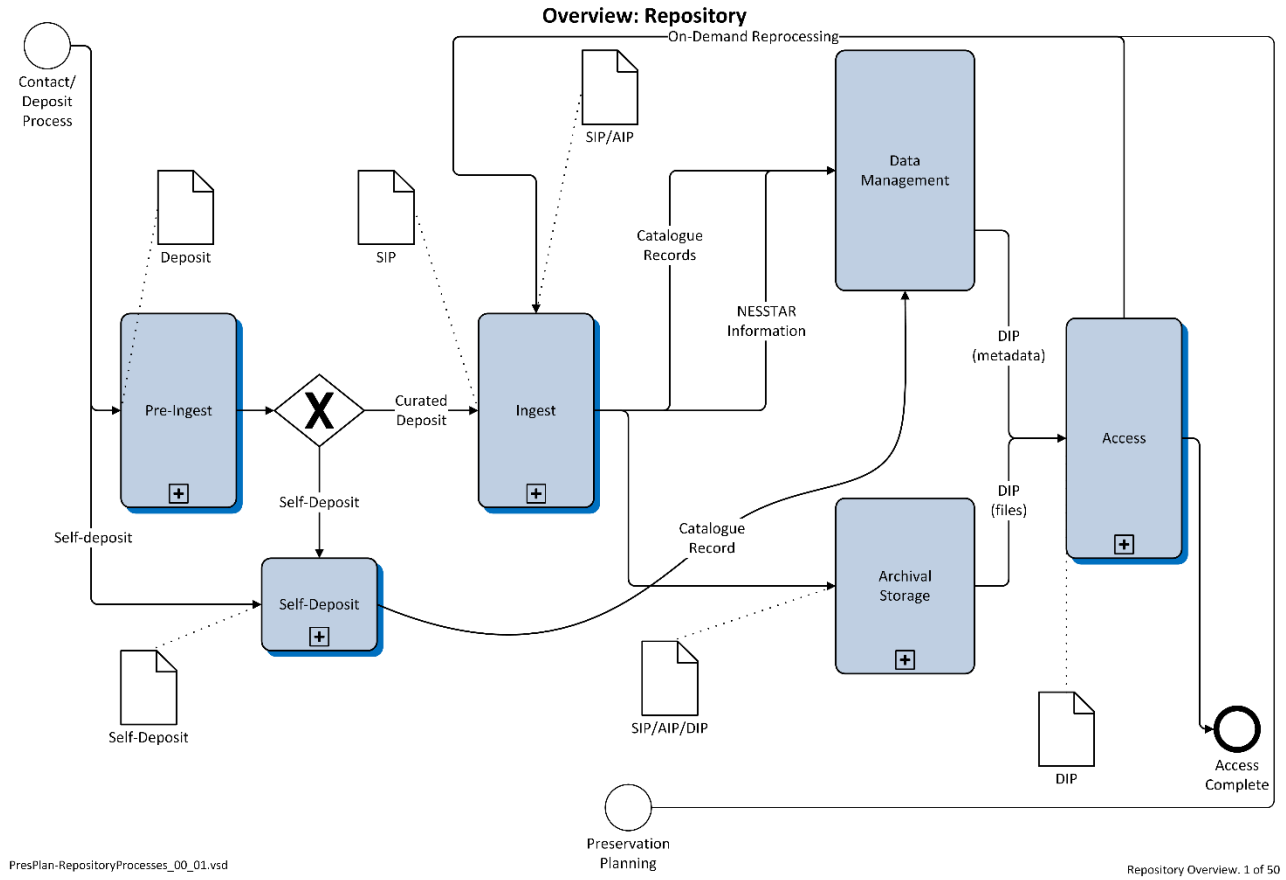


Figure 25: Repository Overview (Case Study UK Data Archive)

Lifecycle Planning

Some planning activities continue in parallel to the data/metadata creation and amendment lifecycle, primarily guiding an understanding of that lifecycle in terms of stakeholders (community) and preservation goals as well as guiding the design and implementation of actions and events which recur throughout

the lifecycle.

Community Watch and Participation
This activity covers our interactions with our community including direct contact and training events. The community of stakeholders includes data producers and data users as well as funders and government departments. It also includes a degree of technology watch as we monitor use of standards (including metadata standards) software and formats within the community.

The UK Data Archive monitors the Designated Community (OAIS) continuously but could usefully make the outputs easier to consume. This might not be directly associated with metadata quality but this monitoring could certainly trigger changes in the approach to metadata including new standards, new technologies, increased granularity, other formats etc.

We could usefully improve our community watch and participation outputs by collating information from our producer relations interactions and contacts (surveys, questionnaires, workshops) with users alongside other analysis and reporting information (see Monitoring, Appraisal and Disposition 0).

Data Management Planning/Preservation Planning
The term 'data management' planning incorporates metadata planning in the pre-archival phase of the data/metadata lifecycle. Preservation Planning refers to the OAIS Preservation Planning function and also covers OAIS Administration functions.

The Archive works with Data Producers directly and in more general training scenarios through our Producer Relations section to support good data/metadata practices prior to the archival phase of the lifecycle and we also maintain with contacts with funders who have a role in requiring and defining data management plans.

We have more direct control over the archival phase of the data lifecycle, Information from Community Watch and Participation (0) and from more general Technology Watch (See Monitoring, Appraisal and Disposition (0)) is incorporated to inform the development of repository standards and strategies covering data and metadata including fixity, preservation events, legal and rights metadata impacting preservation and access. This work, includes developing packaging (SIP, AIP, DIP designs) and plans for migration so may require the redesign of data and metadata management practices (See Metadata design/redesign(0)) and ultimately managed change to data/metadata (See Metadata Change and Change Management (0)).

Recurrent Actions and Events
The design, redesign, change, change management, monitoring appraisal and disposition, as well as standards for custody transfer are all driven by data management/preservation planning and community and technology watch processes.

Custody Transfer
Any change of custody is a high risk point for data and metadata.

Custody transfers are recurrent events in the sense that they may occur numerous times during a digital objects' lifecycle. The key relevant custody transfer for the UK Data Archive is deposit which takes place during pre-Ingest but there are arguments for applying some custody transfer good practices at each 'internal custody transfer' during the repository workflow including from pre-Ingest to ingest, from ingest to Archival Storage.

While some archives simply take a single clear custody transfer for deposit (e.g. a data dump from an earth observation satellite) the UK Data Archive maintains close communications with depositors of government and academic data and multiple files or versions of files (data and metadata) may be submitted over time, This improves the speed of ingest and availability and improves the quality of supporting metadata as revisions are made but does require a more 'porous' boundary between depositor and archive.

Standard metadata is required for each 'deposit' but decisions over the level of control and documentation required for each 'batch' of information deposited must take account of the administrative overhead on staff.

Monitoring, Appraisal and Disposition
Various individuals or organisations may have responsibility and processes (more or less formal) to monitor metadata, appraise it as fit for purpose (or otherwise) and trigger disposition decisions (e.g. retain, improve or delete).

Historically the Archive has developed internal software analogous to the OAIS functions Ingest (0) and Access (0) with the addition of a Pre-Ingest function analogous to the PAIMAS processes (Producer-Archive Interface Methodology Abstract standard)73 This has separated data and metadata use across the functions which has a tendency to build knowledge silos which become less tenable as we take a more lifecycle-aware approach to software development and repository planning.

Current work to update software will support more accessible monitoring of archival practice to support appraisal and disposition.

Considering the monitoring, appraisal and disposition of all important repository assets (data, metadata, records etc.) as a whole improves the internal knowledge base and reduces communications barriers.

Monitoring
Ongoing review of existing metadata standards and practices including internal archival practice and guidance provided to depositors earlier in the metadata lifecycle.

[73] http://public.ccsds.org/publications/archive/651x0m1.pdf

Software systems under development will provide centralised collation of all repository processes to support better analysis and reporting.

To improve our responsiveness to our stakeholders and streamline our activities we will move to provide common access to business intelligence such as business process information, archival storage monitoring, community watch and participation, impact assessments etc. These will be aligned with our ISO27000 Information Security certification processes.

A critical part of the input into data/metadata related decisions is 'Technology Watch' undertaken by our technical services section.
Appraise/Reappraise
Reappraise: "Return data which fails validation procedures for further appraisal and reselection." (DCC)

In addition to the appraisal of 'offers' of data for deposit the monitoring of our archival storage will support appraisal of formats currently stored with regard to their fitness for preservation (format risk) and their fitness for use by the Designated Community. Such appraisal/reappraisal processes might trigger metadata design/redesign (in turn triggering a metadata change), a custody transfer or a managed change.

Metadata reappraisal would normally be a review of our metadata profile as metadata standards are changed. Business records, business processes and our metadata profile are also subject to monitoring and appraisal.

The outcome of an appraisal is a 'disposition' decision.

Disposition
Disposition does not only imply the destruction of data or metadata, the appraisal process may include decisions to take no action or may imply the need to make a managed change. In the case of metadata containing personal data there may be a legal requirement for destruction or secure destruction which will be taken in line with our retention schedule for business records.

During the appraisal of an offer of data during pre-ingest data may be rejected, selected for the fully curated collection or directed to our self-archiving system.

At the UKDA the 'delete' issue seldom impacts us but we do have de-cataloguing procedures and de-archiving procedures and a decision to migrate metadata to a new format or to amend and existing format implies that we follow metadata change management procedures.

Metadata Design/Redesign
Metadata design/redesign applies to the archives overall approach to data/metadata which is then implemented during pre-ingest, ingest and access processes as managed changes (0) to particular data collections.

New metadata designs may be required to handle new data formats or data from new disciplines while metadata redesign may be necessary to meet the needs of the stakeholders to respond to new or updated metadata standards which form part of our workflows. A decision to redesign metadata may be triggered as a result of Community/Technology Watch. If a redesign implies change this will follow change management practices during metadata versioning or data migration.

The UK Data Archive was a participant with the DDI Alliance in the original DDI standard and current database systems are strongly aligned with DDI version 2.1 elements. We have now transitioned to DDI2.5 which is intended to be a transitional schema between DDI2 (Now DDI-C for collection) to DDI3.x (Now DDI-L to reflect a greater focus on the full lifecycle.

During the ongoing development of new metadata management systems we will be taking an approach which is 'schema-agnostic' (with the word schema applying to a particular XML or other set of metadata elements rather than in the sense of 'database schema') letting us support a wide range of metadata standards within the database.

A metadata profile manager will support the alignment of elements in the database with elements in one or more relevant metadata standards to support the export of XML for harvesting (including via OAI-PMH) and as 'snapshots' to be retained alongside data in the archival storage system.
Changes to the profile, and by extension to the database fields supporting the relevant elements, will be subject to formal change management requests (0. Controlled vocabularies will be managed through an analogous process.

Metadata Change and Change Management
Change is inevitable for metadata as it is for data and implies the need for versioning which is critical to quality, not least in terms of authenticity and provenance.

Procedural changes at the UK Data Archive are managed through the records management of 'controlled documents' which are formally reviewed and approved by a Governance Oversight Committee or by an Information Security Management Group if they have information security implications. In future repository business processes will be more formally managed.

The UK Data Archive provide for minor changes (no DOI update) and major changes (DOI update) in data collections both of which are accompanied by a change log. Major/minor changes decisions are made by the Ingest team based on their expected impact on users but a new DOI may be minted for a data or a metadata change.

In addition to change and change management of metadata on individual collection items changes to our overall metadata profile must be managed, these are currently applied alongside software development and maintenance practices but once new repository software development is completed change

requests for the metadata profile and controlled vocabularies will be made via an issue tracking system and assessed for repository-wide impact before approval and implementation. Approval will involve a repository developer, the preservation planning manager and the standard manager and Relevant experts from all UK Data Service sites across Pre-Ingest, Ingest and Access will be consulted.

Projects and development work will be asked to include any variations from the current profile in their proposals.

Maintenance requests via an issue tracker (JIRA) will be asked to flag requests as having a potential impact on the metadata profile. Developers will be asked to add a flag if it turns out a change to the metadata profile is required.

Once standard compliant metadata snapshots are produced it will be critical that changes don't impact their validity. It will be possible to enforce change management procedures for metadata called directly by products from the central database but for products using other databases we will need to rely on cooperation from developers of other in scope systems to keep their changes synchronised with the metadata profile.

Migrate
The UKDA addressed migration during the ingest process (to acceptable preservation formats) but also during 'on demand reprocessing' which might occur as a result of a user request or if a format is designated as at risk (for preservation purposes) or as no longer the best format for the designated community.
Migration of a format may imply migration of metadata, or metadata may be migrated independent of a format or data change.

Sequential Actions

Conceptualise
The UK Data Archive works to improve overall Data Management in the stages prior to ingest through its Data Management and Sharing materials and workshops which include guidance on creating high quality metadata records for data that comply with international social science archival practice.

For data generated as a product of research there is often a need to seek funding which may imply additional steps (see below) but for other data which are of relevance to research there may have been little or no consideration of the wider metadata needs or the lifecycle of the data beyond original conception.

Examples:

Social Media data: metadata is designed to support the functions of the service and to report on behaviours which will inform future service provision or monetisation of the service.

Administrative data: metadata is designed to support the business processes identified and any critical analysis and reporting at a higher level. Metadata to support late research may not be an issue.
Funding: for data derived from funded research it is increasingly likely that there will be a Data Management Plan including plans for metadata, which are a requirement for funding. As the UK Data Archive is funded by the ESRC we may have some influence over their data management plan metadata requirements but this is not the case for most research data we receive.

As noted above the metadata will simply be a by-product of the immediate need to support the collection/reporting processes so funding (and planning) is less likely to be addressed independently of planning around the data. The cost of metadata creation is part of the overall costs and not split out.

Automated integration of funding metadata is part of our existing processes where possible.

Create/Capture
The initial creation of metadata can also be defined as the first point of capture.

In an ideal world we have a creation process that follows all of the stages and actions defined in the conception stage. But during the generation of data there will almost always be a need to change or revise the data/metadata structures. In a creation environment it remains unlikely that these changes and the justifications and agents involved will be recorded to a level which aligns with best practice for authenticity and provenance.

For the UK Data Archive we need to take account of the very limited influence we have over the creation stage of the process but of course we will 'enrich' the metadata after initial Capture.

Custody Transfers
The Archive must take into account that there may have been several custody transfers prior to 'deposit' at the Archive. Though improved information about the standards applied during these processes would improve overall provenance we accept that this is unlikely in the immediate future.

For UK Data Archive see Pre-Ingest and Deposit below, for inclusion in a wider definition of custody transfer see 0 above.

Producer/Archive Interface
Approaches differ depending on whether the data producer is an ESRC Award Holder, a regular depositor or a new depositor http://ukdataservice.ac.uk/deposit-data/how-to.aspx.

At the UK Data Archive we refer to all points from the initial contact with a potential depositor (who may be the Data Producer or another custodian) as 'Pre-Ingest'.

Contact detail metadata may be the first collected, then sufficient metadata to support Appraisal and Selection. The Pre-Ingest process is effectively a negotiation with the depositor which includes descriptive and rights metadata as well as the appropriate licences at the deposit stage. A pre-ingest negotiation is managed via an 'Acquisition identifier' which, if successful will be mapped to a Study Number (data collection identifier).

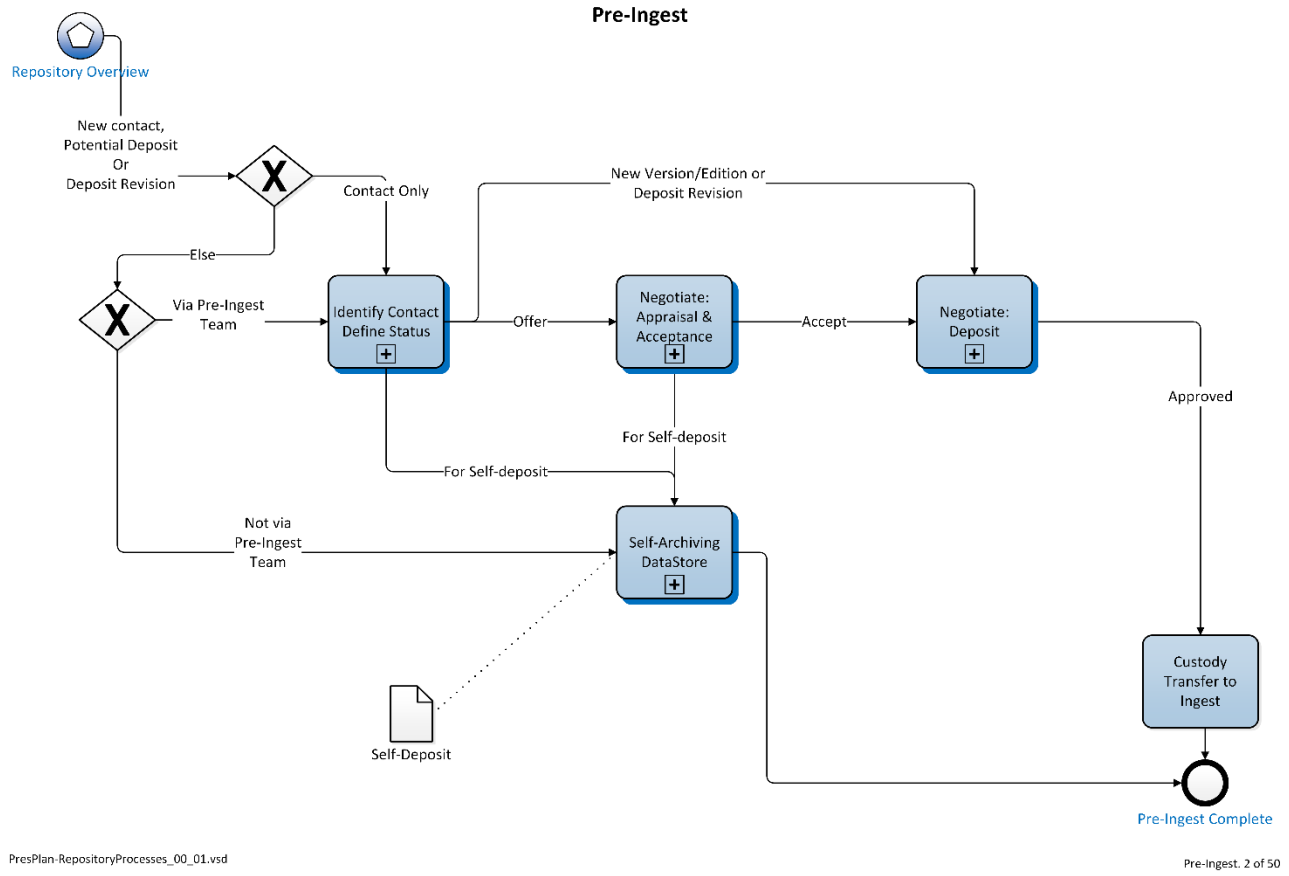


Figure 26: Pre-Ingest (Case Study UK Data Archive)

UK Data Archive Pre-Ingest Team key metadata-related issues

Strengths:

- High quality metadata from knowledgeable researchers who understand the importance of metadata
- Archive easy to use deposit form/metadata creation tool. Make sure what's typed in matches the collection
- Good data practices throughout research avoids doing the metadata afterwards
- Clearly defined standard metadata profile applicable across the board
- Validation on form entry fields
- Research stage: systems that capture the metadata early
- Integrated systems because cutting and pasting between systems is inefficient
- Controlled vocabularies of entries for metadata/input programmes promotes standardisation but we also need to have flexibility to let us evolve over time.

Weaknesses:

- Bad data practices. Researchers not knowing or understanding the value of good metadata practice
- Longevity: there can sometimes be a lack of foresight over what we'll need in the future.
- (not just what we need now)
- Too much focus on one discipline rather than looking at other disciplines to make sure your metadata profile is widely applicable Version control is challenging
- Time series metadata risks getting bulky with add-ons but there are administrative overheads in constant re-editing.
- RELU project suffered from a weakness that the study level was not necessarily the best level to describe an object (as opposed to project level or data collection level or survey level
  - implies need for clear object model.

Opportunities:

- Improve capture methods
- Data citations and data publishing promotes and incentivise good as they showcase good researcher metadata
  - Ties in with researcher training
- Development of mining systems, data citation index and registries
- Portability of metadata.

Threats:

- Time delays between the end of a project and the arrival of information
- Capture tools
- Metadata pulled from one system to another causes reduction of quality, precision and errors including transcription errors (special characters etc.)
- Metadata provided on behalf of another e.g. Government departments with one representative implying dilution of information and information 'lost in translation'
- Tendency to assume metadata can be copied directly from earlier parts of a time series
- Pro-active approach to creation a metadata record as part of the creation process, even with a repeated survey would be best.

Appraisal and Select

When data are offered to the Archive sufficient metadata is necessary to for the Data Appraisal Group to:

Understand the content, scope and value of the data being offered: Ensuring it aligns with our core mission and to define whether it should be rejected or accepted into the fully created collection or our self-archive.

Define the rights and access criteria: Ensuring the data is not over-encumbered with rights issues and can be accessed by appropriate methods

which may range from open access to permission only access via a secure server or safe room.

Deposit
For the UK Data Archive the critical Custody Transfer process is a 'Deposit' which is the result of a Data Appraisal Group decision. A deposit may be as a result of an 'offer' of data or as mandated as a condition of project funding.

Deposit metadata is controlled via required and optional fields in a data deposit form which are machine validated and undergo manual quality assurance.

A deposit may be a single event for a data collection or it may be a longitudinal collection with planned deposits over time. The latter case triggers the need to handle versions and editions and to 'clone' metadata from previous deposits to avoid repeated data entry while permitting entry for metadata which changes from deposit to deposit.

For some archives the deposit process is strongly focussed on the integrity of the deposited object, e.g. for large scale physics experiments or earth observation satellite data large quantities of data are deposited and the primary goal is to ensure they remain unchanged. In contrast, at the UK Data Archive the deposit process may be more porous with corrections, changes or enrichments to an original deposit arriving over time either un-prompted or as a result of quality assurance work undertaken by the pre-ingest or ingest teams. While increases the value of the data it also increases the complexity and cost of the producer-archive interface as metadata inevitably changes.

Ingest
Once all initial quality assurance are addressed the deposit is designated a formal Submission Information Package or SIP (OAIS). Throughout this period study metadata is entered in line with DDI2.5.

As well as quality assurance extensive 'curation' occurs. The data may be enriched, and standardised to a level which aligns with our standards. Changes to copies of the deposited data collection may be made either via a re-deposit or via agreement with the depositor.

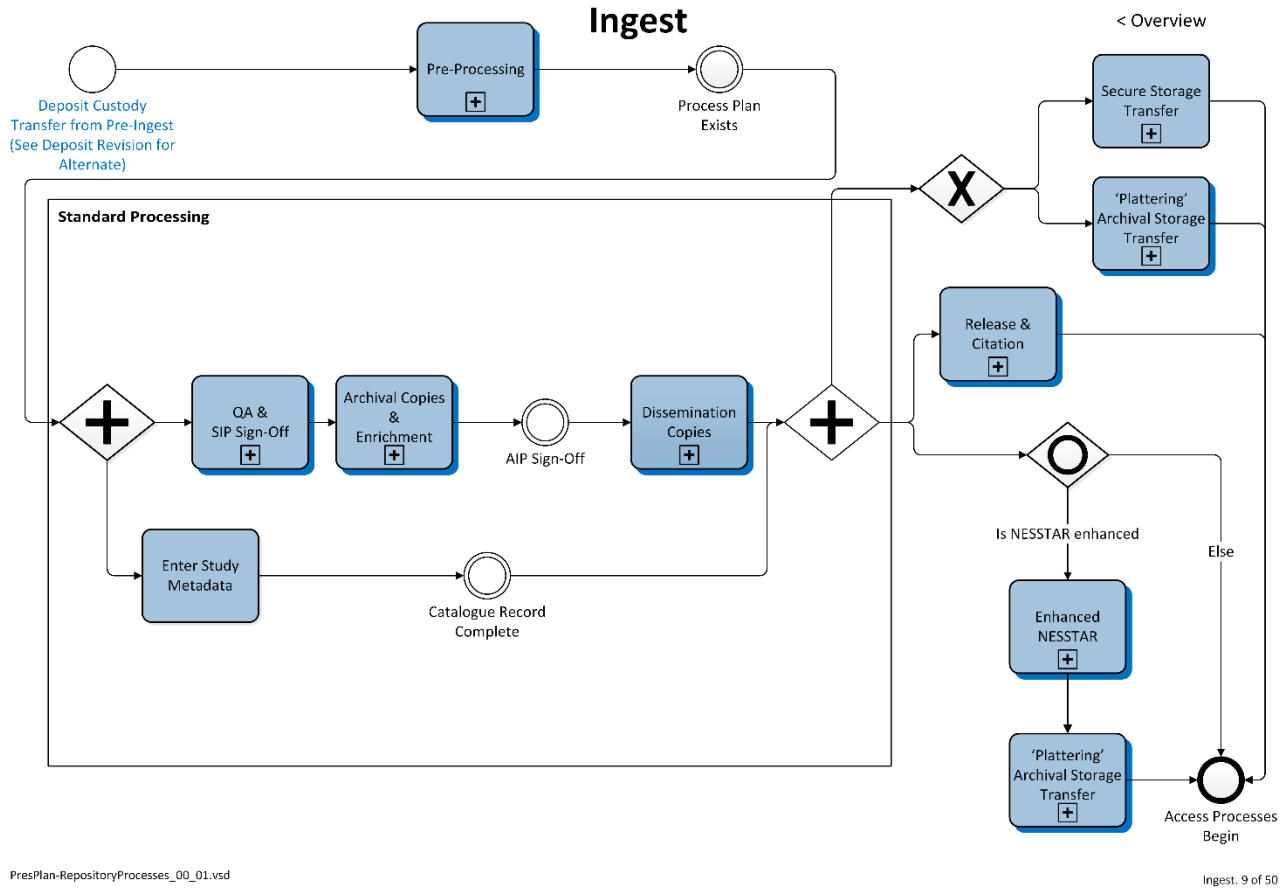


Figure 27: Ingest (Case Study UK Data Archive)

http://www.data-archive.ac.uk/curate/archive-quality

UK Data Archive Ingest key metadata-related issues

Strengths:

- All of our systems have adhered to some form of relevant standards for as long as they have existed for both 'standardisation' and interoperability purposes
- We monitor new standards as they appear and have a good international perspective; limiting to a national view isn't appropriate tin this area
- Flexibility including our ability to deal with new types of access (Secure and Open) and cross-disciplinary work (e.g. RELU) which lets us enhance our practices to deal with the needs of different kinds of researchers.

Opportunities:

- Improve our understanding and descriptions of the objects we curate through alignment with current international projects on metadata modelling (including lifecycle modelling) such as SDMX and DDI.
- Benefit from the expertise of others through this and through organisations like CESSDA.
- Such alignment improves opportunities for automation
- Big Data on the horizon implies new kinds of metadata, different

expertise and potential changes to our designated community
- Social Media provides new kinds of metadata and needs for metadata management.

Weaknesses:
- Flexible version control is challenging, users need to understand what changes have been made but curators need the space to differentiate between a spelling mistake and a policy-impacting statistical error
- outdated bespoke systems make it harder to adapt to new metadata (mitigation is a schema-neutral approach).

Threats:
- A tendency (especially in a world where academic funding is cyclical and metadata experts are often focussed on this) to try to be all things to all people creates a tension that can lead to an over complex approach to metadata
- A simplest is best approach simplifies things for users and minimises the long term maintenance burden.

Pre-Processing
The Acquisition ID is mapped to a new or existing Study identifier. A standard directory structure is developed to align with our Archival Storage system; the physical structure is expected to be replaced by offset metadata mapped to files over time. A processing plan for the data collection is created and initial quality assurance is undertaken.

Preservation Action
Migration to preservation formats in undertaken and the events recorded. Our goal is to increase the standardisation of such preservation action descriptions i.e. reduce prose and increase the granularity and reduce the prose element of such descriptions. This reduces the chances of human error and reduces the overall workload.

Store
An Archival Storage Transfer is treated as a custody transfer within the Archive. Any previous editions/versions of data collections are moved and new, approved versions editions are integrated and pushed out for dissemination.

Extensive metadata is in place within the Archival Storage system to manage multiple copy integrity.

There is clearly scope to align Archival Storage best practices with the earlier Store actions in the metadata lifecycle.

Access
Relevant descriptive metadata to support resource discovery and selection is pushed to our 'Discover' catalogue search system.

Dissemination Information Packages are pushed from the Archival Storage System (for objects made available via digital downloads or made available for online browsing, or accessed via secure servers or safe rooms. In all cases 'Access Requests' are driven from the Discover system.

Extensive access criteria are attached to each data collection which interact with variables attached to user accounts when managing access.

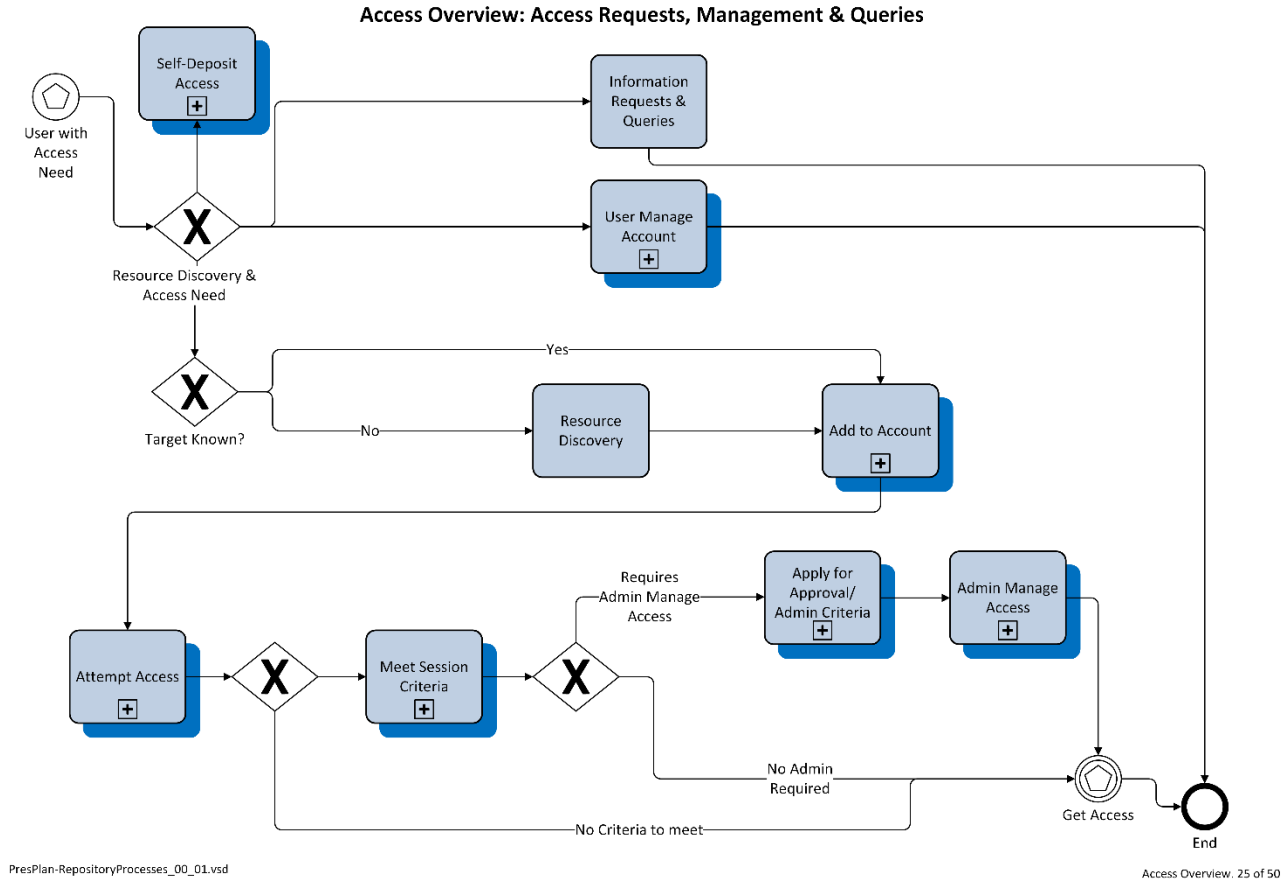


Figure 28: Access (Case Study UK Data Archive)

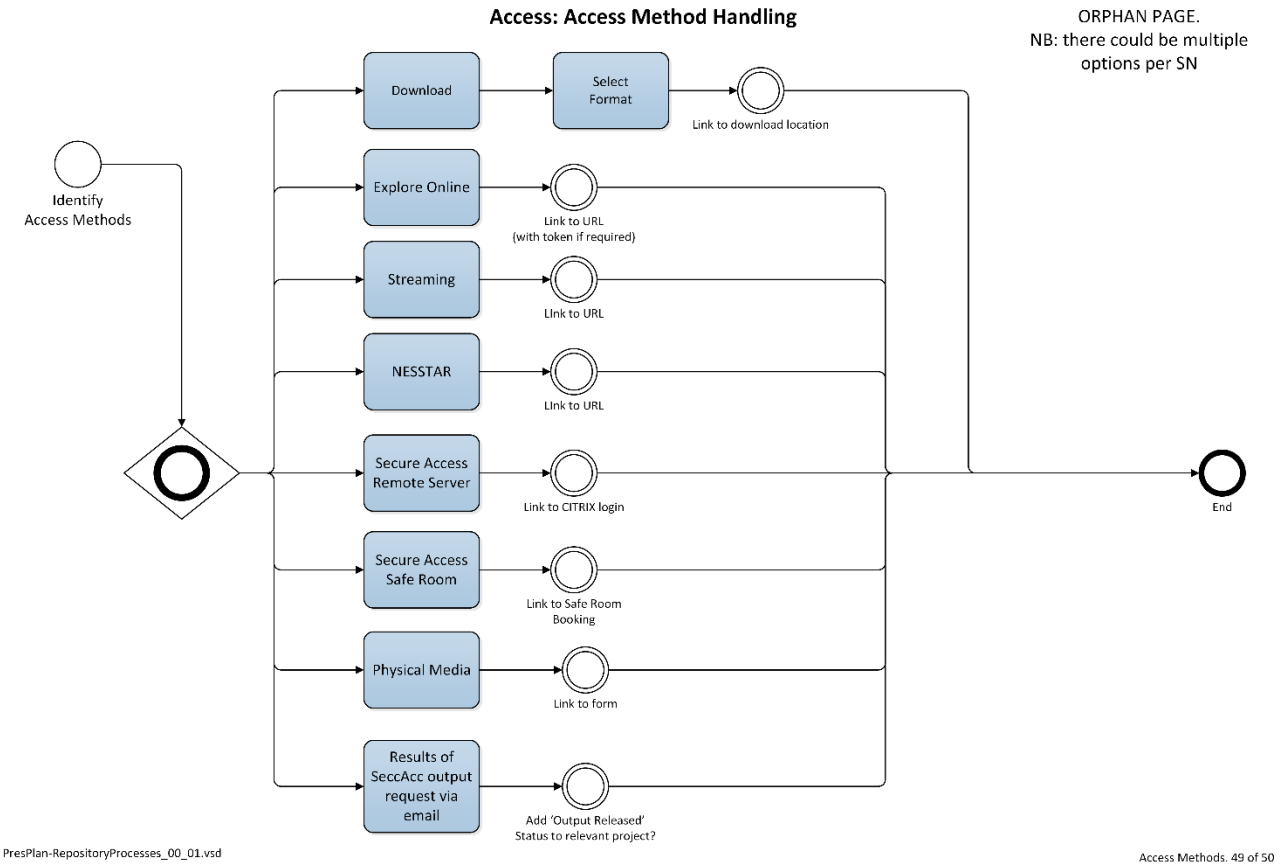


Figure 29: Possible Access Methods (Case Study UK Data Archive)

UK Data Archive Access key metadata-related issues

- Identity is a challenge when dealing with registrations but a federated service like Shibboleth where participating institutions have a closer link to the individual increases confidence that we know who we're dealing with
  - We can't meet everyone face to face and ask for a passport so there's always a small risk around identity but we haven't recorded an incident in this area.
  - Without global use of digital signatures we will retain the need for physically signed licence objects which must in turn be managed.
- The application of access criteria is always complex as this revolves around interactions between a user account/status (variables about the user) and conditions applied to the particular dataset (variables around the data collection)
  - A structured approach such as describing rights and restrictions in open digital rights language (ODRL) might simplify the link between intention and technical implementation.
- Considering the long term implications of access and rights negotiations is also critical. We need to consider the impact on resources and also the impact of for example, the death of a key contact on our ability to offer access or renegotiate access criteria over time
- We can be constrained by the original conditions and find ourselves without an easy route to renegotiation.
- A strict approach to rights and access criteria (i.e. minimal number of

conditions as simple as possible) minimises risk
- A loose approach to rights and access criteria (i.e. per study negotiations, a proliferation of criteria) increases complexity, increased resources required and makes it much less likely that a 3rd party system will meet the organisations needs which could necessitate a complex and resource hungry bespoke information management system

Use and Re-Use

We need to retain contact with the data/metadata at the point of use/re-use, either through monitoring citations of our data including through the number of times our DataCite DOIs have been resolved, or through direct contact with the depositor. This supports efficient re-capture of amended metadata when derived research is redeposited.

Example Catalogue XML

<stdyInfo>
<subject>
<keyword>
</keyword>
<topcClas Vocab="unknown">Economic processes and indicators - Economics</topcClas>
<topcClas Vocab="unknown">Economic systems and development - Economics</topcClas>
<topcClas Vocab="unknown">General - Employment and labour</topcClas>
<topcClas Vocab="unknown">Elites and leadership - Social stratification and groupings</topcClas>
<topcClas Vocab="unknown">Management and organisation - Industry and management</topcClas>
</subject>
<abstract>This project aims to develop knowledge and understanding of the contemporary globalization of the headhunting industry in Europe and its implications for new forms and geographies of executive search and selection. Europe has become the most complex and sophisticated pan-regional market for executive search, fuelled by free labour mobility within the EU, thereby offering a unique environment in which to study the changing practices of the headhunting industry.</abstract>
<sumDscr>
<nation>UK</nation>
<geogCover>London</geogCover>
<nation>France</nation>
<geogCover>Paris</geogCover>
<nation>The Netherlands</nation>
<geogCover>Amsterdam</geogCover>
<nation>Germany</nation>
<geogCover>Frankfurt</geogCover> <nation>Belgium</nation>
<geogCover>Brussels</geogCover> <universe>Executive search consultants, researchers and associations in London, Paris, Frankfurt, Amsterdam and Brussels, 2006-2007</universe>
<anlyUnit>Individuals</anlyUnit>
<anlyUnit>Institutions/organisations</anlyUnit>
<collDate event="start" date="01/2006">January 2006</collDate>
<collDate event="end" date="08/2007">August 2007</collDate>

<timePrd event="start" date="01/1980">January 1980</timePrd>
<timePrd event="end" date="08/2007">August 2007</timePrd>
</sumDscr>
<notes>
</notes>
<method>
<dataColl>
<sources>
<dataSrc rule="Sources used">The Executive Grapevine, The Directory of Executive Recruitment, published by The Executive Grapevine International Ltd. Editions consulted: 1980, 1985, 1990, 1994, 2000, 2005</dataSrc>
<dataSrc rule="Source location and access">Copies are held at the British Library and the most recent edition is available for private purchase.</dataSrc>
</sources>
<collMode>Face-to-face interview</collMode>
<collMode rule="Other">Time series for search firm and office data collated from the Executive Grapevine Directories of International Recruitment</collMode>
<sampProc>Purposive selection/case studies</sampProc>
<timeMeth>Cross-sectional (one-time) study</timeMeth>
</dataColl>
</method>
<othrStdyMat>

Appendix D: Case study DANS

D.1 Introduction

This deliverable about metadata quality improvement describes in sections 5 to 7 the different metadata strategies of CLARIN, DARIAH and CESSDA and possibilities for cross fertilisation. To illustrate how this works at the repository level, we performed 4 case studies, which describe in more detail the data management at the individual institutes. This case study describes the situation at Data Archiving and Networked Services (DANS) in the Netherlands. DANS is special in that respect, in that it is engaged in several disciplines and therefore takes part in all three infrastructures: CLARIN, DARIAH, as well as CESSDA. This involvement in multiple disciplines influences the metadata management procedures at DANS.

This case study starts with a general description of DANS, followed by a description of metadata management. Coming developments will be discussed, followed by plans to enhance the quality of metadata.

E.2 General Information about DANS

History

DANS was established in 2005. The institute is a successor of four Dutch organisations in the field of data archiving and data dissemination in the Netherlands: the Steinmetz Archive in the field of the social sciences, the Netherlands Historic Data Archive (NHDA), the Scientific Statistical Agency (WSA) and the Electronic Depot of the Netherlands' Archaeology (EDNA). The collections and activities of these predecessors were transferred to DANS.

In 2011, NARCIS – the National Academic Research and Collaborations Information System – became a service of DANS.

Organisational context

DANS is an institute of the Royal Netherlands Academy of Arts and Sciences[74] (KNAW) and the Netherlands Organisation for Scientific Research[75] (NWO). As the forum, conscience, and voice of the arts and sciences in the Netherlands, the KNAW promotes the quality of scientific and scholarly work and strives to ensure that Dutch scholars and scientists make the best possible contribution to the cultural, social, and economic development of Dutch society. NWO funds scientific research at Dutch universities and institutes by means of more than a hundred different types of grants.

National and International Infrastructures

[74] https://www.knaw.nl/en

[75] http://www.nwo.nl/en

DANS takes part in several projects and infrastructures that aim to promote the scientific data infrastructure in the Netherlands and Europe. An overview of the international collaboration is available on the DANS website[76].

Mission

DANS promotes sustained access to digital research data. 'Digital research data' is understood to mean: research information, research data (such as databases, spread-sheets, text, images, video and multimedia) and digital publications (including preprints and reports). For this purpose, DANS encourages researchers to archive and reuse data in a sustainable manner.

As part of its mission, DANS supports the Open Access principle, while being aware of the fact that not all data can be freely available and without limitations at all times. Even so, it is important that research data that are not available (yet) or only available to a limited degree are archived in a sustained manner. Therefore, DANS adheres to the principle 'Open if possible, protected if necessary'.

Information related to the institutional identity of DANS is published on the website[77] and in the DANS Strategy Policy 2011-2015[78].

Main activities

DANS is primarily a service institute. DANS provides services in the fields of archiving, reuse, and training and consultancy. In addition, DANS performs research into sustainable access to digital information and takes part in national & international projects and networks. The English summary of the DANS Strategy Policy 2011-2015 describes in more detail the content of these services[79].

DANS provides services for data access and preservation for the social sciences, humanities and adjacent disciplines. For this purpose the online archiving system EASY[80] was developed. EASY presently contains more than 27.000 datasets for the disciplines of archaeology, social and behavioural sciences, history and geospatial sciences.

DANS also provides access, through the national portal for scientific information NARCIS[81], to more than 30.000 datasets, almost 900.000 e-publications and other research information in the Netherlands.

To ensure that archived data can still be found, accessed and used in the

[76] http://dans.knaw.nl/en/content/about-dans/more-information/international-co-operation

[77] www.dans.knaw.nl

[78] http://dans.knaw.nl/content/strategie-en-beleid

[79] http://www.dans.knaw.nl/sites/default/files/file/jaarverslagen%20en%20strategienota/Samenvatting%20strategienota_UK_DEF.pdf

[80] https://easy.dans.knaw.nl/ui/home

[81] http://www.narcis.nl/

future, DANS co-founded and participates in the Data Seal of Approval consortium[82]. This data seal can be requested and granted to data repositories that meet a number of clear criteria in the field of quality, preservation and accessibility of data.

Number of employees
More than 50 people work at DANS on a permanent, temporary and/or freelance basis. An overview of staff members and their contact details can be found on the DANS website[83].

The archive department is responsible for data management. Ten employees of this department are engaged in data management for (a small or bigger) part of their regular activities. The total capacity for data management is 2 full-time equivalent (fte). Additional work on data management is being done for special projects.

D.3 Metadata production

Concept of self-archiving
DANS applies the concept of self-archiving: data producers deposit their research data with DANS themselves using the deposit service in the EASY archive[84]. During the depositing procedure the users are being guided through a couple of forms, in which they have to document their data. Not all forms are mandatory. At the end of the procedure, the user is asked to assign the access level of the data set, to accept the License agreement and to upload his data. DANS's data managers can provide assistance during the depositing procedure.

After the data have been deposited in EASY, a data manager at DANS will process them in accordance with a standard data processing protocol. The purpose of this protocol is to ensure that the data will be findable, accessible, and understandable for re-use in the longer term. A key element of this protocol is, where applicable, the verification of privacy-sensitive data. This applies, in particular, to survey data and interviews. On the basis of this protocol, the following types of verification are performed:

- Verification of completeness of the dataset, with regard to both the data files deposited and the accompanying documentation files;
- Verification of the readability of the files;
- Verification of the file format. In the future, it should still be possible to open and use the data files, as well as the documentation files. The verification is performed on the basis of a list of preferred file formats.
- Verification of the description of the dataset for completeness and accuracy, and improvement of the presentation of the description.

[82] www.datasealofapproval.org
[83] http://www.dans.knaw.nl/en/content/contact/staff-members
[84] https://easy.dans.knaw.nl/ui/home

- Verification of the presence of privacy-sensitive data, both in the files and in the metadata. If necessary, anonymisation of privacy-sensitive information is carried out. Verification for the presence of informed consent forms for privacy-sensitive data.
- Verification of the clarity of the directory structure. If this structure is not sufficiently clear, it will be adjusted.
- For archaeological data: Verification of completeness and correctness of the list of files. This list includes a short description of each file of the dataset.

Upon archiving, an automatically generated Persistent Identifier is attached to each dataset. The Persistent Identifier enables identification of the dataset, independent of its location (on the web). DANS assigns URN-NBN[85] identifiers for preservation purposes. Because of the high demand for data citation using DOI, DANS will start implementing DOI identifiers from DataCite[86] in 2014.

After processing by a DANS data manager, the datasets are published. All actions performed by the data manger are documented in the administrative metadata.

Certain research projects have an obligation to deposit data as part of from their funding agreement, for example, research funded by the Netherlands Organisation for Scientific Research (NWO). Dutch archaeology adheres to a national regulation (the Kwaliteitsnorm voor de Nederlandse Archeologie, KNA quality norm), which determine that all digital documentation from archaeological research projects needs to be deposited for long-term preservation. Apart from the obligatory deposits, other researchers are welcome to deposit their research data at DANS.

Types of metadata that play a role within the work processes of DANS

Descriptive metadata
A dataset in EASY is disclosed by terms from the EASY Metadata schema[87] (EMD). This schema is based on the Dublin Core Metadata Initiative (DCMI) terms[88], with some additional options from Qualified Dublin Core, additional granular fields for creator and contributor, and additional geospatial information. EMD is mainly descriptive metadata. The metadata is completed by the data depositor and checked by a DANS's data manager before the dataset is published.

Some research institutes have special agreements with DANS for the automatic depositing of data and metadata into EASY. For this translocation DANS uses SWORD[89], a lightweight protocol for depositing content from one location to another. The DANS SWORD protocol makes use of the DANS Dataset Metadata

[85] https://wiki.surfnet.nl/display/standards/URN-NBN
[86] https://www.datacite.org/whatisdoi
[87] http://easy.dans.knaw.nl/schemas/md/emd/2012/11/emd.xsd
[88] http://dublincore.org/documents/dcmi-terms/
[89] http://swordapp.org/

(DDM) Schema[90]. EMD and DDM contain the same elements, although EMD is more application specifically oriented.

Contextual metadata
The content of contextual metadata is more discipline specific. Because DANS serves several disciplines, the approach regarding this kind of metadata varies by discipline. In most cases contextual metadata is available in the form of codebooks. These codebooks are uploaded in EASY as Portable Document Format (PDF) files within the dataset.

A selection of social science data is published on the DANS NESSTAR server[91]. For the documentation of these files, DDI codebook[92] is being used as metadata schema. The DANS NESSTAR server is harvested by the CESSDA catalogue[93]. These NESSTAR DDI files, and other DDI files submitted by data depositors, are uploaded as xml-files in EASY for preservation purposes. They are recognisable as 'additional-metadata.xml' in the file folder.

Like DDI files for the social sciences, the CLARIN community uses CMDI[94] as a metadata format. CMDI is a flexible metadata format, which allows data creators to define their own CMDI profile. The CMDI metadata files are produced by the depositors, and sustainably stored in the dataset file folder 'NIET-DC-metadata' within EASY. The files are harvested by CLARIN, and are made available at the Virtual Language Observatory[95], which provides services in the field of linguistics.

Technical metadata
Upon ingesting files in EASY, technical metadata is added to each file. It includes the size, the MIME-type, and the file-ID in the Fedora system. DANS does use an internal schema for technical metadata.

Structural metadata
EASY files are archived in a virtual file folder structure. Each file automatically gets a data identifier, a folder identifier, and a file identifier. Files that belong together are placed in the same virtual folder. This is done automatically upon ingesting. When checking the dataset, the data manager ensures that the representation of the files is clear to the user. Sometimes the data manager restructures the folder structure, or changes file names or folder names. The original file structure remains preserved in the folder 'original'.

Preservation metadata
Preservation metadata is closely linked with technical and administrative data metadata; it records information that maintains the longevity of a digital data

[90] http://easy.dans.knaw.nl/schemas/md/2012/11/ddm.xsd
[91] http://nesstar.dans.knaw.nl/webview/
[92] http://www.ddialliance.org/
[93] http://www.cessda.net/catalogue/
[94] http://clarin.eu/content/component-metadata
[95] http://clarin.eu/content/virtual-language-observatory

object for future use. Up to now DANS has not used a specific schema for capturing this kind of metadata. At the moment (Spring 2014) we are looking at PREMIS[96]. The PREMIS Data Dictionary for Preservation Metadata is the de facto standard for metadata to support the preservation of digital objects and ensure their long-term usability.

Administrative and provenance metadata
In EASY, administrative metadata is captured about the users of the system. The data manager can record provenance metadata about the dataset, using a checklist and a free-text field. We use an internal schema for administrative and provenance metadata, we do not make use of a specific standard. Information related to the processing of datasets is published in the provenance document 'Provenance Document: the Processing of Datasets by DANS' on the DANS website.[97]

Controlled vocabularies
A custom made controlled vocabulary is available for the field 'audience' defined by DC terms[98]. Other controlled vocabularies are only available for the metadata fields 'Subject' and 'Coverage' for deposits of archaeological datasets. These vocabularies[99] contain terms from a national dictionary of standardised archaeological codes (ABR, Archeologisch Basisregister[100]). These vocabularies are not mandatory. For other disciplines, controlled vocabularies are not being used so far.

Granularity of the metadata descriptions
Descriptive metadata in EASY is used and stored with datasets. A dataset contains one or more folders and files. Technical metadata is created and stored on a file level.

Interoperability
At the moment there are 4 Metadata schema(s))[101] available, which are exposed by DANS for OAI-PMH harvesting: OAI-DC, CARARE, CMDI and NL-DIDL. OAI-DC is the basic metadata schema required by OAI-PMH. CARARE is the metadata format for the CARARE portal, a service that brings together digital content for archaeological monuments and historic sites interoperable with Europeana[102]. CMDI is being used within the CLARIN network. NL-DIDL is a set of guidelines for the use of DIDL and MODS by institutional repositories in the Netherlands to allow NARCIS[103] to harvest rich bibliographical metadata.

[96] PREMIS (Preservation Metadata: Implementation Strategies) see: http://www.loc.gov/standards/premis/
[97] http://www.dans.knaw.nl/sites/default/files/file/Provenance%20document_120823_UK.pdf
[98] http://easy.dans.knaw.nl/schemas/vocab/2012/10/narcis-type.xsd
[99] http://easy.dans.knaw.nl/schemas/vocab/2012/10/abr-type.xsd.
[100] http://www.den.nl/standaard/166/Archeologisch-Basisregister
[101] http://easy.dans.knaw.nl/oai/?verb=ListMetadataFormats
[102] http://www.carare.eu/eng
[103] The portal provides access to (open access) publications from the repositories of all the Dutch universities, KNAW, NWO and a number of research institutes, datasets from Dutch data archives as well as descriptions of research projects, researchers and research institutes. http://www.narcis.nl/?Language=en

DANS also participates in the OpenAIRE+ project (Open Access Infrastructure for Research in Europe) [104]. To become harvestable by the OpenAIRE aggregator, DANS has to make the EASY metadata in compliance with the OpenAIRE guidelines for Data Archives[105]. This metadata format will be available by the end of 2014.

Metadata creation
As mentioned in the section on the concept op self-archiving, the creation of metadata is mainly the responsibility of the data creator and data depositor. Besides the checking of metadata provided by the data depositor and some editing, data managers at DANS do not create metadata themselves. However, for special projects DANS can be involved in metadata creation.

Instruction material and training
Every metadata field in EASY is accompanied by instructions and examples of how to use the field. Instruction material for the use of EASY is available on the DANS website[106]. This webpage contains PDF documents with detailed instructions per discipline. In addition to these instructions, information about preferred formats, licenses, data processing and data management plans is provided on the website.

Within the context of Research Data Netherlands[107] (RDNL), DANS is involved in the organisation of training courses for data librarians. These trainings are a combination of face-to-face meetings and a website[108] containing online information.

D.4 New developments

Front-Office Back-Office model
Within the context of RDNL, a Front-Office Back-Office model is being developed. RDNL is an alliance between 3TU.Datacentrum and DANS with a mission to promote long-term archiving and reuse of research data. The coalition was founded in 2013 and is also open to other Dutch parties. The partners in RDNL aim to fulfil a back-office function in the process of data curation - in other words to ensure that the research data delivered to them is permanently archived. The so-called front offices (university libraries, research institutes) are much closer to the actual research and are more involved in the actual data management and metadata creation of the research projects. This cooperation model will be rolled out in the coming years.

Surveydata.nl
For survey data, DANS collaborates with CentErdata, a research institute

[104] https://www.openaire.eu/openairefactsheet-40
[105] https://guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_Data_Archives These guidelines are based on the DataCite Metadata Schema v2.2 http://schema.datacite.org/meta/kernel-2.2/index.html, with some adjustments.
[106] http://www.dans.knaw.nl/en/content/data-archive/depositing-data
[107] http://www.researchdata.nl/en/
[108] http://www.researchdata.nl/en/activities/cursus/

specialised in online survey research. This cooperation will be realised in surveydata.nl, a service in which CentErdata is responsible for the documentation and dissemination of the survey data, DANS is responsible for the long-term preservation of these data. For this project we will make use of Questasy, a web application developed by CentErdata to manage the documentation and dissemination of data and metadata for (panel) surveys. The metadata within Questasy is DDI-lifecycle compliant. A SWORD-interface facilitates the exchange of data and metadata between the Questasy servers and EASY. The aim of survey.nl is to build a network of Dutch survey research projects. All these projects will be documented and disseminated by dedicated Questasy servers and aggregated by the surveydata.nl portal. The aim is that in the (near) future this aggregation by surveydata.nl will replace the current DANS NESSTAR server.

D.5 Plans to enhance the quality of metadata

DIN certification

In 2014 DANS has started a project to become certified according the DIN Standard 31644[109]. This standard is developed by the DIN working group on Trustworthy Digital Archives of Nestor, the leading competence network for digital preservation in Germany. The standard consist of 34 criteria, which can be used by archives both for self-evaluation and for certification. These criteria are related to a broad range of topics: organisational aspects, staffing, financial and legal aspects, archival processes, IT-infrastructure, risk management, etc. Five criteria cover aspects of metadata (C28 Descriptive metadata, C29 Structural metadata, C30 Technical metadata, C31 Logging the preservation measures, C32 Administrative metadata).

As a consequence of the DIN self-assessment, we thoroughly looked at the metadata and the documentation procedures within DANS and made recommendations how these could be improved. These recommendations are being converted into an overall project plan to upgrade the archive procedures. This project will start at the beginning of 2015.

D.6 Strengths and weaknesses

Because of the mission of DANS, to serve a broad range of disciplines, we have a slightly different approach concerning metadata compared with the other institutes whose case studies are described in this deliverable. DANS doesn't work for one designated community, therefore we cannot be compliant to a single metadata standard. The metadata within EASY needs to be a kind of common denominator for the metadata used by all the communities where DANS wants to be of service. That is the reason DANS has chosen for Dublin Core, because the Dublin Core elements fulfil the basic minimal requirements to describe a dataset properly.

[109] http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html

We are aware of the fact that within DARIAH, CESSDA as well as CLARIN, Dublin Core is not considered sufficient for the various communities. In the same way, the EASY metadata would be insufficient from the perspective of a user. For example, a social scientist looking for specific variables can't be served by a search in EASY. The same goes for a linguist who is looking for specific services in the field of linguistics, he will also not be fully satisfied using the EASY system.

However in the perspective of the mission of DANS, to serve a broad range of disciplines, we think choosing for Dublin Core as the overall metadata standard, complemented by specific metadata standards like DDI for special cases, is an optimal strategy. It will be never possible for a relatively small organisation like DANS to be fully compliant to all different discipline specific metadata standards. We therefore form alliances with organisations that are specialised in a specific field, for example the collaboration in surveydata.nl with CentERdata. In these collaborations DANS will focus more and more on its role as a sustainable archive. For this role we mainly have to focus on descriptive metadata (for resource discovery), technical, structural, preservation and administrative metadata.

Appendix E: Case Study Austrian Academy of Sciences, Institute for Corpus Linguistic

E.1 Background

History of your institute (when is it established)

The Institute of Corpus Linguistics and Text Technology of the Austrian Academy of Sciences was founded in 2010 and was granted permanent status by the Academy in 2012. The department had two institutional predecessors: the *Commission for Functional Literary Text Types* (90ies of the past century) and the AAC – Austrian Academy Corpus. Most of the involved researchers worked at the AAC which was an early innovative digital Humanities initiative funded by the Austrian Federal Ministry for Science and Research. The AAC created a sizeable corpus of digital full texts representing language and literature of the 19th and 20th centuries. In 2013, a number of researchers of the former Institute for Lexicography of Austrian Dialects and Names (DINAMLEX) joined the ICLTT and have pursued since then their linguistic and lexicographic research and investigations on linguistic variation in the framework of the ICLTT's research groups.

Organizational Context

The ICLTT is part of the Austrian Academy of Sciences, Austria's largest non-university research facility. It is based in Vienna.

Part of which infrastructure(s)

The ICLTT has been participating in a number of Humanities infrastructures. It has coordinated the Austrian CLARIN and DARIAH activities together with the Centre for Translation Studies (University of Vienna) for three years. As of January 2014, these activities were merged in a new common initiative *Digital Humanities Austria*, for which the ICLTT has taken over as the sole coordinating institution.

Mission

The Institute for Corpus Linguistics and Text Technology pursues a wide range of interests all of which belong into the realm of eHumanities. Proceeding from a long-standing tradition of corpus-based linguistic and literary research, most research projects deal with digital language resources, the creation and adaptation of corpora and dictionaries as well as technologies for building, accessing and exploiting such data for a wide range of SSH disciplines. The ICLTT's scholars investigate lexical semantics and standards for eLexicography, they work on language documentation and do text technological research and development.

Main activities

In the past three years research activities were organised in 5 major research areas which as of November 2013 were reduced to three research priorities

that are reflected in corresponding working groups:

1. Literature in Transition & AAC – Austrian Academy Corpus
2. Text Technologies and Research Infrastructures
3. Corpus-based Linguistic Research

Working group 1 focuses on DH research. A major point of interest are scholarly digital editions.

In **WG 2** research questions circle around infrastructure components needed in digital humanities in general. This covers a wide range of interests: the researchers deal with issues of indexing, searching, metadata, encoding, but also with more applied issues such as digital repositories. Standards have played an important role, a particular focus are standards for digital lexicography. A very important aspect of more recent research activities are semantic technologies, knowledge representation in RDF and SKOS, and controlled vocabularies.

The **WG3** focuses on corpus-based linguistic research, and the exploitation of corpus data for lexicographic purposes. While earlier corpus activities were restricted to corpora of written language, a recent new field of activities is the build-up of speech corpora. In the past years, WG3 could acquire public funds to conductresearch into non-standard linguistic varieties (ABaC:us, TUNICO).

The developments of the past years have shown an expansion of the research scope from written Standard German towards research into other languages as well.

The ICLTT holds a large number of digital resources, at the core of which is the *Austrian Academy Corpus*. The vast bulk of this corpus dates from the first half of the 20th century. Currently, the corpus consists of about 500 mil. tokens and most of the texts contained in the collection do not strictly belong to the sphere of what traditionally would be described as *belles lettres*. As the texts were collected with a socio-historic, literary and lexicographic perspective, the corpus also contains a considerable amount of functional and informational texts. Roughly half of the data is made up of periodicals, not large size daily newspapers, but rather medium and small size weekly and monthly publications. There are many collective publications such as yearbooks, readers, commemorative publications, almanacs and anthologies which, in themselves, cover a wide range of writers, topics, types of texts and genres.

Since the foundation of the ICLTT, its project planning has been characterised by a strong commitment to the ESFRI projects CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities). The participation in initiatives directed towards common infrastructures for all kinds of language resources was a natural choice for the ICLTT with its traditionally wide range of research

areas. Research based on digital corpora and other digital language resources has been adopted as a set of methodologies in many disciplines within the humanities and social sciences. The ICLTT's involvement in DARIAH and DASISH (Data Service Infrastructure for the Social Sciences and Humanities) stems from the department's manifold projects which have been characterised throughout by considerable efforts in bringing together up-to-date ICT and cross-disciplinary approaches.

The ICLTT's staff has a long-standing tradition in computational lexicography both creating content and creating up-to-date NLP technology. In recent years, the ICLTT's activities in this field have been extended from text lexicography with its focus on particular literary texts to more general linguistically oriented research objectives. Research is conducted on both monolingual and bilingual resources. Lexicographic research is complemented by terminographic and terminological studies. Digital lexicography research has been focused on tools for lexicographic and terminological research and the development of respective standard procedures.

Information related to the institutional identity of the organisation
All published information is found at the website of the institute http://www.oeaw.ac.at/icltt.

E.2 Metadata production

In general, there has been a strong awareness of the importance of metadata at the institute from its early days. The researchers of the department have tried to capture relevant metadata very early on in all their projects.

Types of metadata that play a role within the work processes
The types of metadata that have been created and maintained over the past two decades comprise descriptive, administrative and structural metadata.

Metadata schema(s) are being used?
Traditionally, the ICLTT has been making use of TEI headers. More recently, the policy has been changed towards CMD (Component Metadata) in combination with METS (Metadata Encoding and Transmission Standard).

Context and purpose of the schema
From the very beginning of their digital activities, the metadata policy of the group of involved researchers has been based on **TEI headers** for descriptive purposes. Each digital object, representing a physical item such as a book / a bound volume, was provided with a TEI header. In terms of number of metadata records TEI headers still constitute the main body of metadata information.

All recently produced data (i.e. as of 2013, running projects, data in the *Language Resources Repository*), is now being described making use of CMD (Component Metadata).CMD allows for flexible descriptive metadata, and also

features provisions for structural metadata, allowing to capture collection hierarchies, typed links to actual resources and even arbitrary relations between resources.

The ICLTT was actively involved in the development of the CMDI, both the infrastructural components as well as on the level of modelling the so-called CMD profiles that translate into custom XML-Schemas. Given the institute's stock of teiHeader records, the team joined the international efforts in converting teiHeaders into CMD records, cooperating with research groups of the University of Copenhagen (CLARIN-DK), Oxford Text Archive and Berlin-BrandenburgerischeAkademie der Wissenschaften(BBAW), Berlin.
It turned out that the vast variability of the TEI schema cannot be easily captured in one CMD profile, however at least the teiHeader records of CLARIN-DK (University of Copenhagen) and CLARIN-AT (ICLTT) could be unified into one profile/schema (teiHeader (CLARIN-DK)). The BBAW uses its own custom teiHeader CMD profile (teiHeader DTA).

A summary (made by ICLTT colleague) of the teiHeader related CMD profiles is available online:
http://clarin.oeaw.ac.at/smc-browser/docs/smc-report_teiHeader.html

Aside from these teiHeader emulations, some dedicated CMD profiles turned out to be better suited for describing some types of resources (like lexical resources or collections/corpora).Thus, next to the *teiHeader* profile, ICLTT adopted the *LexicalResourceProfile, collection profile* and*TextcorpusProfile,* all recommended profiles developed by other groups and used for description of resources at other centres as well. The LexicalResourceProfile and TextCorpusProfile were created in the context of the initiative NaLiDa by the University of Tübingen based on interaction with actual users. Accordingly they are well designed and very comprehensive, allowing to describe all relevant aspects of the resources like contact information, access modalities, project context, resource size, technical aspects of the resource (encoding, annotation levels) etc.

Administrative data have been encoded in a XML vocabulary that was designed at the institute for the particular purpose many years ago. It has long since been planned to replace this format by something more conformant to standards. However, the fact that the workflow management is accomplished by software which could not be substituted over the past few years due to financial and logistical circumstances, has prevented any move in this direction.

Concurrently with the move towards CMD for descriptive metadata, the ICLTT's technical crew havestarted to employ the widely used METS Schema (Metadata Encoding and Transmission Standard) for structural metadata. This is a XML vocabulary that allows the comprehensive description of data, metadata and its underlying file structure in a standardized, implementation independent manner. This step was an essential step forward as we were aiming at a tight

integration of *cr-xq*(the content repository component of the *corpus_shell*[110]) with the fedora commons repository which constitutes the core of the ICLTT's mid and long term preservation infrastructure.

URLs

Detailed information on teiHeaders can be obtained from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html the teiHeader section of the P5 version of TEI.

For the CMD profiles, the descriptions are maintained by the Component Registry:

- - TextCorpusProfile
- - LexicalResourceProfile
- - collection
- - teiHeader (CLARIN-DK)

The available Metadata formats can also be consulted at the OAI-PMH endpoint of the institute's repository:

http://clarin.oeaw.ac.at/oaiprovider/?verb=ListMetadataFormats

Type of material to which the metadata relate

The material the metadata relates to is quite varied. There are several digital text collections involved. Most of the older materials are digitised printed publications. In addition to literary works and different types of monographs, there are large amounts of periodicals, journals and magazines. The historical data were collected as part of a program investigating the German language between 1848 and the late 20th century. Notwithstanding this scope, the large bulk of data stems from the first half of the 20th century.

The largest collection is the Austrian Media Corpus (AMC),which consists in roughly 8 billion tokens. This corpus covers the entire Austrian media landscape of the last two decades. All of this material is digitally born.

In recent year, the agenda of the institute has been opened to work on other languages than German. The collections available at the institute now also contain digital dictionaries. While we have been working on digitising print dictionaries on a small scale, most of the lexicographical resources are digitally born and follow a strictly defined TEI schema which is meant to ensure cross-dictionary searches and the application of tools to a number of lexicographic resources.

Granularity

110 corpus_shell is a service-oriented architecture for distributed and heterogeneous language resources. Its core functionality is encapsulated in self-contained components exposing well-defined interfaces based on acknowledged standards. The principle idea behind the architecture is to decouple the modules serving data from the user-interface components. To achieve this end, a number of basic requirements are imposed on the system: dynamic configuration of data sources, dynamic configuration of front-end layout, support for different protocols and support for different data formats. It is built on and around standards and protocols endorsed by and used in the CLARIN infrastructure.

Metadata describe digital objects on three layers: collections, resources and resource fragments. Collections are understood as resources in their own right, both collections and individual resources published via the institutional repository are furnished with PIDs.

There is metadata records for collections and resources. Resource fragments are only identified by a fragment identifier, but are not accompanied by a separate metadata record.

The issue of granularity has been dealt with extensively in the context of the creation of the AAC. Given the fact that more than half of this collection is made up of periodicals and other collective publications which only provides metadata at the level of the physical item, the percentage of identifiable texts is very low. Our knowledge of writers, genres, topics, etc. is restricted to those parts of the corpus where metadata could be attributed to the highest-level digital object (the above-mentioned book and volume levels). To put it another way: the corpus contains a comparatively large amount of textual data and a comparatively scarce amount of metadata. Even though we are in possession of a unique and very large collection of texts, with regard to the journalistic texts even our specialists in literary and history studies have only a fuzzy picture of what this treasure trove really contains. Thus, at present, we are unfortunately unable to provide our users, e.g. with a comprehensive list of authors whose works appear in the journals.

Who are involved in metadata creation, within and outside your institute?
The creation of descriptive metadata has been performed by scholars of the department working on the various research projects. The creation of administrative metadata was performed in a semiautomatic way that was integrated in the corpus creation workflow.

Usually, the corpus/collection records are authored completely by hand, capturing the project context, access modalities and such. The records for individual text have been captured in a separate database at the beginning the digitization process.

Number of employees working at metadata
Over the past ten years, two colleagues were involved with the creation of metadata records. Descriptive records for the AAC were edited manually. The creation of administrative metadata has been supervised by the same two colleagues who have accomplished the task alongside other responsibilities.

More recently, another 2 persons are responsible for further developing and maintaining the metadata information, who interact with members of individual research project to agree on the best suited schema for the metadata and deliver the metadata records for the resources being created in those projects.

Which tools and techniques are being used?
In general, the primary format for all data at the institute is XML, implying the

use of cognate technologies (XSLT, XQuery, XML Schemas). Various tools developed at the department have been used to create and edit the data.

Editing of metadata has been performed making use of a range of different technologies. Originally, the metadata records were stored in a dedicated database with means to be exported into teiHeader records.

Furthermore XML database eXist is used for storing and editing data and metadata. More recently, stable data + metadata is deposited the fedora-commons based digital object repository.

What are the procedures?
Administrative metadata were created by means of the *TaskEditor*, a tool that was created in the early phases of the project. *TaskEditor* assigns particular workflow steps to editors. The tool manages the assignment of the data, retrieves them, takes care of archiving and stores data describing the involved workflow steps (who, when, what).

Descriptive metadata were edited for a long time in a simple database frontend, the data was stored in a relational database (MSSQL). The *corpusEditor*, a corpus editing and management tool also developed at the department, transforms the data in the database and adds them on a regular basis as well-formed and valid TEI headers to the digital objects.

The CMD metadata of recently produced resources have been produced making use of standard XML editors such as oXygen, with the primary persistence layer being XML-DB eXist.

Are controlled vocabularies being used?
In those parts of the corpora where metadata were assigned on the level of resource fragments (AAC), *Dewey Decimal Classification* (DDC, version 22, German) has been utilised. There were a number of arguments in favour of this system: it covers much–though not all–of what was needed to classify the texts and contents at hand, it has been translated into over 30 languages and can thus be easily mapped from one language onto another, it has an ever-growing international community and there are a number of projects working on DDC interfaces with other systems.

Another type of controlled vocabularies are the ISO language-codes, both 639-2 and 639-3 have been used as defined in BCP 47. The system also makes used of private use subtags (BCP 47, 2.2.7) as some linguistic varieties in the collections did not fit into the 639-2/3 scheme.

Is there instruction material? Is there training? Who is the intended audience for this material and training?
There are some internal material and samples. There have been plans to create publicly available documentation of the metadata creation procedures. However, tight project schedules and lack of resources have not allowed us to go about this task. So far, relevant know-how has been passed on in an

informal manner within the team.

With the start of *Clarin Centre Vienna* (early 2014) staff has begun to compile instruction material on how to create metadata which is intended for potential users of the *Language Resources Portal*, the first Austrian repository for language resources.

Which procedures are taken to enhance the quality?
Technically, the quality of metadata is ensured by means of validation through XML Schema and manual inspection.

On the organisational level, there is ongoing effort to better document and formalize the process of metadata creation.

Do you have plans to change workflow and procedures?
There is an ongoing effort to establish integrated procedures for the whole life-cycle of authoring/generating and storing/publishing resources and their metadata. This involves the development stage, where the resources and metadata can be edited by the project team and the production stage, where the resources are assigned a PID, published and stored in a stable content repository.

One goal is also to harmonize the metadata landscape and have CMD and METS records for all available resources.

E.3 Additional notes on metadata production in DARIAH

There is so far no unified strategy on metadata within the whole of DARIAH, however individual sub-communities obviously use various metadata formats to describe research data and there are efforts to come to a common ground within the DARIAH community as well. One proposed solution is the collection registry by DARIAH-DE, offering a simple metadata search over resources on collection level (format Dublin Core application profiles). The French *HumaNum*group running the DH knowledge portal rechercheisidore.fr(RDF-triple-store based) proposes a light-weight approach collecting information about DARIAH in-kind contributions: providers would enrich the web-pages describing their resources with RDFa, these pages get harvested/crawled and the conveyed information can be ingested into a common triple-store. However, all of this is a still very experimental setting.

Although there are many institutions with repositories, most of them are rather meant as publication archives/repositories, offering research results (publications, papers) rather than research data. Some, however, are designed as general purpose archives/repositories, allowing their users to store both publications and data. This brings up the question of how to deal with such mixed repositories, as it is usually not possible to distinguish between publications and research data at the OAI-PMH endpoint. Thus, the harvesting

would have to be either all (filtering only after harvesting) or nothing. Another consideration is, that publications could be seen as research data themselves.

One important aspect within the large DARIAH community is the relationship to the so-called "memory" institutions (libraries, museums, archives) that are seen as potential content providers for DH research. These institutions already mandate aggregated information about their collection by way of various partly long lasting initiatives, such as *WorldCat*,*DBIS*, *OBVSG (Federation of Austrian libraries)*, *The European Library* and *Europeana* (all of them memory institutions). Thus it seems worthwhile for infrastructure projects to consider direct cooperation with these aggregators instead of duplicating the tasks of searching and collecting individual repositories. It is interesting that some institutions already providing metadata for *Europeana* and being ostensibly keen to support DH research, appear to be reluctant to make public their OAI-PMH endpoints (even though they don't see any access/licensing restrictions on the metadata, it is rather to limit the administrative effort).

All in all, the situation in DARIAH indicates that there is a growing awareness of the importance of Semantic Web technologies (and consequently RDF-triple stores as a repository solution), which also seems the appropriate response to the great heterogeneity of resources and their descriptions encountered in the context of the arts and humanities.

E.4 SWOT analysis

Strengths: characteristics that give our approach an advantage over others.
The hybrid approach to metadata production allows for high flexibility in resource descriptions. The granularity and specificity of the (descriptive) metadata can be tailored to the specifics of the resource described and the needs of the project / research task for which it is used.

The growing community making use of CMD is a strong argument in favour of CMD. While sustainability of efforts with respect to metadata creation is difficult to assess, one may expect that – given the large communities involved – the dual approach of making use of TEI headers and CMD should be seen as a viable solution.

Weaknesses: characteristics that place our approaches at a disadvantage relative to others
The structure of the metadata landscape at the department is still quite heterogeneous which is due to the great number of projects, the divergent objectives and different stages of development.

Metadata creation and management is not handled coherently enough, partly due to legacy material, partly due to too diverging needs and situations / organisational setups in individual research projects. In this respect, a more

integrated approach needs to be sought.

Opportunities: elements that our project could exploit to its advantage
The landscape of descriptions for language resources has evolved dramatically over the decade with strong standardization efforts on a broad international consensual basis. Our institute is actively involved in these processes, which allows us to feed our experience into the process and evaluate in practical use cases format proposals early in the development process.

Threats: elements in the environment that could cause trouble for the project
Partly multiple parallel standardization activities are ongoing pursuit by different groups. Although there are efforts to talk to each other over the boundaries and "build bridges", this can potentially lead to new incoherence in the metadata landscape. Also one has to be aware of the (fundamental) gap between the XML-based and RDF-based approaches and it is not clear how they will evolve and co-exist.

While CMD has been submitted officially to ISO to become a standard, it is still under development. The number of competing profiles makes it difficult to anticipate which CMD profiles will be the ones to survive.

Appendix F: Case Study of the CLARIN-DK Repository at University of Copenhagen

F.1 Introduction

This case study describes the metadata workflows within the CLARIN-DK repository at University of Copenhagen. The CLARIN-DK-UCPH repository is a CLARIN B centre.

CLARIN (http://clarin.eu) [111] is one of the Research Infrastructures that were selected for the European Research Infrastructures Roadmap by ESFRI, the European Strategy Forum on Research Infrastructures. It is a distributed data infrastructure, with sites all over Europe. Typical sites are universities, research institutions, libraries and public archives. They all have in common that they provide access to digital language data collections, to digital tools to work with them, and to expertise for researchers to work with them. The CLARIN Governance and Coordination body at the European level is CLARIN ERIC[112], and its members are governments or intergovernmental organisations.

The Danish government are supporting the CLARIN-DK-UCPH repository, which is hosted by the Faculty of Humanities. The development of the repository is carried out by a team at the department Centre for Language Technology, CST, at the faculty of Humanities.

The case study has the focus to describe the types of metadata in use, the workflows and procedures in which metadata plays a role, the different roles and responsibilities of the involved people. Besides the description, the aim is also to sum up weaknesses and threats in the procedures to ensure metadata quality.

F.2 Background

History and Organisational Context

Centre for Language Technology, CST, is a department of the University of Copenhagen (UCPH), Faculty of Humanities since a formal merge as of January, 2004.

CST was established in the early eighties as a temporary centre at the University of Copenhagen with the purpose of fulfilling the Danish obligations as a partner in the large European Machine Translation Project, EUROTRA. In 1991 the centre was granted status as a research centre at the University of

[111] Paragrph cited from: http://clarin.eu/content/general-information

[112] An ERIC is a new type of international legal entity, established by the European Commission in 2009.

Copenhagen, and in 1996 it was formally granted status as an independent government research institution and as the Danish national centre for language technology. In the statutes established at the merge with UCPH in 2004, CST maintains its status as a national centre for language technology.

CST has a strong engagement in the development of research infrastructures for the Humanities, both on national and European level.

The CLARIN-DK-UCPH repository was created in the national project DK-CLARIN supported by a grant of approx. 2 million € from the research infrastructure programme of the Danish Agency for Science, Technology and Innovation. The grant was for construction of a Danish research infrastructure for the humanities integrating written, spoken, and visual records into a coherent and systematic digital repository during the period 2008-2011. CST was the leading partner of DK-CLARIN, and therefore CST took over responsibility for the repository when the project funding ended.

Vision and Mission of CST
CST aims to be an important player in providing good language technology for Danish users and other users of the Danish language, and to bring new knowledge to Denmark through international collaboration, as well as contributing to the international scientific development in the field.

Main activities
CST employs a staff of around 20 scientists (linguists, computational linguists, engineers, computer scientists) working in many areas of language technology. In addition to basic research in lexicography, formal grammar, deep semantic analysis, machine translation (MT), machine learning, and multi-modality, UCPH has considerable experience in the development and evaluation/validation of a variety of Human Language Technologies (HLT) applications such as MT, and in collecting data resources such as large corpora, both in EU projects, in national research projects, and for commercial customers.

Infrastructure participation
CST's engagement in infrastructures for the Humanities includes participation in a number of infrastructures. As already mentioned CST – as part of UCPH - is the Danish institution delivering the Danish CLARIN-ERIC membership contributions, and before that CST was involved in the European preparatory CLARIN project.

Besides the DASISH project to which this report is a contribution, CST has also been involved in META-SHARE[113] through the META-NORD[114] project, and is also providing a META-SHARE repository for corpora and tools for language

[113] A network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources, see more at http://www.meta-share.eu

[114] META-NORD: Baltic and Nordic Parts of the European Open Linguistic Infrastructure, see more at: http://www.meta-nord.eu/

technology developers.

Mission of CLARIN-DK-UCPH
The mission of the repository is to be the Danish node in the European CLARIN-ERIC, and thus provide easy and sustainable access for scholars in the humanities and social sciences to digital language data (in written, spoken, video or multimodal form) and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of where they are located. Digital archiving and long-term preservation and easy and sustainable digital access to data resources and tools will offer new possibilities for the scholars to develop new research methods and ask new types of research questions, and it will support and enhance their participation in collaborative international research.

The CLARIN Centre at the University of Copenhagen promulgates all aspects of this mission through publications, conference attendance, organization of PhD courses and other courses and workshops, e.g. in collaboration with other Danish Universities through the national digital humanities collaboration, DIGHUMLAB115. Employees at the CLARIN Centre at the University of Copenhagen are active participants in both national and international fora that aim to establish standards for best practices and infrastructures for digital archiving. Mission statement can be found here: http://info.clarin.dk/en/overblik/datamanagement/

Main activities of CLARIN-DK-UCPH
The CLARIN Centre at the University of Copenhagen focuses on serving the Faculties of Humanities in Denmark for spoken and written languages. It is therefore promoted as a standalone initiative and not as a part of CST.

The team working on tasks connected to the CLARIN-DK-UCPH repository counts one fulltime and six part time persons. The development work on the repository follows two tracks: further development of the CLARIN centre, and assisting and facilitating use of the repository through dialogue with users. The CLARIN centre was granted the Data Seal of Approval in January 2014, and is now a CLARIN-B centre.

Users can make a metadata search for data, and if they have the right permissions, they can download data. Academic users can deposit new data. A number of tools can also be accessed through the repository.

A number of help desks are set up to help researchers and scholars in use of the repository and tools, and to promulgate the use of the repository among Danish users.

Materials
The CLARIN-DK-UCPH encourages data owners and producers to deposit data

[115] *The CLARIN Centre at the University of Copenhagen is financed with public national funding through the national infrastructure collaboration DIGHUMLAB (http://dighumlab.dk/)*

and their corresponding research material (documents and annotations) in the repository and provides data management consultation and support in connection with the deposit. The focus is on written and spoken resources, but all the following types of resources can be deposited:

- Texts, with or without scanned images of original book pages
- Text annotations of the text
- Video
- Audio
- Video and audio annotations
- Lexica
- Tools
- Data resources (resources that are not one of the other types above, e.g. wordlists, compiled data needed for tools)

We are currently preparing the administration of the resource type collections, which will be a resource type that can contain a collection of other resource types. For the text, video and audio resources and their annotations, the main focus is written and spoken Danish language, but material in other languages can also be found in the repository.

F.3 Metadata Production Overview

Different metadata formats are used for different resource types. For text resources a subset of the TEI metadata standard is used. For Lexica a subset of TEI is also used, with a large overlap with the TEI standard subset for text. For Video, audio and annotations of these data types the IMDI standard it used. For tools and data two CMDI[116] profiles are defined, which reuses the OLAC standard but also extents it with special information relevant for the two resource types. This design principle has been decided together with the user group established in the design phase of the CLARIN centre, and as the users come from different communities, they prefer different standards for documenting their metadata, as no standard covers all needs for all the user groups.

To make it easier to search among a number of resource types in the same metadata search, a number of obligatory metadata elements are used by all resource types.

Currently, metadata production is done by the data owner, often in close collaboration with the CLARIN centre staff. It is our experience that metadata is an abstract matter for many users, and structuring information in metadata is for many researchers a new kind of work.

We believe that the data owner is the person best suited to fill in the metadata, as s/he has the detailed information about the resources, but we

[116] The profiles were defined in an early version of CMDI, they will be upgraded to CMDI version 1.1 during 2014.

fully recognise the need for assistance and that we as a data centre will benefit from these dialogues that also can lead to extensions of metadata or changes in interface or data visualisation and retrieval.

In the deposit workflow, the metadata has to validate with the schemas relevant for the resource type. The users need to choose which resource type the data they are going to deposit conforms to, and to follow the deposit workflow for this resource type.

More information about deposit and validation can be found here, together with links to currently used validation schemas:

- http://info.clarin.dk/en/deponer-resurser/
- http://info.clarin.dk/en/deponer-resurser/validationrequirements

Types of Metadata in Play

In the following, the metadata categories used for the resources in the data centre are described. In this description some metadata may be mentioned more than once, but in the repository each metadata is of course only found once.

Only a minor part of the metadata is obligatory, and the obligatory amount of metadata depends on the resource type. For all resources the following metadata elements are required:

- Title
- Type of Resource
- Creator
- Creation date
- Description
- Format
- Publisher/Depositor
- Publication date
- Language
- Conforms To Standard

For all resource types and all used metadata standards, a joint list of metadata elements counts 82 different elements, but we will not go into that detail in the following, but only list the most common metadata elements.

Descriptive metadata describes the characteristics of data that will help with resource discovery: Main metadata in this group include *title, creator, creation date, publication date, depositor, subject*, and *keyword***s.** But in the data centre you can search on almost[117] all metadata for discovery, and in the resulting display of resources all metadata can be seen as descriptive.

[117] Some metadata *notes* fields about the production process is not searchable

Administrative metadata: provides information that helps with the management of a resource, such as when and how it was created (*creator, creation date*), when and by whom it was deposited (*publication date, depositor),* who can access the resource (*availability*).

Technical Metadata: refers to the technical processes used to produce, or required to use a resource, and relationship to other resources. It includes file format, used standard for content and language. Relationships to other resources are stored in a separate triple store, and not explicitly as metadata.

Structural metadata: describes the structure of data so that it can be interpreted correctly and viewed in the intended order. It includes information about *file format, used standard for content* and *used languages.* These are also mentioned as technical metadata. Also the *description* can contain information about the structure of the data.

Preservation Metadata: No metadata can be said to belong to this category alone, but part of the administrative and technical metadata can be seen as preservation metadata.

Metadata Standards and Context
Information about resource types and validation schemas can be found here: http://info.clarin.dk/en/deponer-resurser/ and http://info.clarin.dk/en/deponer-resurser/validationrequirements. For each resource type a specific metadata schema based on a standard has to be used. These standards are used (see the next section for more details):

OLAC, Open Language Archives Community[118] and DC.

When metadata is deposited, a part of them is transformed to different formats. As a part of this process, all objects are given an OLAC record which includes DC information that can be harvested by the OAI-PMH protocol.

TEI P5, Text Encoding Initiative[119]
Metadata for text resources, text annotations, and lexicons are deposited in TEI P5.

IMDI, ISLE Meta Data Initiative[120]
The IMDI format is used for audio, video and annotations of these resources.

Currently we are creating a CMDI version 1.1 (Component MetaData Infrastructure [121]) metadata record for all resources too. These records are viewable in the CLARIN ERIC Virtual Language Observatory at http://catalog.clarin.eu/vlo.

[118] (www.language-archives.org
[119] (http://www.tei-c.org/Guidelines/P5/)
[120] (http://en.wikipedia.org/wiki/IMDI)
[121] http://www.clarin.eu/content/component-metadata

The resources from CLARIN-DK-UCPH can be found at: http://catalog.clarin.eu/vlo/;jsessionid=11D7BC3362029E0E7D5CF2FB3043B2BF?fq=nationalProject:CLARIN-DK-UCPH.

Metadata Granularity
Metadata is connected to a resource. Depending on the resource type, the resource may contain more files. Below is a short overview of the possible contents of the resource types.

Texts, with or without scanned images of original book pages - Texts to be deposited have to be formatted in TEI P5 with both a teiHeader part for metadata and a body part with the text in TEIP5 format. In the body part links to scanned images of original book pages can be added and the scanned images then have to be included in the zip file together with the texts. Metadata is created for each text.

Text annotations of the text -Each annotation of a text has its owns metadata attached.

Video and audio - Metadata is specified in IMDI metadata.

Video and audio annotations- Metadata is specified in IMDI metadata.

Lexica -Lexica have a TEI P5 metadata file attached. But a Readme file can also be attached.

Tools - Tools deposited for download are deposited with a metadata file in CMDI format. The metadata is an extended version of OLAC. A Readme file can be attached.

Data - Other resource types are deposited with a metadata file in CMDI format. The metadata is an extended version of OLAC. A Readme file can be attached.

Metadata Methods and Process
Metadata are created by the data provider. The data centre team assists when necessary in the creation process.

Normally an xml-editor like Oxygen, that can validate with a schema as you type, is used. For IMDI files the Arbil (http://tla.mpi.nl/tools/tla-tools/arbil/) editor can also be used.

A web-based metadata editor is planned to be implemented that gives the user guidance via help texts and pick lists.

Instruction material and training
The normal procedure is that a new data provider asks for help for understanding what is needed to deposit data. Then meetings and workshops can be arranged – depending on the need and type of data.

Links to a number of documents that advices on how to fill in the metadata files are also provided, see https://www.clarin.dk/documentation/ and http://info.clarin.dk/.

Controlled vocabularies
The schemas have pick lists for a number of metadata information types. The pick lists are defined in the schemas that are available at: www.clarin.dk/schemas/ and more info can be found at http://info.clarin.dk/kom-godt-i-gang/valideringskrav/. A separate set of controlled vocabularies is not used.

Procedures to enhance the quality
In 2014, procedures to enhance the quality will be implemented. We will define a new procedure for deposits. New metadata will be inspected and reviewed by automatic scripts and the extracted information will go through a quick manual inspection to make sure consistent naming is used.

Plans to change workflow and procedures
The plans about changing the workflow currently include:

- A metadata editor
- Easier user guidelines
-

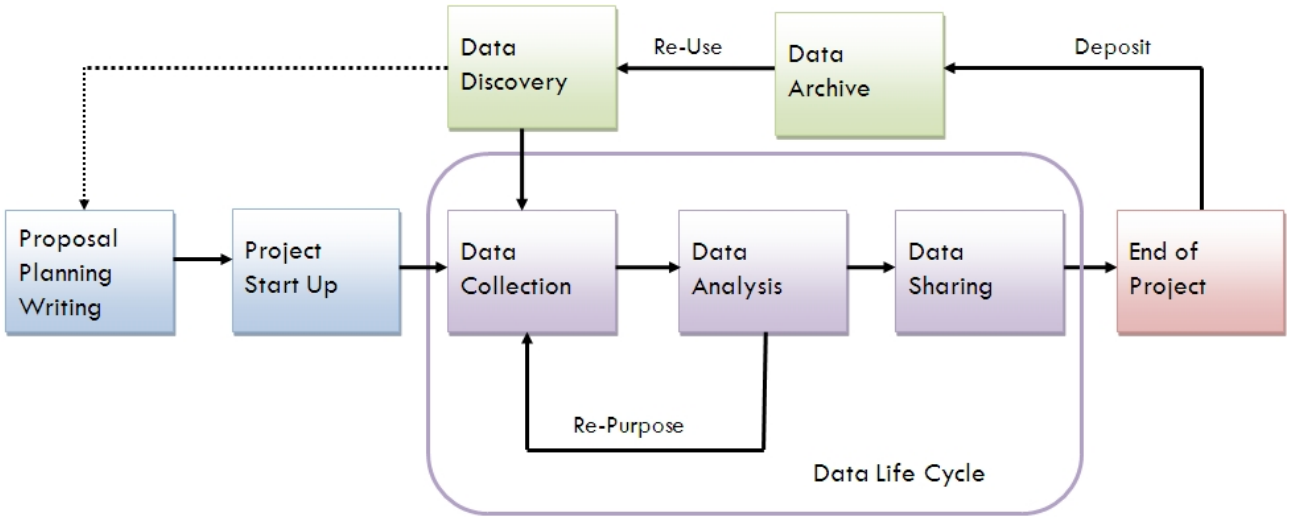
The metadata values are currently a mix of Danish and English. Descriptions have a language code attached to specify the language of the content. All pick lists are in English except the subject area list that refers to a Danish standard for subject domains.

The focus is that the metadata to the greatest possible extend should be machine actionable, so closed pick lists are preferred to open fields.

F.4 Mapped to Data Lifecycle

Here we will first give a description using the data lifecycle[122] used at University of Virginia.

[122] http://dmconsult.library.virginia.edu/lifecycle/



The main part of the resources stored in the CLARIN-DK-UCPH data centre have been created with the clear focus that the resources should be deposited in the data archive. Therefore this data lifecycle does not fit well with the main part of the resources in the data centre, as they were deposited before the end of the project.

After depositing the resources with their metadata it became obvious - as part of data sharing - that a number of inconsistences occurred in the metadata. To some extend it was possible to correct those errors as a part of the project. But other issues had to be handled after the end of the project.

This data lifecycle model seems to lack a data curation/validation phase.

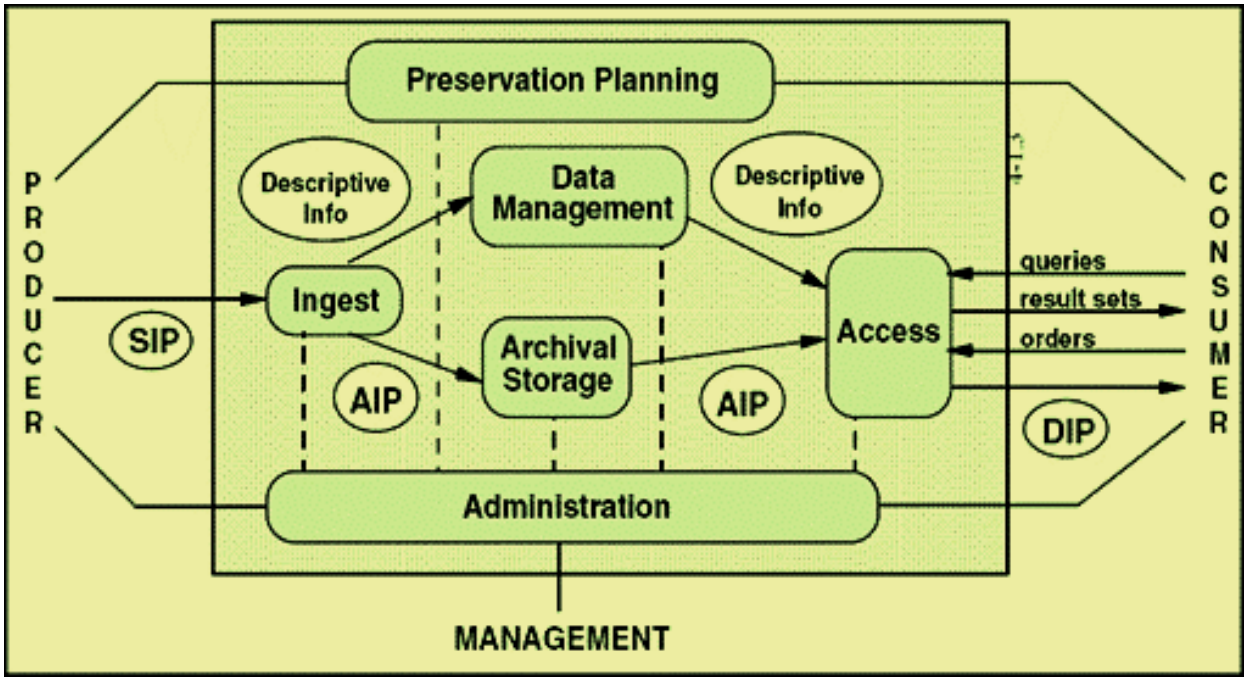
Currently when integrating more data, metadata creation starts just before depositing, and metadata is deposited to facilitate data sharing and search. This is not modelled in this lifecycle model either.

Another issue is that if the data provider has no funding for the metadata creation – if it has to be made after the project ends - it might not be done on a very detailed level, and the metadata might be partly unusable.

The CLARIN-DK-UCPH repository aims at being as conformance to the OAIS reference model's[123] tasks and function as possible.

[123] http://public.ccsds.org/publications/archive/650x0m2.pdf

However, due to the complexity of the OAIS reference model, the repository only implements some of the work flow.



Ingest: The repository uses the national identity federation WAYF.dk to support single identity and single sign-on operation based on SAML2.0 and trust declarations. Those users that are defined as researchers by their home institutions can ingest a *Submission Information Package* (SIP) to the repository. To submit a SIP the user selects and accepts the licence under which the data will be deposited. The SIP has to fulfil a number of requirements to be accepted. The metadata format and content included in the SIP has to comply with a defined list of standards for which there are defined xml schemas that will be used to evaluate the metadata contained in the SIP. After validation of the SIP, the deposit service handles the transformation of the SIP to the *Archival Information Package* (AIP). The procedures for checking the SIP before creating the AIP will be extended in the future. Scripts will be applied to extract, and structure the metadata to reveal if some of the extracted information needs to be extended or changed, making sure consistent naming is used in the metadata.

Archival Storage: The repository is using the Fedora Commons Repository Software with the eSciDoc (The Open Source e-Research Environment Processes) extension. Backup of the repository is carried out on a daily basis, and backup storage is done on an external location.

Data management: Both eSciDoc tools and the standard Fedora Commons tools, in combination with a specific administration application are used for data management. Metadata is distributed via the OAI-PMH protocol, supporting selective harvesting as well. The OAI-PMH supplied metadata, the Fedora Commons tools, and the administration tool are used to report on the status of the data.

Preservation Planning: The metadata contained in the SIP is preserved unchanged. It is an important issue that the data should be preserved, but the procedures are not yet defined. This work is in progress.

Administration: Contract agreements with the Data Producer are created when the SIP's are ingested. Administration staff includes a content manager who is dedicated to issues about the content administration and validation.

Access: The Dissemination Information Packages (DIP) and query responses are delivered to users, who have the rights to access the data. Metadata are publicly available, content data can require public, academic or restricted access permissions. A user interface available on clarin.dk allows the user to search metadata. The user can also inspect some of the content types online and download the content if the access requirements of the content have been met by the rights of the user. We do not handle micro data that contains sensitive information. The digital objects are in the process of being available for reading access via their Persistent Identifiers (PID) for authorized users, based on the national AAI infrastructure. The PIDs will be available in the metadata, which can be harvested via OAI-PMH (e.g. by the VLO http://catalog.clarin.eu/vlo/).

Deposit workflow: A large part of the data of the repository is collected as part of a former project which had as the aim to collect resources and prepare metadata for them. New data can be added by researchers at a Danish research institution.

The repository implements an explicitly defined workflow described on our website in several pages.

The workflow consists of:

- Deposit
  - Guidance: see http://info.clarin.dk/deponer-resurser/vejledning/
    - You first need to obtain a depositor role. Contact address admin@clarin.dk.
    - Ensure that your data is valid. Resources need to be prepared in one of the formats that the repository accepts.
    - Package the data
    - Make data available at a web-location from where the deposit service can download the data
    - Log in using WAYF or another Identity Federation
    - Choose resource access (Public, Academic or Restricted)
    - Choose the file to deposit
    - Accept deposit
- Archiving
  - Description, see http://info.clarin.dk/en/overblik/datamanagement/
- Access

- Who can use the repository and access the data: see http://info.clarin.dk/overblik/hvem/
- How to search: see http://info.clarin.dk/soeg-resurser/soeg/ (CLARIN-DK), http://info.clarin.dk/soeg-resurser/vlo/ (VLO)
- Overview of Licences for viewing and downloading data: http://info.clarin.dk/overblik/licenser/

F.5 SWOT analysis

Only two parts of the SWOT analysis will be taken into account here: weaknesses and threats. The focus is to address gaps and challenges for the metadata to be well written and complete.

Weaknesses

A diversity of resource types gives a diversity of metadata information
In our work with metadata we have tried to collect metadata for a number of different resource types. This is both a challenges and a weakness of the repository.

When creating search and display interfaces it is an extra demanding task to make them display the metadata in a uniform or at least understandable way. Information expressed for one resource type will be neither relevant nor available for all other resource types.

A diversity of focus areas for the data providers and users
Even for the text resource type, it became clear that different researchers have a number of different views on the importance and needed details for some metadata.

As an example a user collecting texts for a corpus of special kind of texts, are tempted to give a text a description that states that this text is a part of this special corpus xyz with this description. Where the researcher that are working with a text from a certain author, will be more focussed on describing this specific text in the light of this authors work in the description metadata field.

Even when guidelines are given for adding metadata, this is not always solving this challenge as people have diverse foci on what to fill in.

Lack of knowledge and use of standards
The researcher will in a number of cases be unexperienced in the field of metadata creation and in the use of the specific standard. As the standards have been created with different goals and often to handle a lot of different types of situations, it is not simple to choose the right field to fill in with the right information.

For TEI P5 - as an example - only a few fields have specific restrictions about the content. The user can be given a more restricted schema for the data, but

we think that a specific metadata editor could be a good solution to limit the fields the user can fill in and to add restrictions to the values of these fields.

When the metadata editor is implemented, it will be possible to restrict the values of certain metadata types such as: *language*, *format*, *resource type* and the syntax for *date*. But for the open metadata fields it can still be difficult for the users to differentiate between the content of e.g.: *creator*, *data provider* and *publisher* or *title* and *source title*. Here good guidelines can be a good help.

Lack of validation of content of open metadata fields
As mentioned above, one weakness is that some metadata fields accept free text. Since there is no validation of the *content* of the metadata for these fields, variation or even wrong use of metadata opens up for deficient metadata search results.

Lack of explaining precisely the intension of a metadata field or explaining the suggested pick list
Only good guidelines when creating metadata and when doing search can help. These guidelines should be integrated with a metadata editor and the search interface to explain what is the intension of a metadata field or an explain of the suggested pick list.

Threats
Standards compliance is a difficult issue: standards might change and that might complicate things, and these changes are neither defined of the users nor the data centre. These kinds of changes are difficult to keep up with.

In the end there is nothing to prevent the user for writing rubbish in open fields as can happen when creating metadata for thousands of data records automatically. A random test set of resources has to be selected for validation for each data provider to make sure that the metadata are in a good shape and that nothing important is missing or misunderstood.

PART B APPENDICES

Appendix G: List of SSH Metadata Providers

CESSDA

Institute	OAI-PMH endpoint	Formats	Schema	no records
GESIS	http://oai.datacite.org/oai?verb=ListIdentifiers&metadataPrefix=oai_dc&set=GESIS.ARCHIV	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	6208
		oai_datacite	http://schema.datacite.org/oai/oai-1.0/oai.xsd	6210
SND	http://oai.datacite.org/oai?verb=ListIdentifiers&metadataPrefix=oai_dc&set=SND.SND	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	2223
		oai_datacite	http://schema.datacite.org/oai/oai-1.0/oai.xsd	2225
Språkbanken	http://spraakbanken.gu.se/resources/oai-pmh?verb=Identify	ddi_3_1	http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/XMLSchema/instance.xsd	116 116
DANS	http://easy.dans.knaw.nl/oai/?verb=ListMetadataFormats	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	28073
		carare	http://www.carare.eu/carareSchema	timeout
		nl_didl	http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd	timeout
UKDA	http://oai.esds.ac.uk/oai.asp	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	5741
		marc	http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd	5741
		DDI/Version1-2-2	http://www.icpsr.umich.edu/DDI/Version1-2-2.xsd	5741
	http://oai.datacite.org/oai?verb=ListIdentifiers&metadataPrefix=oai_dc&set=BL.UKDA	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	6279
		oai_datacite	http://schema.datacite.org/oai/oai-1.0/oai.xsd	6279
ISSDA	http://nesstar.ucd.ie/oai-pmh/	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	?
		DDI/Version1-2-2	http://www.icpsr.umich.edu/DDI/Version1-2-2.xsd	?
LiDA	http://www.lidata.eu/oaiprovider	marcxml	http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd	544
		oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	563
		DDI/Version1-2-2	http://www.icpsr.umich.edu/DDI/Version1-2-2.xsd	545
The Danish data archive	http://ddaonline.dda.dk/oai-pmh/?verb=Identify	DDI/Version1-2-2	http://www.icpsr.umich.edu/DDI/Version1-2-2.xsd	115

CLARIN

Centre	End-point	Metadata formats	website	Number of records
MPI for Psycholinguistics	http://corpus1.mpi.nl/ds/oaiprovider/oai2?verb=Identify	cmdi	http://corpus1.mpi.nl/	150000
DANS - Data Archiving and Networked Services, The Hague	http://oai.clarin-beta.dans.knaw.nl/OAIHandler?verb=Identify	cmdi	https://easy.dans.knaw.nl/ui/home	1000
Universität des Saarlandes	http://fedora.clarin-d.uni-saarland.de/oaiprovider/?verb=Identify	cmdi	http://fedora.clarin-d.uni-saarland.de/	108
Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)	http://fedora.dwds.de:8088/oaiprovider/?verb=Identify	cmdi	http://fedora.dwds.de/	1699
DK-CLARIN	http://clarin.dk/oaiprovider/?verb=Identify	cmdi	https://clarin.dk/clarindk/forside.jsp	13320
LINDAT, Charles University Prague	http://lindat.mff.cuni.cz/repository/oai/request?verb=Identify	cmdi	http://lindat.mff.cuni.cz/lindat/	1090
CSC, the Finnish IT Center for Science + University of Helsinki	http://metalb.csc.fi/cgi-bin/que?verb=Identify	cmdi	https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/FinClarinEsittely	231
INL	http://repository.dev.clarin.inl.nl/oai/oai_server.php?verb=Identify	cmdi	https://portal.clarin.inl.nl/	16
EKUT, Univ Tuebingen	http://weblicht.sfs.uni-tuebingen.de/oaiprovider/?verb=Identify, http://repository.sfb833.uni-tuebingen.de:8080/oaiprovider/?verb=Identify	cmdi	http://www.sfs.uni-tuebingen.de/ascl/clarin-center/repository.html	48
Meertens Institute	http://www.meertens.knaw.nl/oai/oai_server.php?verb=Identify	cmdi	http://www.meertens.knaw.nl/cms/en/	249658
Bayerisches Archiv für Sprachsignale	http://www.phonetik.uni-muenchen.de/cgi-bin/BASRepository/oaipmh/oai.pl?verb=Identify	cmdi	http://www.en.phonetik.uni-muenchen.de/research/bav_arch_spsig/index.html	22433
ASV Leipzig - Abteilung Automatische Sprachverarbeitung, Universität Leipzig	http://clarinoai.informatik.uni-leipzig.de:8080/oaiprovider/oai?verb=Identify	cmdi	http://asv.informatik.uni-leipzig.de/	6135
Hamburger Zentrum für Sprachkorpora (HZSK)	http://virt-fedora.multilingua.uni-hamburg.de:8080/oaiprovider/?verb=Identify	cmdi	http://virt-fedora.multilingua.uni-hamburg.de/drupal/fedora/repository	21
IDS - Institut für Deutsche Sprache, Mannheim	http://repos.ids-mannheim.de/oaiprovider/?verb=Identify	cmdi	http://repos.ids-mannheim.de/	22435
IMS, Universität Stuttgart	http://clarin04.ims.uni-stuttgart.de/oaiprovider/oai?verb=Identify	cmdi	http://www.ims.uni-stuttgart.de/forschung/projekte/ClarinD.html	30
CELR, Estonia	http://register.keeleressursid.ee/oaiprovider/oai?verb=Identify	cmdi	https://register.keeleressursid.ee/fedora/describe	55
CLARIN Center Vienna	http://clarin.oeaw.ac.at/oaiprovider?verb=Identify	cmdi	http://clarin.oeaw.ac.at/ccv/	7
Wroclaw University CLARIN-PL	https://clarin-pl.eu/oai/request?verb=Identify	cmdi	http://clarin-pl.eu/en/	37
Huygens- NG	http://oaipmh.huygens.knaw.nl/oai?verb=Identify	cmdi	https://www.huygens.knaw.nll/	1509

DARIAH

Institute	OAI-PMH Endpoint	Formats	Schema	Website	Number of records
DARIAH-EU	http://demo2.dariah.eu/colreg/OAIHandler	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://demo2.dariah.eu/colreg/	92

		dclap	http://schema.dariah.eu/colreg/dclap/dclap.xsd		92
Language Resource Portal, Austrian Academy of Sciences	http://clarin.oeaw.ac.at/oaiprovider	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://clarin.oeaw.ac.at/ccv	12
		cmdi_lexRes	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1290431694579/xsd		3
		cmdi_teiHdr	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1380106710826/xsd		0
		cmdi_textCorpus	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1290431694580/xsd		1
		cmdi_collection	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1345561703620/xsd		3
University of Vienna	http://fedora.phaidra.univie.ac.at/oaiprovider/	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	https://phaidra.univie.ac.at/	34460
		mets	http://www.loc.gov/METS/ http://www.fedora.info/definitions/1/0/mets-fedora-ext1-1.xsd		34567
University of Graz	http://gams.uni-graz.at/oaiprovider	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://gams.uni-graz.at/	9459
		oai_europeana	http://www.europeana.eu/schemas/ese/		9459
restricted access	http://www.dismarc.org/oai/index.php	?			
DAI (Deutsches Archäologisches Institut)	http://arachne.uni-koeln.de/OAI-PMH/oai-pmh.xml	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://arachne.uni-koeln.de/drupal/	331498
		rdf_dc	http://purl.org/NET/crm-owl		-
		prometheus	http://www.prometheus-bildarchiv.de/		-
		enrich	http://tei.oucs.ox.ac.uk/ENRICH/ODD/RomaResults/enrich.dtd		40
		origin	http://www.arachne.uni-koeln.de/formats/origin/		-
		cidoc_crm	http://purl.org/NET/crm-owl		-
		claros	http://purl.org/NET/crm-owl		-
		mets	http://www.loc.gov/METS/		-
		geo	http://www.arachne.uni-koeln.de/formats/geo/		98
		carare	http://www.arachne.uni-koeln.de/formats/carare/		100
Demo_instance_for_the_imeji_community	http://demo.imeji.org/fledgeddata/oai/	imeji	http://colab.mpdl.mpg.de/mediawiki/Imeji/item	http://demo.imeji.org/imeji/	100
		oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd		100
MPDL, MPG (Max-Planck-Gesellschaft) / Max Planck	http://pubman.mpdl.mpg.de/escidoc-oaiprovider/	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://pubman.mpdl.mpg.de/escidoc-oaiprovider/	3940

Society					
Konrad Zuse Internet Archive	http://zuse.zib.de/fledgeddata/oai/	imeji	http://colab.mpdl.mpg.de/mediawiki/Imeji/item	http://zuse.zib.de/	100?
		oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd		100
Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)	http://edoc.bbaw.de/oai2/oai2.php	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://edoc.bbaw.de/	2321
		epicur	http://www.persistent-identifier.de/xepicur/version1.0/xepicur.xsd		2317
		oai_pp	http://www.proprint-service.de/xml/schemes/v1/PROPRINT_METADATA_SET.xsd		2321
		xMetaDiss	http://files.dnb.de/standards/xmetadiss/xmetadiss.xsd		0
		xMetaDissPlus	http://files.dnb.de/standards/xmetadissplus/xmetadissplus.xsd		0
University of York, Archaeology Data Service	http://archaeologydataservice.ac.uk/oai/archives/OAIHandler	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://archaeologydataservice.ac.uk/archives/	431
		ads_archive	http://archaeologydataservice.ac.uk/catalogue/schema/archive.xsd		431
		nerc_dss	http://www.isotc211.org/2005/gmx/gmx.xsd		0
		medin_dpp	http://www.isotc211.org/2005/gmx/gmx.xsd		0
Journal_Metadata_Record_Catalogue	http://archaeologydataservice.ac.uk/oai/journals/OAIHandler	oph	http://www.editeur.org/onix/serials/SOH		9
	http://archaeologydataservice.ac.uk/oai/imagebank/OAIHandler	oai_imagebank	http://www.openarchives.org/OAI/2.0/oai_dc.xsd		1433
STAR_depot_national_des_theses_electroniques_francaises	http://staroai.theses.fr/OAIHandler	tef	http://www.abes.fr/abes/documents/tef/recommandation/tef_schemas.xsd	http://www.theses.fr/	31323
		oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd		31304
Calames OAI Serveur ABES	http://www.calames.abes.fr/oai/oai2.aspx	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://www.calames.abes.fr/pub/	184891
CLEO	http://oai.openedition.org/	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://www.openedition.org/	277270
		qdc	http://dublincore.org/schemas/xmls/qdc/2008/02/11/qualifieddc.xsd		277270
		mets	http://www.loc.gov/standards/mets/mets.xsd		8399
		tei	http://lodel.org/ns/tei.openedition.1.2.xsd		
		basictei	http://lodel.org/ns/tei.openedition.1.2.xsd		
SDLR-LPL	http://sldr.org/oai-pmh.php	olac	http://www.language-archives.org/OLAC/1.1/olac.xsd	http://sldr.org	233
		oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd		233

SIDIH portal (EtnoInfoLab)	http://www.etnoinfolab.org/oai/oai.php	oai_dc	http://www.openarchives.org/OAI/1.1/dc.xsd	http://www.etnoinfolab.org/	1754
	http://www.arzenal.si/oai/arzenal	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://www.arzenal.si/	error
ISN_ZRC_SAZU	http://isn3.zrc-sazu.si/etnofolk/OAI-2.0/oai.php	eef	http://etnofolk.aipberoun.cz/schema/eef.xsd	http://isn3.zrc-sazu.si/etnofolk/	241
		oai_dc	http://www.openarchives.org/OAI/1.1/dc.xsd		241
Sistory si OAI Repository	http://www.sistory.si/oai.php	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://www.sistory.si/	25873
e-codices	http://www.e-codices.unifr.ch/oai/oai.php	oai_dc			1173
Göttinger Digitalisierungszentrum (GDZ)	http://gdz.sub.uni-goettingen.de/oai2/	oai_dc		http://gdz.sub.uni-goettingen.de/gdz/	191529
		mets	http://www.loc.gov/mets/mets.xsd http://www.loc.gov/standards/mods/v3/mods-3-2.xsd		50894
Virtuelle Fachbibliothek Kunstgeschichte	http://archiv.ub.uni-heidelberg.de/artdok/cgi/oai2	oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	http://archiv.ub.uni-heidelberg.de/artdok/	2670
		XMetaDissPlus	http://files.dnb.de/standards/xmetadissplus/xmetadissplus.xsd		2667
		didl	http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd		2667
		epicur	http://www.persistent-identifier.de/xepicur/version1.0/xepicur.xsd		2667
		europeana_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd		2667
		oai_wgl	http://www.openarchives.org/OAI/2.0/oai_dc.xsd		2667
		rdf	http://www.openarchives.org/OAI/2.0/rdf.xsd		2667
		uketd_dc	http://naca.central.cranfield.ac.uk/ethos-oai/2.0/uketd_dc.xsd		2667
Visual Library Server der Universitäts- und Landesbibliothek Düsseldorf	http://digital.ub.uni-duesseldorf.de/theaterzettel/oai	epicur	http://www.persistent-identifier.de/xepicur/version1.0/xepicur.xsd	http://archiv.ub.uni-heidelberg.de/artdok/	error
		oai_dc	http://www.openarchives.org/OAI/2.0/oai_dc.xsd		error
		mets	http://www.loc.gov/standards/mets/mets.xsd		error
		mods	http://www.loc.gov/standards/mods/v3/mods-3-0.xsd		error
		marcxml	http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd		error

Appendix H: List of facets with Definitions

Facets/Fields	Semantic definition of the catalogue's properties	Facet or Field	Included
Collection	A named aggregation of resources. A body which the resource is part of.	field	Y
Contributor	An entity, that is: a person, an organisation, or a service, participating in the creation of a resource.		Y
Country	Country where the data described by the metadata was collected or created	facet	Y
CreationDate	The date at which the data described by the metadata was created.	facet, if a suitable UI element for selection can be provided else may to CreationYear	Y
Creator	Writer, generator or producer of the resource described by the metadata. Since multiple authors can be associated with a resource, the field can have multiple values. In creating a list of authors, the policy is to avoid duplicates.	facet	Y
DataProvider	The organization or archive that makes the data available.	facet	Y
Discipline	Primary audience for the resource described by the metadata, a specific branch of knowledge.	facet	Y
Format	The organisation of information according to preset specifications (computer processing) or tradition (books, physical representation of content). For example: file format, physical medium, or	field	Y

	dimensions. Multiple field values are allowed, since sometimes a metadata description applies to more than one resource.		
Language	Specification of the language used in the resource. For example, an English book about Japanese language, will have 'English' as the value of the language field.	facet	Y
MetadataSchema	A URI to an elaborate description of the form the metadata can take. This could be a website, a document, or a more formal normative schema backing up description.	field	Y
Notes	A description or abstract of the data set. For ckan itself, the field is optional.	field	N
ID	An unique identifier for the resource.	field	N
PublicationDate	The date at which the resource was made known to the public.	facet, if a suitable UI element for selection can be provided else may to PublicationYear	N
ResourceType	Broad categorisation of the subject of description into for example: text image, audio, video, object, collection, event, interaction, and or data set.	facet	N
Rights	Any rights information for this resource (from DC)	field	N

SpatialCoverage	The extent or scope in space about which the data is focused or where the data was gathered. Multiple field values could be mapped.	field	N
Subject	A representation of the resource in term of keywords, key phrases, or classification codes. Preferably these should be part of a closed vocabulary.	facet	N
TemporalCoverage	The space of time the data refers to with regard to contents or date (or period) of the data collection. The coverage is expressed by two dates: the begindate and the enddate. The dates need to be specified in UTC format. Only a single value can be associated with each of these fields.	field	N
Title	A name given to the resource (from DC).	field	N
URL	A URL for the source of a data set.	field	Y

The latest version of the facets & field document can be found at https://github.com/DASISH/jmd-documents

Appendix I: List of Mappings

This appendix consists of a list of descriptions of mappings used in the metadata catalog. Here, 'mapping' means an association of a type of metadata external to the catalogue with the fields that are known within the catalog. In the software, the mappings take the form of XML files feeding the mapper component, that part of the catalog responsible for converting the external metadata to what a user will see appearing in the catalog. Both the formal mapfiles and mapping descriptions are available on Github here[124] because the development of the catalog will not cease after the publication of the report. The latest versions of the mapping descriptions & definitions in the tables below and their implementation can be found at:

DC mapping

DC or DCMI

Is usually mean to mean the set of 15 original elements. In the course of the years many new elements have been added (DCTerms). OAI-PMH requires only the core-elements to be available.

Some of the facets, on mapping, do not receive a value. The cause of this is twofold. Firstly, there might simply not be a term defined by the Dublin Core standard that can be mapped onto it, or secondly, the data harvested does not provide suitable values for the facet.

DC definition referred to: http://dublincore.org/documents/dcmi-terms/ issued on 2012-06-14

DC schema usually encountered: http://www.openarchives.org/OAI/2.0/oai_dc.xsd

Property in ckan	**DC Element**		**DC Definition**
	Property		
Collection	-	-	-
Contributor	Contributor	/dc:contributor	An entity responsible for making contributions to the resource.
Country	-	-	-
CreationDate	Date	/dc:created or /dc:date	Date: A point or period of time associated with an event in the lifecycle of the resource. Created (refined): Date of creation of

[124] [descriptions] https://github.com/DASISH/md-mapping/tree/master/doc and [mapfiles] https://github.com/DASISH/md-mapping/mapfiles

			the resource.
Creator	Creator	/dc:creator	An entity primarily responsible for making the resource.
DataProvider	Publisher	/dc:publisher	An entity responsible for making the resource available.
Discipline	-	-	-
Format	Format	/dc:format (except values that contain the term 'Bytes')	The file format, physical medium, or dimensions of the resource.
Language	Language	/dc:language	A language of the resource.
MetadataSchema	http://www.openarchives.org/OAI/2.0/oai_dc.xsd	-	-
Notes	Description	/dc:description	An account of the resource.
PID	Identifier	/dc:identifier	An unambiguous reference to the resource within a given context.
PublicationDate	-	No equivalent in DC Core, DCterms has "issued"	The year when the data was or will be made publicly available.
ResourceType	Type	/dc:type	The nature or genre of the resource.
Rights	Rights	/dc:rights	Any rights information for this resource.
SpatialCoverage	-	-	

Subject	Subject	/dc:subject	The topic of the resource.
TemporalCoverage	-	-	-
Title	Title	/dc:title	A name given to the resource.
URL	Identifier	/dc:identifier (provided the value conforms to an URI)	An unambiguous reference to the resource within a given context.

DataCite 3.0 mapping based on Joan Starr, Jan Ashton, Amy Barton, Jannean Elliott, Marie□Christine Jacquemot□Perbal, Merja Karjalainen, Lynne McAvoy, Elizabeth Newbold, Sebastian Peters, Madeleine de Smaele, Natalija Schleinstein, Wolfgang Zenk-Möltgen, and Frauke Ziedorn: DataCite metadata schema for the publication and citation of research data: Version 3.0., 2013, doi:10.5438/0008 and http://schema.datacite.org/meta/kernel-3/metadata.xsd				
Property in ckan	**DataCite 3.0 Element**			**DataCite 3.0 Definition**
	ID	**Property**		
Collection	17	Description, if descriptionType 17.1=seriesInformation	resource/descriptions/description[@descriptionType=”SeriesInformation”]	Information about a repeating series, such as volume, issue, number.
Contributor	7	Contributor	-	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.
	7.2	contributorName	resource/contributors/contributor/contribu	The name of the contributor.

			torName	
Country	-	-	-	-
Creation Date	8	Date, if dateType 8.1 = created	resource/dates/date[@dateType="Created"]	The date the resource itself was put together.
Creator	2	Creator	-	The main researchers involved in producing the data, or the authors of the publication, in priority order.
	2.1	creatorName	resource/creators/creator/creatorName	The name of the creator.
DataProvider	4	Publisher	resource/publisher	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role.
Discipline	-	-	-	-
Format	14	Format	resource/formats/format	Technical format of the resource.
Language	9	Language	resource/language	The primary language of the resource.
Metadata Schema	-	http://schema.datacite.org/meta/kernel-3/metadata.xsd	-	-

Notes	17	Description, if descriptionType 17.1=Abstract	resource/descriptions/description[@descriptionType="Abstract"]	A brief description of the resource and the context in which the resource was created.
PID	1	Identifier	resource/identifier[@identifierType="DOI"]	The Identifier is a unique string that identifies a resource. Currently only DOI is allowed
PublicationDate	5	PublicationYear	resource/publicationYear	The year when the data was or will be made publicly available.
ResourceType	10.1	ResourceTypeGeneral	resource/resourceType[@resourceTypeGeneral]	The general type of a resource.
Rights	16	Rights	resource/rightsList/rights	Any rights information for this resource.
SpatialCoverage	18	GeoLocation	-	Spatial region or named place where the data was gathered or which the data is about.
	18.3	geoLocationPlace	resource/geoLocations/geoLocation/geoLocationPlace	Description of a geographic location.
Subject	6	Subject	resource/subjects/subject	Subject, keyword, classification code, or key phrase describing the resource.
TemporalCoverage	8	Date, if dateType 8.1 = collected	resource/dates/date[@dateType="Collected"]	The date or date range in which the resource content was

				collected.
Title	3	Title	resource/titles/title	A name or title by which a resource is known.
URL	1	Identifier	resource/identifier[@identifierType="DOI"]	The identifier is a unique string that identifies a resource.

based on www.ddialliance.org/sites/default/files/Version1-2-2.dtd and http://www.icpsr.umich.edu/DDI/Version1-2-2.xsd				
Property in ckan	**DDI 1.2.2 Element**			**1.2.2 Definition**
	ID	**Property**		
Collection	A.6.5.1	Series Name (serName)	d:codeBook/d:stdyDscr/d:citation/d:serStmt/d:serName	The name of the data/documentation series to which the data/documentation belongs.
Contributor	A.6.2.2	Other identifications and acknowledgments (othId)	d:codeBook/d:docDscr/d:citation/d:rspStmt/d:othId	Give statements of responsibility not recorded in the title statement of responsibility area. Make notes on persons bodies connected with the work, or significant persons or bodies connected with previous editions and not already named in the description.
Country	2.2.3.3	Country (nation)	d:codeBook/d:stdyDscr/d:stdyInfo/d:sumDscr/d:nation	Indicates the country or countries covered in the file.
CreationDate	A.6.3.3	Date of Production (prodDate)	d:codeBook/d:docDscr/d:citation/d:prodStmt/d:prodDate	Date the data/documentation was produced (not archived).

Creator	A.6.2.1	Authoring Entity (AuthEnty)	d:codeBook/d:docDscr/d:citation/d:rspStmt/d:AuthEnty	The person, corporate body, or agency responsible for the data/documentation's substantive and intellectual content.
DataProvider	A.6.3.1	Producer (producer)	d:codeBook/d:docDscr/d:citation/d:prodStmt/d:producer	The producer is the person or organization with the finacial or administrative responsibility for the physical processes whereby the data/documentation is brought into existence.
Discipline	-	-	-	-
Format	3.1.5	Type of File (fileType)	d:codeBook/d:fileDscr/d:fileType	Types of data files include raw data (ASCII, EBCDIC, etc.) and software-dependent files such as SAS datasets, SPSS export files, etc.
Language	0.0	xml:lang attribute for codeBook element	d:codeBook/@xml:lang	(There is no place for getting the language for the resource. The only place would be the top level elements xml:lang attribute which describes the general language of the metadata.)
MetadataSchema	-	http://www.icpsr.umich.edu/DDI/Version1-2-2.xsd	-	-
Notes	2.2.2	Abstract (abstract)	d:codeBook/d:stdyDscr/d:stdyInfo/d:abstract	An unformatted summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and what questions the PIs attempted to answer when they conducted the study. A listing of major variables in the

				study is important here.
PID	A.6.1.5	Identification Number (IdNo)	d:codeBook/d:docDscr/d:citation/d:titlStmt/d:IDNo	Unique string or number (producer's/archive's number).
	A.6.2.2	Other identifications and acknowledgments (othId), if type=DOI	d:codeBook/d:stdyDscr/d:citation/d:rspStmt/d:othId[@type="DOI"]	-
PublicationDate	A.6.4.5	Date of Distribution (distDate)	d:codeBook/d:docDscr/d:citation/d:distStmt/d:distDate	The date the data/documentation became operational in a computerized form and available for distribution/presentation.
ResourceType	2.2.3.8	Kind of Data (dataKind)	d:codeBook/d:stdyDscr/d:stdyInfo/d:sumDscr/d:dataKind	The type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data/ratings, process-produced data, etc.

Rights	A 6.3.2	Copyright (copyright)	codeBook/d:stdyInfo/d:citation/d:prodStmt/d:copyright	Copyright Statement.
	2.4.1.3	Availability Status (avlStatus)	codeBook/d:stdyDscr/d:dataAccs/d:setAvail/d:avlStatus	Statement of collection availability. An archive may need to indicate that a collection is unavailable because it is embargoed for a period of time, because it has been superseded, because a new edition is imminent, etc.
Spatial Coverage	2.2.3.4	Geographic Coverage (geogCover)	d:codeBook/d:stdyDscr/d:stdyInfo/d:sumDscr/d:geogCover	Information on the geographic coverage of the data. Include the total geographic scope of the data, and any additional levels of geographic coding provided in the variables.
Subject	2.2.1.1	Keywords (keyword)	d:codeBook/d:stdyDscr/d:stdyInfo/d:subject/d:keyword	Words or phrases that describe salient aspects of a data collection's content. Can be used for building keyword indexes and for classification and retrieval purposes. A controlled vocabulary can be employed
	2.2.1.2	Topic Classification (topcClas)	d:codeBook/d:stdyDscr/d:stdyInfo/d:subject/d:topcClas	The classification field indicates the broad substantive topic(s) that the data cover.
TemporalCoverage	2.2.3.1	Time Period (timePrd)	d:codeBook/d:stdyDscr/d:stdyInfo/d:sumDscr/d:timePrd	The time period to which the data refer. This item reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected.

Title	A.6.1.1	Title (titl)	d:codeBook/d:docDscr/d:citation/d:titlStmt/d:titl	Contains the full authoritative title of the data/documentation.
URL	2.4.1.1	Location (accsPlac)	d:codeBook/d:stdyDscr/d:dataAccs/d:setAvail/d:accsPlac/@URI	Location where the data collection is currently stored.

based on http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/XMLSchema and http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/XMLSchema/FieldLevelDocumentation		
Property in ckan	**DDI 3.1 Element**	**DDI 3.1 Definition**
Collection	s:StudyUnit/r:SeriesStatement/r:SeriesName	Series name
Contributor	s:StudyUnit/r:Citation/r:Contributor	The name of a contributing author or creator, who worked in support of the primary creator given above.
Country	s:StudyUnit/c:ConceptualComponent/c:GeographicLocationScheme/r:GeographicLocation[r:Values/r:GeographyValue/r:GeographyCode/r:Value/@codeListID="ISO3166-1"]/r:Values/r:GeographyValue/r:GeographyName	Textual description of the particular geographic entity/code.
CreationDate	s:StudyUnit/d:DataCollection/d:CollectionEvent/d:DataCollectionDate/r:StartDate	Start of a date range.
Creator	s:StudyUnit/r:Citation/r:Creator	Person, corporate body, or agency responsible for the substantive and intellectual content of the described object.
	a:Archive/a:OrganizationScheme/a:Individual/a:IndividualName	Full name of the individual.
	a:Archive/a:OrganizationScheme/a:Organization/a:Individual/a:IndividualName	Full name of the individual.

	a:Archive/a:OrganizationScheme/a:Organization/a:OrganizationName	The official name of the organization.
DataProvider	s:StudyUnit/r:Citation/r:Publisher	Person or organization responsible for making the resource available in its present form.
Discipline	-	-
Format	s:StudyUnit/pd:PhysicalDataProduct/pd:PhysicalStructureScheme/pd:PhysicalStructure/pd:Format	Description of the physical format of data file (e.g., SAS save file, Delimited file, Fixed format file).
	s:StudyUnit/a:Archive/a:ArchiveSpecific/a:Item/a:Format	Describes the item's format. Can be repeated to support different languages.
	s:StudyUnit/r:OtherMaterial/r:MIMEType	Provides a standard Internet MIME type for use by processing applications.
Language	s:StudyUnit/r:Citation/r:Language	Language of the intellectual content of the described object, expressed either as a two-character ISO language code or as a pair of two-character codes indicating language and locale, as per ISO 3166.
MetadataSchema	http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/XMLSchema/instance.xsd	-
Notes	s:StudyUnit/r:Abstract/r:Content	A human-readable abstract of the study unit describing the nature and scope of the data collection, special characteristics of its content.
PID	s:StudyUnit/r.Citation/r:InternationalIdentifier	ISBN, ISSN or similar designator.

Publication Date	s:StudyUnit/r:Citation/r:PublicationDate/r:SimpleDate	The date of publication.
ResourceType	s:StudyUnit/s:KindOfData	Describes, with a string or a term from a controlled vocabulary, the kind of data documented in the logical product(s) of a study unit. Examples include survey data, census/enumeration data, administrative data, measurement data, assessment data, demographic data, voting data, etc.
Rights	s:StudyUnit/a:Archive/a:Access/a:AccessConditions	Describes conditions for access.
	s:StudyUnit/r:Citation/r:Copyright	The copyright statement.
SpatialCoverage	s:StudyUnit/r:Coverage/r:SpatialCoverage/r:Description[@xml:lang='en']/text() If no lang: s:StudyUnit/r:Coverage/r:SpatialCoverage/r:Description	Provides a human-readable summary of the information included in Geography and Geography Reference. It may include information on all levels of spatial coverage, in addition to the overall coverage. This field can map to Dublin Core Coverage, which does not support structured strings.
Subject	s:StudyUnit/r:Coverage/r:TopicalCoverage/r:Subject	A subject or list of subjects that indicate the topical coverage of the data described in a particular module/section.

	s:StudyUnit/r:Coverage/r:TopicalCoverage/r:Keyword	A keyword (which can be supplied in multiple language-equivalent forms) to support searches on topical coverage.
TemporalCoverage	s:StudyUnit/r:Coverage/r:TemporalCoverage/r:ReferenceDate/r:SimpleDate	A single point in time.
	s:StudyUnit/r:Coverage/r:TemporalCoverage/r:ReferenceDate/r:StartDate and s:StudyUnit/r:Coverage/r:TemporalCoverage/r:ReferenceDate/r:EndDate	Start of a date range. End of a date range.
Title	s:StudyUnit/r:Citation/r:Title	Full authoritative title. Field may be repeated to document multiple languages.
URL	s:StudyUnit/a:Archive/a:ArchiveSpecific/a:Collection/a:URI	The URL or URN for the collection.

based on data provided by GAMS, Graz
http://gams.uni-graz.at/oaiprovider/?verb=ListRecords&metadataPrefix=oai_europeana
Based on Europeana Schema Elements (ESE – the old europeana data model http://www.europeana.eu/schemas/ese/) using fields also from dublincore and dublincore terms schemas

Please note that Europeana ESE definitions consist of Dublin Core/terms definitions and comments and notes added to these.

Property in ckan	Fields in ESE	ESE definition
Collection	dcterms:isPartOf	A related resource in which the described resource is physically or logically included. Refinement of dc:relation. See also dcterms:hasPart. Note: Use for the name of the collection which the digital object is part of.
Contributor	dc:contributor	An entity responsible for making contributions to the resource.

		Note: The name of contributors to the original analog or born digital object. This could be a person, an organisation or a service. Map each name to a separate repeated contributor element if possible. Ideally choose a preferred form of name from an authority source. If you do not use an authority source, use a consistent form of the name e.g. Shakespeare, William.
Country	dcterms:spatial	Spatial characteristics of the resource. Refinement of dc:coverage. Note: Information about the spatial characteristics of the original analog or born digital object, i.e. what the resource represents or depicts in terms of space. This may be a named place, a location, a spatial coordinate or a named administrative entity.
CreationDate	-	-
Creator	(dc:contributor)	See above.
DataProvider	europeana:dataProvider europeana:provider dc:publisher	
Discipline	-	-
Format	-	-
Language	dc:language	A language of the resource. Comment: The recommended best practice is to use a controlled vocabulary such

		as RFC 4646 (http://www.rfc-archive.org/getrfc.php?rfc=4646) which, in conjunction with ISO 639, defines two- and three-letter primary language tags. Either a coded value or text string can be represented here. Note: Use this element for the language of textual objects and also where there is a language aspect to other objects e.g. sound recordings, posters, newspapers etc). If there is no language aspect to the digital object (e.g. a photograph), please ignore this element. This element is not for the language of the metadata of a resource, which may be described in xml:lang attribute. See europeana:language.
MetadataSchema	http://www.europeana.eu/schemas/ese/ http://www.europeana.eu/schemas/ese/ESE-V3.4.xsd	
Notes	-	-
PID	dc:identifier	An unambiguous reference to the resource within a given context. Note: Use europeana:isShownBy for the URL of the provided digital object. If the URL is already included in the dc:identifier element in the existing metadata, keep it and repeat the information in europeana:isShownBy.

PublicationDate	-	-
ResourceType	dc:type, europeana:type	The nature or genre of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content. The recommended best practice is to select a value from a controlled vocabulary (for example, the DC Type vocabulary is available at http://dublincore.org/documents/dcmi-type-vocabulary/). Note: The type of the original analog or born digital object as recorded by the content holder, this element typically includes values such as photograph, painting, sculpture etc. Although the portal needs normalised values to support type-related functions it is desirable to keep the original local values as well. Thus, all these original values should be mapped to this element. A separate europeana:type element has been added to contain the normalised value.
Rights	europeana:rights	Information about rights held in and over the resource. Note: This is a free text element and should be used for information about intellectual property rights or access arrangements for the digital object that is **additional** to the controlled value provided in europeana:rights.

		Compare the use of this element with europeana:rights before making a mapping decision. A record may contain both elements but do not duplicate values in both elements: <europeana:rights>http://www.europeana.eu/rights/rr-f/</europeana:rights> <dc:rights>Kilmarnock House Trust (David Jones)</dc:rights>
SpacialCoverage	dcterms:spatial	Spatial characteristics of the resource. Refinement of dc:coverage. Note: Information about the spatial characteristics of the original analog or born digital object, i.e. what the resource represents or depicts in terms of space. This may be a named place, a location, a spatial coordinate or a named administrative entity.
Subject		
TemporalCoverage	dcterms:temporal	Temporal characteristics of the resource. Refinement of dc:coverage Note: The temporal characteristics of the original analog or born digital object, i.e. what the resource is about or depicts in terms of time. This may be a period, date or date range.
Title	dc:title	A name given to the resource. Typically, a Title will be a name by which the resource is formally known. Refined by:

		dcterms:alternative Note: The title of the original analog or born digital object. Please use xml:lang attributes for direct translations of the title.
URL	europeana:isShownAt, europeana:object	An unambiguous URL reference to the digital object on the provider's web site in its full information context. Refinement of dc:relation. See also europeana:isShownBy. Note: This is a URL that will be active in the Europeana interface to give access to the provided digital object displayed on the provider's web site in its full information context. Use europeana:isShownAt if you display the digital object with extra information (such as header, banner etc)or if the object is only accessible by clicking another icon on the local page or for digital objects embedded in HTML pages (even where the page is extremely simple).

Appendix J: CKAN Performance Tuning

CKAN is used by a number of organizations and governments including the UK, Canada and US governments. The UK Government (data.gov.uk) uses CKAN to provide a central access to government data with the objective of making data "easy to find, easy to license, and easy to re-use". The Canadian government (data.gc.ca) uses CKAN to provide one-stop access to the Government of Canada's searchable open data with the objective of enhancing transparency and accountability. Similarly, the US Federal Government (data.gov) uses CKAN to provide a single portal where data from different portals, sources and catalogs (over 200 publishing organizations) is displayed in a standardized user interface allowing users to search, filter and facet through thousands of datasets.

All of these public CKAN installations have datasets in thousands, not in millions. The Canadian national portal (data.gc.ca) has less than 200,000 datasets. The US government portal (data.gov) has less than 100,000 datasets and the UK government portal (data.gov.uk) has less than 20,000 datasets.

Before making a choice for CKAN as a platform, we tested CKAN performance with respect to size and speed. With our tests, we found that CKAN with default configuration performs adequately with ten thousands of datasets (records) but became too slow with millions of records (see Fig. 30). To make it work for millions of datasets, performance optimization is necessary.

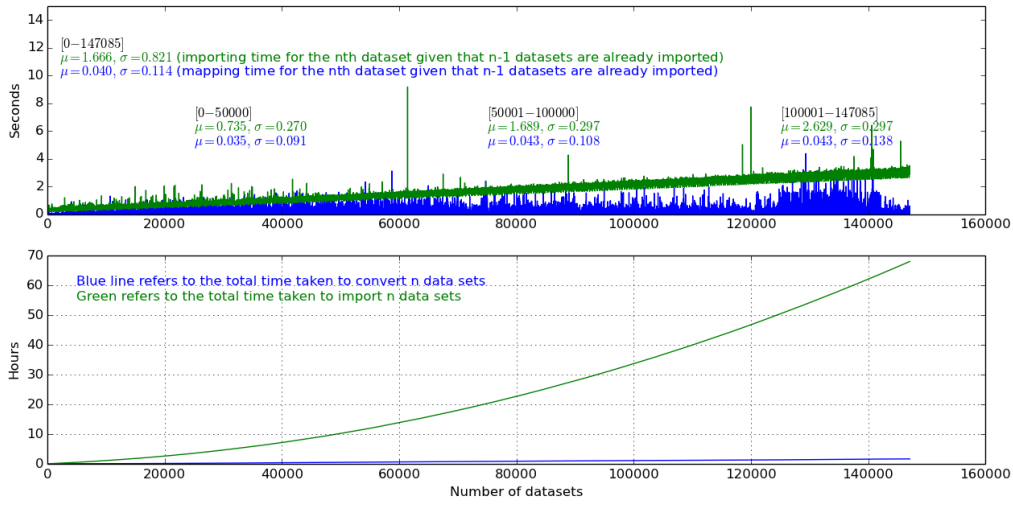


Figure 30: CKAN performance with default configuration values

time taken to import datasets into CKAN as a function of the number of datasets already imported. For the generation of this figure, CKAN is used with default configuration values.

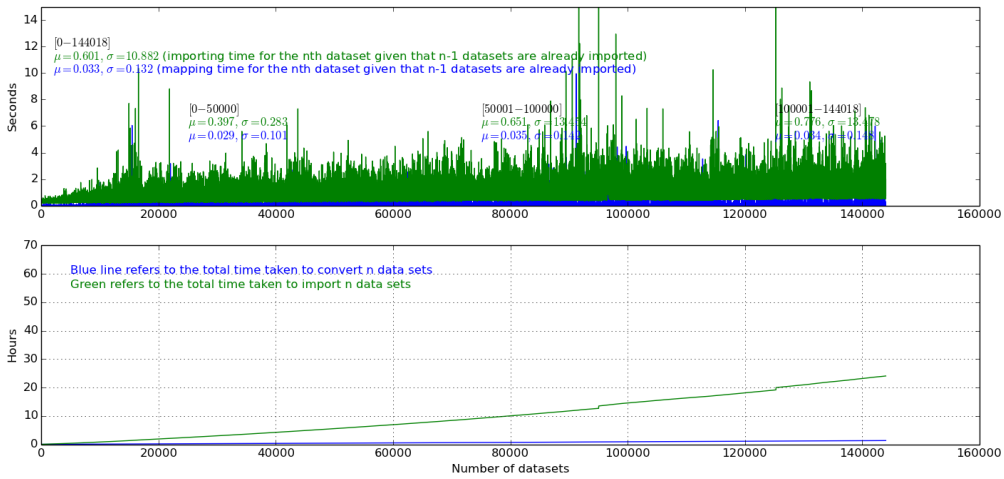


Figure 31: CKAN performance after configuration changes

time taken to import datasets into CKAN as a function of the number of datasets already imported. For the generation of this figure, CKAN is used after configuration changes (performance tuning).

Hence, we carried out such performance tuning at three levels. First, we changed the CKAN configuration file. The changes here involve delaying Solr indexing/committing and stopping activity streaming. Second, we changed some designs in the PostgreSQL database tables based on tips from and our observations. The changes here involve removing constraints and adding/removing database indexes). Third, we changed a few PostgreSQL (postgresql.conf) configurations to take advantage of available memory and CPU. With these changes, we imported 2 million datasets into CKAN in less than 2 weeks. Without these changes, it would have taken over a year. This we estimated based on a trend seen on 150,000 datasets on a machine with 8GB RAM and 2.67Ghz CPU (dual core).

The three level performance changes are shown in detail below.

1. CKAN config file changes. The purpose of these changes is to delay or stop unnecessary operations.
    - ckan.search.automatic_indexing = false
    - ckan.search.solr_commit = false
    - ckan.activity_streams_enabled = false

2. Postgres database design changes. The purpose of these changes is to remove constraints and add or drop indexes. Indexes are added to make select sql statements faster. Indexes are dropped to make insert sql statements faster. Depending on which operations we do more of, we can decide to add or drop indexes.
    - (sudo -u postgres psql ckan_default # login to database ckan_default)

- \i <path-to-sql-statements>constraints.sql # constraints.sql is a file containing sql statements [https://github.com/okfn/ckan/wiki/Performance-tips-for-large-imports]
- \i <path-to-sql-statements>what_to_alter.sql # what_to_alter.sql is a file containing sql statements [https://github.com/okfn/ckan/wiki/Performance-tips-for-large-imports]
- \i <path-to-sql-statements>eric_create_indexes.sql # from Eric, a colleague at MPI. <eric_create_indexes.sql> is a file that contains the following sql statement
  - create INDEX idx_pr_package_id ON package_role (package_id);
  - create INDEX idx_mr_table_id ON member_revision (table_id);
  - create INDEX idx_mr_continuity_id ON member_revision (continuity_id);
  - create INDEX idx_mr_revision_id ON member_revision (revision_id);
  - create INDEX idx_per_revision_id ON package_extra_revision (revision_id);
  - create INDEX idx_pe_package_id ON package_extra (package_id);

3. Postgresql config file changes. The purpose of these changes is to take advantage of the available memory. Make the following changes to postgresql configuration file named postgresql.conf
   - shared_buffers = 200MB (default is 24MB) # for caching
   - work_mem = 512MB (default is 1MB) # for in-memory sorts,...
   - maintenance_work_mem = 512MB (default is 16MB) # memory for operations like VACUUM, CREATE INDEX, and ALTER TABLE ADD FOREIGN KEY
   - max_stack_depth = 6MB (default is 2MB) # memory for stack-based operations
   - synchronous_commit = off (default is on)
   - full_page_writes = off (recommended when synchronous_commit is off)
   - checkpoint_segments = 64 (default is 3) # for heavy-memory disk writes
   - effective_cache_size = 4096MB (default is 128MB) # memory for disk-caching
   - log_min_duration_statement = 1000 (default is -1 (disabled)) # logging statements longer 1000 milliseconds
   - log_temp_files = 256 (default is -1 (disabled)) # logs temp files bigger that 256K

Appendix L: Normalization

The normalization (also referred to as harmonization) postprocessing script performs two tasks: replace an old value with a new one and change date formats to UTC format (YYYY-MM-DDThh:mmTZD). Important input to the normalization script is a configuration file. The user writes rules or actions to be performed on a specified set of datasets.

Usage
From the user point of view, the usage and configuration of the harmonizer (postprocessor) should be as simple as possible. In this spirit, the harmonizer is designed to be a command to be executed with the following three arguments.

Input: *JSON file* (the input comes from the output of the mapper)
Input: *Configuration file* (this is a text file, where actions or rules are specified, see below for more)
Output: *JSON file* (the output of the postprocessor is another JSON file ready to be validated and/or uploaded to CKAN)

Configuration file
Out of the three arguments, the configuration file needs more explanation. The purpose of the configuration text file is to inform the postprocessor what actions to take on which dataset or datasets. For example, if we want the postprocessor to change language values (such as 'en', 'eng', 'fr', 'fre','de','ger') to a closed vocabulary set (such as 'English', 'French','German') , we should be able to edit the configuration file and specify the actions to be performed. For the configuration file to be understood and edited by humans and still be understood by the postprocessor, it has to have a simple format. An action or a rule is a line and has the following configuration format.

GroupName,,datasetName,,facetName,,old_value,,new_value,,action

Each line has six fields separated with a double comma (,,). The purpose of the double comma is to make it possible for easy parsing by the postprocessor (not too verbose for humans). Each field has the following semantics.

GroupName - this field specifies to which datasets (i.e. group name) an action should be taken on (for example, CLARIN).

datasetName - this field specifies which dataset an action should be taken on.

facetName - this field specifies which facet an action should be taken on (for example, language).

old_value - this field specifies what the old value of the given facet is.

new_value - this field specifies the new value of the old_value of the given facet

action - this field specifies the action (e.g. replace, change2UTC-format, etc)

Examples
How do we use the above configuration in practice?

Here is an example of how to specify rules to change different language codes to a closed vocabulary.

```
*,,*,,Language,,en,,English,,replace
*,,*,,Language,,eng,,English,,replace
*,,*,,Language,,fr,,French,,replace
*,,*,,Language,,fre,,French,,replace
*,,*,,Language,,de,,German,,replace
*,,*,,Language,,ger,,German,,replace
```

The first line says for any groupName and for any datasetName with a facetName="Language", take the action "replace", which takes "en" replaces it by "English". The rest of the lines can be interpreted in a similar way.

Changing date to UTC format
,,,,PublicationTimestamp,,*,,UTC,,changeDateFormat

The above line in a configuration file tells the harmonizer to change any date format to UTC. More specifically, the line says for any group, for any dataset and for facet PublicationTimestamp, take "changeDateFormat" action from any date format to UTC. The harmonizer uses a regular expression to check the date format of the given time-related facet (for example, PublicationTimestamp). If the date is already in UTC format, then it extracts the most relevant parts and returns a new date in the following format: YYYY-MM-DDThh:mm:ssZ. If the date is not already in UTC format, it extracts the YYYY part and appends to it -07-01T11:59:59Z (July first). If YYYY cannot be extracted, the date field is set to empty.