Interdisciplinary Interoperability

Nikolaos Beer, Kristin Herold, Wibke Kolbmann, Thomas Kollatz, Matteo Romanello, Sebastian Rose, Niels-Oliver Walkowski

DARIAH-DE Working Papers Nr. 3 (DARIAH-DE Report)

Beer, N. ; Herold, K. ; Kolbmann W. ; Kollatz, T. ; Romanello, M. ; Rose, S. ; Walkowski, N.-O.:
Interdisciplinary Interoperability
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2014
(DARIAH-DE working papers 3)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2014-1
URN: http://nbn-resolving.de/urn.nbn.de.gbv:7-dariah-2014-1-0

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE Working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Stefan Schmunk und Christof Schöch

Abstract: The exchange and reusability of data used for research in the humanities is one of the goals of DARIAH. To increase the interoperability of data sets between disciplines we present an overview and recommendations of measures to achieve this. We account for the finding and fetching of data with legal aspects in mind. This is achieved through standardized methods of discovery and transfer via interfaces on the web. Furthermore, we consider syntactic and semantic interoperability of data for use in different fields of study. Standardized metadata sets are one way to achieve this and we present some of them in this paper. The importance for scholars to find and be able to process data that is be relevant to their work is the main motivation of this document and for the aspects of the digital humanities covered within. We present options for each of the four aspects that we identified (APIs and Protocols, Standards, Identifiers and Licensing).

Keywords: Digital Humanities, Interoperabilität, Metadaten, Standards
Digital Humanities, Interoperability, Metadata, Standards

Interdisciplinary Interoperability

Nikolaos Beer (Musikwissenschaftliches Seminar, Detmold/Paderborn), Kristin Herold (Musikwissenschaftliches Seminar, Detmold/Paderborn), Wibke Kolbmann (DAI), Thomas Kollatz (STI), Matteo Romanello (DAI), Sebastian Rose (HKI), Niels-Oliver Walkowski (BBAW)

Nikolaos Beer, Kristin Herold, Wibke Kolbmann, Thomas Kollatz, Matteo Romanello, Sebastian Rose, Niels-Oliver Walkowski: "Interdisciplinary Interoperability". *DARIAH-DE Working Papers* Nr. 3. Göttingen: DARIAH-DE, 2014.

URN: urn:nbn:de:gbv:7-dariah-2014-1-0.

Dieser Beitrag ist ursprünglich im Februar 2013 als Report 3.3.1 im Rahmen von DARIAH-DE (BMBF, Förderkennzeichen 01UG1110A-M) entstanden.

Abstract

The exchange and reusability of data used for research in the humanities is one of the goals of DARIAH. To increase the interoperability of data sets between disciplines we present an overview and recommendations of measures to achieve this. We account for the finding and fetching of data with legal aspects in mind. This is achieved through standardized methods of discovery and transfer via interfaces on the web. Furthermore, we consider syntactic and semantic interoperability of data for use in different fields of study. Standardized metadata sets are one way to achieve this and we present some of them in this paper. The importance for scholars to find and be able to process data that is be relevant to their work is the main motivation of this document and for the aspects of the digital humanities covered within. We present options for each of the four aspects that we identified (APIs and Protocols, Standards, Identifiers and Licensing).

Table of Contents

1 Introduction

1.1 Intended Audience and Use

In addition to defining and populating discipline-specific data standards, the field of interdisciplinary usage of data is one of the main topics in information technology and especially in (digital) humanities research. The present set of recommendations aims to advise and support Humanities institutions and research projects in establishing digital data collections and/or preparing their existing collections for discipline-specific and interdisciplinary usage, mainly in conjunction with the services and tools developed and offered by the DARIAH infrastructure.

After a brief overview of definitions and key concepts of interoperability, some thoughts are given on the approach of DARIAH to interoperability, together with a short description of four exemplary fields in which interoperability is of particular interest in the DARIAH context. These key aspects then serve as the basis for wider survey of practical use cases and the resulting recommendations. Although the focus lies on interdisciplinarity generally, it was inevitable to focus slightly more on those disciplines of which the authors of these recommendations have a deeper knowledge.

1.2 Key Concepts about Interoperability

Interoperability can be defined generally in this context 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" (NISO 2004) or in a more abstract way and concentrated on information as "the ability of two or more systems or components to exchange information and use the exchanged information without special effort on either system" (CC:DA 2000). In the Digital Humanities, this means that data and metadata from different contexts can be used and processed directly in every interoperable environment without any effort to reorganize or reformat the data.

"Interoperability" has to be distinguished from "interchange" (Unsworth 2011), where exchange of information is based on an intermediate process or format, with possible loss of information between the input and the output. Interoperability establishes a direct connection between two or more data sets. But, considering the various heterogeneities of data sets and their structures, successfully achieved interoperability is impeded in many ways on different levels. Therefore, establishing interoperability is a much more sophisticated process (Haslhofer and Neuhold 2011; Unsworth 2011).

As mentioned, interoperability touches different levels of exchange. The most basic of these is the technical or system level, dealing with interoperability of hardware, communication interfaces and software platforms. Focusing on interoperability of data sets and content, this report concentrates on the more sophisticated information-based levels. Haslhofer & Klas (2010) give a brief overview of several approaches to define different interoperability levels, of which the syntactic and semantic levels of interoperability are of particular interest in this context.

Interoperability on the syntactic level corresponds to questions of the formal structure of data sets. For example, the Extensible Markup Language (XML) and all encoding prin ciples based on XML provide schemas to ensure a correct and reliably persistent structure of information and markup. As a prerequisite, syntactic interoperability is neces sary to provide advanced access to the potential meaning of the data on the semantic level. Here, beyond the structural sphere, encoded data elements are enriched by further information to maintain communication and understanding of specified meaning. As for the semantic level, where structures are regulated in schemas, the encoding of different meanings also depends on regulations. Controlled vocabularies as used in many discip lines are one example for such a regulation.

1.3 Rationale

If interoperability is difficult, true interoperability across disciplines is perhaps even more so – particularly when talking about semantic interoperability –, as the wider the applica tion domain is, the lower are the chances of achieving some results. This is the case, for example, when using ontologies for this purpose, as shown by Marshall and Shipman (2003).

Therefore, given the number of domains and disciplines that DARIAH is trying to cater for, the solution of mapping the meaning of content in different collections onto the same ontology or conceptual model soon appeared not to be viable. As Bauman makes clear while discussing the topic of interoperability in relation to the goal and mission of TEI (Bauman 2011), the drawback of adhering to standards for the sake of interoperability is the consequent loss in terms of expressiveness.

Instead, DARIAH's position in this respect is to allow for crosswalks between different schemas: a sort of "crosswalk on demand". Infrastructure users will be able to use the Schema Registry – a tool which is being developed in DARIAH-DE – to create crosswalks between different metadata schemas so that they are made interoperable.

Our main goal was to devise a set of guidelines that is realistically applicable by partner institutions as part of their policies. Therefore, the first preliminary step was to gather and analyze information about the digital collections of the partners with regard to interoper ability. We identified the following key aspects to guide our analysis:

- APIs and Protocols: APIs and protocols are essential as they allow for workflows of data access and exchange not necessarily dependent on human agents. This idea is implied in the notion of "blind interchange" discussed by Bauman with the only difference being that, in our own vision, as little human intervention as possible should be required.
- Standards: using the same standard is in some, if not many cases, not enough in order to achieve true semantic, not just syntactic, interoperability. Therefore we discuss further aspects of standards in relation to interoperability, such as multiple serializations of the same scheme, and the problem of adoption and adaption of schemes to different contexts.

- **Identifiers**: two aspects of identifiers were considered: on the one hand, their persistence over time, which is a crucial aspect for any infrastructure project, and on the other hand the use of common, shared identifiers (e.g. controlled vocabu lary URIs) to express references to the same "things", that is one of the core ideas of Linked Data.
- **Licences**: licences, and specifically their machine-readability, play – perhaps not surprisingly – a crucial role within an interoperability scenario: not only should a licence be attached to any collection as soon as it is published online, but such a licence should also be readable and understandable, for example, to an auto mated agent harvesting that collection.

These four aspects define the context for the use cases that are described in the next section and also define the core aspects that will be covered in the recommendations.

2 APIs and Protocols

2.1 Overview

APIs and protocols are two milestones towards greater interoperability of electronic resource collections and, more generally, of systems that were independently developed. Let us start with some definitions.

In software engineering, an Application Programming Interface (API) describes the func tions (for procedural languages) or methods (for object-oriented languages) that are exposed by a given software, module or library. Such descriptions typically include:

- Information about the input parameters of the function/method, that is their name, number and type;
- A description of the operations performed by such a function/method, such as the algorithm it implements;
- Information about the output that is returned.

However, the term API is often used – particularly since the introduction of the Web 2.0 – to indicate Web APIs, that is a specific kind of APIs which uses the HTTP protocol for the exchange of API-related messages (i.e. requests and replies). In this section, and more generally in these guidelines, when we refer to APIs we tend to mean Web APIs mainly because it makes sense for us given the distributed nature of the collections we are dealing with, an issue that can be overcome by focusing on Web APIs.

A more than legitimate question that one might ask is "why do we need APIs?". To answer this, let us take as an example the implementation of a search functionality across several collections, that we call "generic search" for the sake of brevity. The way this typic ally works is by indexing the content of all the collection items: search terms are then matched against this index in order to retrieve the search results. To implement such a generic search, one needs to be able to index collections that may be stored in several locations – this is the case with DARIAH, where many different partner institutions provide their data – in a largely automated way. Being able to do so automatically is

essential for the generic search to be scalable (i.e. able to work with a large number of data or collections) and always up-to-date. Since some collections may change more frequently than others, they need to be harvested (i.e. gathered) and indexed periodically in order to be always up-to-date. Harvesting, that is the capability of accessing and fetching the content of a collection of resources without the need for (much) human intervention, is a key concept related to APIs and, more generally, to interoperability.

APIs allow not only harvesting (i.e. reading) data collections, but also modifying their content, that is creating, updating or deleting one or more items contained therein. The acronym CRUD – which stands for Create, Read, Update and Delete, is used to indicate this very set of operations that are typically made available by APIs and protocols.

2.2 Existing Approaches

In this section we will give a brief overview of some of the existing APIs and protocols that can be used in order to expose data in a machine-actionable way.

2.2.1 Atom Syndication Format

The Atom Syndication Format (from now onwards just Atom) is probably the most light weight and low-barrier approach to expose the contents of a collection of resources on the web. It was published in 2005 as an Internet Engineering Task Force (IETF) RFC standard (Nottingham and Sayre 2005). Among the advantages of using this format there is the wide variety of software, including web browsers, that support it natively.

The first use case for which Atom was employed was, as described by the RFC standard, "the syndication of Web content such as weblogs and news headlines to Web sites as well as directly to user agents". Quoting again from the format specifications:

> Atom is an XML-based document format that describes lists of related information known as "feeds". Feeds are composed of a number of items, known as "entries", each with an extensible set of attached metadata. For example, each entry has a title.

Another use case for which Atom has been used is the implementation of OAI-ORE, the Open Archives Initiative Object Reuse and Exchange format (Lagoze et al. 2008). Without going too much into the details of OAI-ORE, its main goal is better to represent Web resources, and particularly aggregated resources, that is resources consisting in turn of a set of resources. The OAI-ORE seeks to provide a machine-readable definition of a resource (e.g. document) and of the elements this is composed of, allowing one to make precise (i.e. granular) statements, for example, about the publication date, version or author of the whole resource as well as of any of its constituents. In other words, OAI-ORE is needed in order to overcome the difficulty, for a machine agent, correctly to understand what is a Web resource, and what are its components. This difficulty, however, does not affect human agents as they can easily tell, for example, what links contained in an article publish on the Web point to subsections of the same article and what links instead point to external resources.

Atom is also often used as an easy-to-consume format to package the reply of an API, such as for instance a list of results for a given query, or as a more machine-readable

serialization format in addition to plain HTML. For example OpenContext – an online open platform to publish archaeological research data – provides three serialization formats for each record in its collections: HTML, ArchaeoML and ATOM ("Item 97-L-19(549)" 2014; Kansa et al. 2010).

2.2.2 OAI-PMH

The "Open Archives Initiative Protocol for Metadata Harvesting" (from now on: OAI-PMH) is a protocol specified by the Open Archives Initiative. It consists of a specification to implement the RESTful API (see next section) and implementation guidelines (Open Archives Initiative 2014).

The purpose of OAI-PMH is to make repositories of data interoperable. A data provider is a repository that makes its metadata available via the OAI-PMH protocol. This data can be harvested by a service provider to create value-added services that allow new interac tion with the metadata previously harvested.

The current specification of OAI-PMH is from 2002 so it can be expected to be stable and well-known within the community. It uses standard technologies like XML and a RESTful API and mandates a minimum set of metadata that has to be exposed via this protocol, but other forms of metadata are allowed (quoting the specification):

> At a minimum, repositories *must* be able to return records with metadata expressed in the Dublin Core format, without any qualification. Optionally, a repository *may* also disseminate other formats of metadata. (Lagoze and Van de Sompel 2014)

Further, OAI-PMH allows for selective harvesting, i.e. the limitation of harvesters to harvest metadata from a repository to only harvest metadata that meets certain criteria.

For repositories that do not consist of large changing sets of data that would warrant an implementation of OAI-PMH for this repository, there is the possibility of using a static repository (Van de Sompel et al. 2004). A small repository (up to 5000 records) can make its metadata available through an XML document at a persistent URL. This URL can then be processed by an implementation of a static repository gateway, a piece of software that mediates OAI-PMH requests that it gets and answers them by using the static XML file that was provided by the repository. This way, small repositories can still expose their metadata via OAI-PMH without the need to implement it themselves. For an example, see the Use Case in section 2.3.

2.2.3 RESTful APIs

A representational state transfer application programming interface (from now on: RESTful API) is a web API that works via a well-known internet protocol: HTTP (Fielding et al. 1999). RESTful interfaces have emerged as a predominant architecture for web-oriented services. It mandates the use of the existing capabilities of HTTP to build an API for a web-oriented service. Resources are identified by their URI and typically consist of a representation of a resource in XML format, though strictly speaking, REST is resource format agnostic (as long as it has a supported MIME type (Freed and Borenstein 1996)). A RESTful web service must support the different HTTP methods (Wikipedia 2014), for example GET or POST to retrieve or create a resource, respectively.

REST is not a standardized protocol, but an architectural choice for a protocol or a web-service to base upon. As such it has found widespread adoption for services access ible over the internet. It reuses other technologies such as HTTP, XML or JSON (Crockford 2006) to facilitate communication between clients and servers.

2.2.4 Canonical Text Services Protocol

The Canonical Text Services protocol (CTS) is an interesting example of a solution specifically devised to tackle an issue that is typical of research in the Humanities, and particularly in Classics. Its main goal is to provide a protocol to translate between several common ways scholars in these fields use to refer to their primary sources, namely ancient texts.

One of the main characteristics of such "canonical citations" is that they allow scholars to cite, in a very precise way, a specific portion of a text without referring to a specific edition, but using instead a canonical citation scheme. This simple yet very interoperable solution allows them to express precise references to texts that everyone can look up in any particular edition of the cited texts.

For example, "Hom. *Il.* I 1-10" refers to the first ten lines of Homer's *Iliad*: The citation remains valid no matter if one is looking it up in a manuscript or a modern edition. The corresponding query to fetch this very passage from a CTS repository is:

```
http://hmt-cts.appspot.com/CTS?request=GetPassagePlus
&urn=urn:cts:greekLit:tlg0012.tlg001.msA:1.1-1.10
```

Let us see in detail how this query string is constructed:

- `http://hmt-cts.appspot.com/CTS` is the address of an existing CTS-compliant repository;
- `?request=GetPassagePlus` indicates the CTS method that is being invoked which in this case is "GetPassagePlus" and returns an XML-encoded response containing the requested text passage as TEI XML together with pointers to the preceding and following passages;
- `&urn=urn:cts:greekLit:tlg0012.tlg001.msA:1.1-1.10` this is the identi fier of the requested text passage expressed by means of a URN that follows a syntax defined within the CTS protocol ("Homer Multitext Project: Documentation" 2014).

CTS combines the FRBR - Functional Requirements for Bibliographic Records - data model together with URNs and a Web API to make a collection of TEI-encoded texts accessible by using the same citation schemes with which scholars in the field are already familiar.

2.2.5 HTTP and Linked Data

Linked Data (LD) also deserves to be mentioned in this section despite being neither, technically speaking, an API, nor a protocol in itself as it relies on the HTTP protocol. In a nutshell, LD is a way of publishing data on the Web which uses Semantic Web technologies to express the semantics of data and HTTP mainly as communication protocol. The main idea of LD is that "things" are identified by URIs and such URIs should be dereferenceable, meaning that by resolving an URI one should get back a representation of the thing that is referred to by that URI.

LD becomes a suitable approach for publishing data online particularly when dealing with decentralized sets of RDF data. This solution may be especially suitable when RDF is already the format of choice and when data are being published under an open license (because of the open and decentralized nature of LD and the HTTP protocol themselves).

A recent example of how to apply this approach in a real-world project is given by Pelagios, which began as a project and ended up being almost a consortium of institutions willing to share data about ancient world places. From a technical point of view, the pillars of Pelagios are:

1. The use of Pleiades URIs to unambiguously refer to geographical places;
2. The use of the Open Annotation Collaboration (OAC) ontology in order to express references to places that are found in the datasets provided by the partner institutions;
3. Decentralized storage of the annotations, meaning that rather than having a single data repository there is a single access point for the Pelagios datasets but each single dataset was stored and looked after by the contributing institution.

The RDF vocabulary used in Pelagios to describe the datasets is the Vocabulary of Interlinked Datasets (VoID) and aims at providing information such as where to find the dataset, who are the authors, which license applies to it, etc. The single access points to all Pelagios annotations can be found at http://pelagios.dme.ait.ac.at/api/datasets.ttl where each dataset contributed by partner institutions is listed together with basic metadata including the `void:dataDump` property which indicates where the annotation triples are to be found.

2.3 OAI-PMH Repository from Static File

As mentioned in the section about OAI-PMH (2.2.4) there is a possibility to provide an OAI-PMH interface to a repository without having to implement the protocol for the repository. For small repositories (fewer than 5000 records) one can use a static file to expose the metadata ("OAI-PMH Static Repository Gateway" 2014). This is called the static repository, which is provided by the repository at a persistent URL. This URL is

given to a web application, running on any server, which answers OAI-PMH requests by looking at the static repository file. This software is called the static repository gateway. There is a C implementation of an OAI-PMH static repository gateway called "`srepod`" (Van de Sompel and Lagoze 2014). It runs on UNIX-like (such as Linux) systems only. The installation procedure is not as straightforward as it could be, depending on the distribu tion. It contains an INSTALL file which lists instructions to install and configure the software within the Apache HTTP server ("The Apache HTTP Server Project" 2014) (special permissions such as root might be required). The main obstacle was the correct configur ation of Apache to use the srepod installation. For reference included here is an example configuration that worked on an Arch Linux installation (paths are subject to change on other systems):

```
<VirtualHost *:80>
    ServerAdmin root@localhost
    DocumentRoot "/srv/http/htdocs/"
    ErrorLog "/var/log/httpd/localhost-error_log"
    CustomLog "/var/log/httpd/localhost-access_log" common
    <Directory /srv/http/htdocs/>
    AddHandler cgi-script .cgi .pl
    Options ExecCGI FollowSymLinks MultiViews +Includes
    Order allow,deny
    allow from all
    </Directory>
    <Directory "/srv/http/cgi-bin/">
    AddHandler cgi-script .cgi .pl
    Options ExecCGI FollowSymLinks MultiViews +Includes
    Order allow,deny
    allow from all
    </Directory>
</VirtualHost>
```

Listing 1: Example static repository

Also included in the Appendix to this report is an excerpt from a real-world example of a static repository, kindly provided by Harald Lordick from the Steinheim Institut, Essen. The procedure is described below.

Static repository (Appendix A) → **Static repository gateway** (srepod instance) → **OAI-PMH API** (exposed by srepod) → **OAI-PMH request** (by someone who wants to query the metadata from the static repository, e.g. an OAI-PMH harvester)

The same principle is described in the following diagram ("OAI-PMH Static Repository Gateway" 2014):

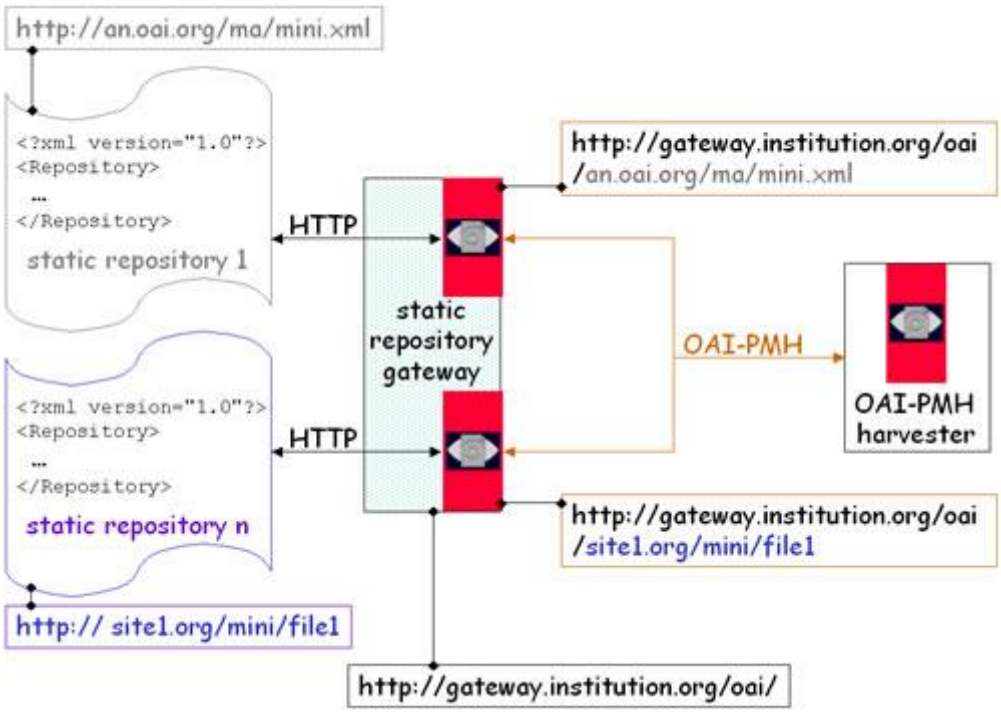

Figure 1: Architecture of the OAI Static Repository Gateway

The static repository gateway takes as input the static repository and answers OAI-PMH requests by accessing the XML file via HTTP. The static repositories have to be registered at the static repository gateway and can also be de-registered so that they are no longer available.

2.4 Recommendations

There is no one-size-fits-all API or protocol to be recommended (as there is not for many other kinds of problems), but rather a variety of possible approaches. Therefore, we strongly recommend that every data collection that is published online is provided with at least one machine interface that allow agents, either humans or machine agents, to fetch and/or to manipulate its content.

When only one API is provided, we recommend this to be compliant with the OAI-PMH protocol. Among the reasons for doing so, there is the existence of a number of open source implementations and client libraries together with the wide adoption of this standard by institutional repositories across disciplines.

If a collection is accessible via other APIs in addition to OAI-PMH, RESTful interfaces are certainly a robust approach as they allow, among other things, multiple serializations, separation between data and presentation of data and thus the transformation of data collections from mere black boxes into more reusable data sources.

3 Standards

3.1 Three Dimensions of Standards for Interoperability

To approach the topic of standards from an interdisciplinary perspective, the different levels and areas of interoperability through standards in an interdisciplinary context have to be clarified.

Scheme. Depending on the material and the research setting in which the material is modeled and described, researchers could choose to integrate their data into a scheme for publication which is widely accepted across domain borders instead of domain dependent or proprietary schemes.

Serialization. Although it is somehow common knowledge to use XML as an inter change format for the exposure of data in the humanities, the situation is not always that simple. Possible serializations can influence the decision for a standard aiming at interdis ciplinary interoperability. The description of serializations of ORE in ATOM, RDF/XML, RDFa, pure HTTP and the attempts to express TEI in RDF and OWL (for the benefits of standoff markup) show that serialization is of major concern when dealing with interdis ciplinary interoperability. This is especially true when limited project resources do not allow for crosswalks at the end of a project for interoperable data exposure.

Adoption. Every schema can be adopted in undefined ways depending on the under standing of the schema, the project background, project principles and so on. The semantics of TEI elements, for example, can be interpreted in a number of ways, thus leading sometimes to multiple equivalent ways of annotating the same phenomenon within different texts. This is also the general experience with Dublin Core, a fact which is its strength as well as its weakness. The awareness of how a scheme is generally used and of the decisions behind its application in a project is essential when thinking of interdis ciplinary interoperability. Consequences in the application of a schema may be to orient oneself to common practice, to interpret the schema in a very general way, or to write down the principles of one's project adoption as metadata for future data integration.

Interoperability is a machine-based concept and has to be distinguished from human-based ideas like being understandable or interpretable. Haslhofer and Neuhold, for example, call a resource interoperable, "if an application is aware of the structure and semantics of data [and if] it can process them correctly" (Haslhofer and Neuhold 2011, 1). The necessary paradox distinction is implicitly present in the word "correctly" because, from a machine point of view, a resource could be processed correctly, meaning in a formally correct way, but a researcher would not call the result correct from a discourse point of view. This could lead to a paradox because formality needs unambiguity and discourse – especially discourse between different interdisciplinary communities – works through balanced contingency. Unambiguity and contingency cannot be served at the same time. This situation needs wise decision-making by the person modeling the data or providing a resource with metadata.

Interdisciplinary interoperability for standards has two perspectives, related to the standard being used (a) to model data and (b) to create metadata. Having summarized

the different areas of tension one may face when dealing with interdisciplinary interoperability, the separation between metadata and data offers the possibility to apply different strategies. To insure expressivity of the semantics of the data, a specific standard or application of a standard related to the specific situation of the research context can be applied while, at the same time, a general metadata scheme to describe the resource can be chosen. Moreover, the metadata description could and should be used to give the necessary information to increase the interoperability of the primary data. With the shift from databases to dataspaces (Franklin, Halevy, and Maier 2005), the task of data integration is widely accepted as a step of the data processing task relieving the person producing the data from the burden of treating interoperability as the most important criterion. In any case, the metadata for a resource must be sufficient to make data integration possible. Difficulties may also arise from the fact that metadata and data are not always clearly separable. For a linguist, TEI Markup is metadata. For a philologist, it is not. This short listing of perspectives makes it clear that an evaluation of standards for interdisciplinary use cannot focus only on the selection of particular standards but must also handle the question of how to deal with and implement standards.

There are many high-level standards which are specifically designed for domain-independent interoperability, like CIDOC-CRM for semantics, OAI-ORE for structures, EDM, METS/(MODS) for cultural heritage, Dublin Core for digital resources of any kind, DDI for Data etc. Most of these standards, as well as many domain-specific standards, permit scalability. This approach allows the user to adapt a standard down to the project level without losing interoperability on a higher level. For example, ORE lets you define and use more concrete or several predicates to explain what `ore:aggregates` means in your context. Nevertheless, interoperability is maintained because these predicates remain subclasses of the class `ore:aggregates` and are therefore aligned automatically to other ORE implementations like `ens:hasView` in EDM. Dublin Core provides a fixed list of refinements for its core elements. MODS differentiates between required and optional elements within an element group. The concept of scalability is a very important one in the context of interdisciplinary interoperability, allowing both to achieve the level of precision needed in a particular research context and to transcend this level and its resulting heterogeneities in an interdisciplinary perspective. But there are also limitations, since scalability is only possible among hierarchical semantics. The unspoken assumption behind hierarchical semantics is that research projects mainly vary in the level of concreteness and so metadata descriptions can be integrated by abstract classes. Of course, research also means disagreement which can result in an opposing view on the use and the semantic of the class structure itself. On the other hand, class structures are only one model of knowledge representation and the effectiveness of scalability declines when one begins to use multi-dimensional knowledge representations.

3.2 Existing Approaches in Interdisciplinary Interoperability

3.2.1 High-Level Metadata Schemas

Dublin Core, the Minimal Consensus

Dublin Core is a metadata standard that was developed in 1994, when the emerging web experienced problems in the findability of web resources. A set of general statements for the classification of these web resources was seen as extremely useful (Dublin Core Metadata Initiative 2014a). Since the approach was a generic one, the fifteen core elements defined by Dublin Core have established themselves as the most widely used metadata description on the Web. Although created for the description of web resources like video, images, and web pages, it is also widely used for the description of physical resources on the web by libraries, museums, and archives. It was endorsed in the standards RFC 5013, ISO 15836-2009 and Z39.85. The implementation of an OAI-PMH interface defines Dublin Core as a mandatory standard for data exposure and, in the Linked Open Data Community, Dublin Core is widely used to achieve semantic interoperability between different datasets. Presenting data in this way using Dublin Core is the essential recommendation in choosing a high-level metadata scheme to assure interoperability.

As mentioned, the core of Dublin Core is a set of 15 metadata elements called Simple Dublin Core: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights. For further information or special needs (which were classified as additional requirements), it is possible to add additional elements through specification of existing elements. Any output that uses these refinements is classified as Qualified Dublin Core. The distinction between Simple and Qualified Dublin Core references an issue which was raised in section 3.1 above, that of abstraction vs. expressivity. For example, qualifiers permit the use of structured or compound values as values of a metadata element. The use of a more expressive yet more complex scheme could lead to technical interoperability problems because a software which is to consume the metadata has to be capable of processing its complexity. On a semantic level, the issue is reflected in the so called "dumb-down" principle. This principle recommends that by using a refinement for an element, this element should also be correct and comprehensive without the refinement. For example, when using the Alternative Title refinement of Title, there still has to be a Title statement and this Title must contain the expression by which the resource is mostly denominated.

By introducing another issue from the general introduction in section 3.1, the discussion becomes even more complex: the topic of the application of a metadata scheme, in this case Dublin Core. While trying to keep the definition of Dublin Core and any other scheme with a generic approach simple and open to assure its adaptability in a huge variety of contexts, this situation often leads to inconsistencies when applied. To reduce interoperability problems of this kind, principles or best practices are documented, if not by the provider, then by the community using the scheme. Dublin Core has two other principles:

- The "one-to-one" principle stating that the metadata description should belong to the resource to which it is attached and not to the object which might be repres ented in the resource, for example a jpg image of a painting.
- The "appropriate values" principle declares that a value for a metadata element should allow a metadata description to be read by machines and humans and should at least on a minimal level be understandable by both. Apart from that it depends on the context of the Dublin Core implementation what values may be appropriate.

The topic of the appropriate value raises the topic of authority files and controlled vocab ularies, that is, the use of a standardized language for values. Some very general vocabularies are listed in section 4. There are also specific schemes to encode formats, for example, a date statement. For interdisciplinary interoperability, the use of these vocabularies is recommended although the principles mentioned above should be reflected. Encoding schemes and controlled vocabularies improve metadata consistency among different resources and give the reader of metadata the opportunity to look up and better understand the meaning of data. To ensure this, every element which uses a specific vocabulary or encoding scheme should name this vocabulary or scheme. For this reason, Dublin Core has a qualifier for each element called scheme where the name or code of the vocabulary or scheme can be given. On the other hand, it is important to keep in mind that choosing a controlled vocabulary may reduce expressivity. The use of a generic controlled vocabulary instead of a domain specific one also makes it easier to find information in an interdisciplinary environment but reduces the quality of the metadata for people within that domain.

In a metadata description using Dublin Core, each element may appear many times, but the sequence of elements has no explicit meaning. Of course, the implementation of Dublin Core in a specific project may consider a project-related meaning for the sequence but this meaning cannot be transported through the definitions of Dublin Core into other environments.

As mentioned in different contexts, the disadvantage of Dublin Core is its abstraction level. The Dublin Core guidelines state that Dublin Core is "easily grasped, but not neces sarily up to the task of expressing complex relationships or concepts" (Dublin Core Metadata Initiative 2014b). By deciding for a high-level scheme to describe metadata across disciplines, one has to decide what information should be transported into which situation. There are also other generic schemes like the Metadata Object Description Scheme which may serve better in particular cases.

CIDOC-CRM and Semantic Interoperability

The refinement strategy of Dublin Core is a way to limit the drawbacks of an approach which tries to define a minimal consensual metadata scheme above any other data and data structures. It comes from the top and tries to find its way down. The CIDOC-CRM addresses the problem from the opposite perspective. By emphasizing that the useful ness of such metadata is a question of scalability (Doerr 2003, 77) it chooses a bottom-up

approach. Scalability in this context means that any more complex queries where more precise results are expected will not lead to these expected results because the simplifica tion within Dublin Core eliminated the necessary information layer. The interoperability strategy of CIDOC-CRM therefore preserves the complexity of the source data "from pure terminology" (Doerr 2003, 76) and defines a way of describing the semantics which are inherent to this terminology. This is done by a conceptual model which is seen as a ground for any schema semantics. In short: while Dublin Core defines a minimal set of semantic terms into which data or metadata is wriggled into, CIDOC-CRM defines an ontological layer (on the basis of some philosophical commitments) to which semantics of a specific domain or a project scheme can be aligned directly from the data level. As there is no simplified or consensual scheme modeled of the source data schemes – like in scheme crosswalks for example – the actual relations within the data are more likely inferred through the CIDOC-CRM conceptual model at query time. CIDOC-CRM calls this "read-only-integration", as the data integration only exists in the result of the query while the data stays in its original state before and afterwards. So data integration becomes an exploratory process respecting the existing differences of source data and facilitating interoperability up to the level where it is possible without repressing these differences.

As mentioned before, CIDOC-CRM is not semantically neutral, although it builds an abstraction level which tries to be as agnostic to semantic decisions as possible. The semantic commitments made by CIDOC-CRM are therefore theoretical and are related to the Cultural Heritage field where it developed. One should therefore consider the commitments before choosing CIDOC-CRM and see if they work in a particular situation. The Cultural Heritage field consists primarily of objects, for example paintings, pottery, books, and so on. But what could be said about these objects is extremely contingent. Objects move around, change over time, and are interpreted in different ways. Around objects there is, therefore, a vast space of information heterogeneity. The first commit ment reflecting this situation which was already transparent in the introduction to CIDOC-CRM above is that this heterogeneity is meaningful and therefore has to be preserved. Following this principle, CIDOC-CRM is an instrument to organize this hetero geneity in an interoperable semantic space. This even allows contradictory pieces of information between different data sources to be modeled. CIDOC-CRM, in this sense, has a different interoperability approach: it does not control semantics for interoperability but derives semantics to create interoperability. This leads to a necessary decision to be made for a project as to whether consistency and semantic control are the goals of implementing an interoperability layer or whether no more interoperability is needed than to have a finding aid (for example by Dublin Core Metadata) or whether the specific situation of the data needs a strategy like CIDOC-CRM to remain meaningful even in an interdisciplinary environment.

The second commitment of CIDOC-CRM is that it defines an object as heterogeneous because it flows through time and space. Therefore, in CIDOC-CRM an information unit is modeled generally as an event which took place somewhere. It is an interesting question whether discourse-oriented approaches to semantic heterogeneity would need other formalizations. In any case, it is important to know that, in contrast to Dublin Core,

CIDOC-CRM has a semantic inner logic one has to consider before applying the scheme. Despite the commitments of CIDOC-CRM, it is in the end a scheme which reflects the situation of humanities research very well, considering that what is true for objects is even more true for symbolic entities or other humanities research objects. Also, the time/space assertion for heterogeneity reflects very well the important role of the source in humanities research. CIDOC-CRM therefore offers an appropriate approach to achieving interoperability for the specificity of humanities research data.

Apart from the aforementioned point that the level of interoperability achieved is not easily predictable, there are also other pragmatic issues to consider before choosing CIDOC-CRM. What should be clear up to this point is that it is not as easily applicable as other interoperability strategies. Many things must be considered in advance and the learning curve is high as the application is work intensive. Because of this, although it is a widely accepted standard, it is not applied to the same extent as other interoperability approaches. Hence one must consider, first whether time resources are sufficient, and second whether the expected audience may benefit from the use of CIDOC-CRM.

3.2.2 Schema Serialization and Data Publication

One level of interdisciplinary interoperability does not so much refer to the semantic dimension of the scheme as to the syntactic level into which the scheme is serialized. This place should not be used to repeat the common experience that even in 2001, XML was seen "as a prevailing data model in humanities computing" (Sahle 2002). XML has the biggest infrastructural support in the Digital Humanities and there are a variety of adja cent techniques to interact, process, and manipulate it. Apart from a few exceptions, one could hardly find a scheme in the humanities that is not a serialization model for XML or where XML is not the main focus. Besides this wide acceptance, another advantage of XML is that it is easily readable for human beings as well as for machines and the rules are limited and simple. Furthermore, there are abundant resources which give recom mendations about using XML.

Hence this chapter would like to move the attention to an approach which has become more and more important over the last several years and which implements the core aspects of cross-domain interoperability. This approach is called Linked Open Data, which is not so much a type of data serialization as it is an idea and a set of best practices to facilitate this idea. The idea is that data should be published in the same way documents are published on the web. This should make data publication and consumption as successful, useful and easy as the publication of websites, leading into an infinite data space of uniquely identifiable resources, interlinked with and describing each other and processed in a common way. The identification and linking are done through URIs, just as for web documents. The processing is also done through HTTP and the description process uses RDF/XML. So, apart from infrastructural measures to be taken, the serializa tion of the data in RDF/XML is a single requirement. RDF/XML can be seen as additional rules for XML where the graph model of RDF – which is supposed to be more expressive than the hierarchical model of pure XML, where overlapping structures are prohibited – is implemented in XML. Loosely speaking, RDF defines a way to model subject, predicate,

and object sentences talking about entities which, when combined, generate a graph of connected assertions.

The general lesson for interdisciplinary interoperability – apart from the recommendation to publish data as linked data – is that, by serializing schemes in a specific data model, one refers to a specific infrastructure, which is used in specific environments, and works with specific schemes. The decision of the serialization of a scheme, therefore, has to reflect the purposes and the audience for which the data is published. Nevertheless, the serialization of the data in XML is almost mandatory and rather concerns offering more than one serialization.

3.2.3 Interpretability of Schemas, the Audience of Data and Scalability

In the introduction to this chapter, the approach was introduced that scalability is a concept that can be used productively to reflect on interoperability and create interoper ability. There are three overlapping perspectives on interoperability from a scalability point of view. First, the more precise a piece of information is, the more one can expect that other people or systems interpret this information differently or would use another value for it. Inversely, the more abstract or integrative a piece of information is, the less useful it can be. This was CIDOC-CRM's critique of Dublin Core. On the one hand, the information value is high, as is the risk of interoperability issues; on the other hand, inter operability is achieved but the meaning of the information may be low. Both a high information value and interoperability are more easily achieved in situations of stable language use which may exist in some contexts. As language is, by definition, contingent, stable language use is a phenomenon of consensual thinking and practices for specific things can be achieved by regulation, as with authority files. Of course, these attempts are always undermined by disagreement or by the lack of definition leading to one side opposed to the other. These three axes build the matrix in which a space for adapting a scheme is created and where a decision has to be made. Referred to the three axes, this decision decides on the granularity with which a scheme is applied, reflecting the creators' knowledge of the usage of the scheme and the audience they aim at. Knowing and informing about the best practices by using a scheme is the first step. Knowing and guessing at the discursive and semantic situation that exists for the audience, the systems used by the audience and how this relates to the best practices is the next. The TEI ecosystem reflects this situation of interoperability by being a matrix and not a reachable goal.

The Text Encoding Initiative Standard could be described as the most supported Data Scheme in the Humanities. Of interest for the present topic is to see that it is also one of the most expressive, with more than 500 elements. This expressivity often leads to critique from the perspective of interoperability, based on the observation that "Even with XML-formatted TEI, a scheme that theoretically should lend itself to interoperability, not all texts are created equal" (Pytlik Zillig 2009, 188). In fact, one can often use different elements to describe the same textual phenomena. On the other hand, the number of elements really make it hard for people to obtain an overview of the TEI, which is needed for a consistent use of the model. In response to this, the TEI consortium defined a signi

ficantly smaller model of some hundred elements which fits "90% of the needs of 90% of the TEI user community" ("TEI: TEI Lite" 2014). Since TEI-Lite was defined on the basis of experiences from projects which used TEI, it should be capable of handling a variety of text-types and of producing reasonable results for them by being as small and simple as it is consistent with the other goals. At the same time, it is still compliant with the TEI-Full definitions. The definition of TEI-Lite, therefore, reflects precisely the situation we described above. Considering the decisions that were made to define TEI-Lite can there fore be a good aid in approaching the interoperability task within one's own project, as TEI-Lite is a good recommendation for the use of TEI in an interdisciplinary environment.

Beside TEI-Lite, there are also other projects trying to deal with the contingency of TEI, both from within and outside of the TEI. The EpiDoc collaborative ("TEI: EpiDoc" 2014), for example, defines a specific set of rules and elements for the tagging of papyrological documents and epigraphical objects, whereas the CEI could be seen as dialect of TEI for the tagging of medieval charters. These are community-specific approaches for building environments stable enough for fostering the definition of more precise applications of a standard that deliver more interoperability. From an interdisciplinary perspective, projects should be aware of such community efforts and should define which communities belong to the audience for which the data could be useful. If these efforts are not form alized, as in the case of CEI or EpiDoc, it might be possible to identify common practices in such communities so that interdisciplinary interoperability could be reached while still producing meaningful data.

The last example from the TEI ecosystem is TEI-ANALYTICS (TEI-A), which is a standard defined not so much for the use of encoders but for the use of linguistic data analysts. The goal is to "work with extremely large literary text corpora of the sort that would allow computational study of literary and linguistic variance across boundaries of genre and geography and over long periods of time" (Zillig: 2009, 188). While consistency is needed to fulfill this task, the reality is an extremely heterogeneous landscape of encoded texts. TEI-A, as a subset of TEI elements for linguistic purposes, is automatically transformed from TEI source data into TEI-A. The source remains untouched while a new TEI docu ment is produced which is compliant to TEI-A and can therefore be gathered together with other TEI-A documents to build a corpus.

The purpose of these recommendations is not to evaluate whether the outcome of such an automatic abstraction can be fair to the source data or whether meaningful insight can be drawn from such an abstraction level; it is, however, a good example since inter disciplinary interoperability can be tackled not only from the data creation side but also from the data integration side. When trying to decide on a strategy for interdisciplinary interoperability in a project, one should know about existing tools and strategies for data integration for the audience which one expects to address. So, by considering computer linguists as possible consumers for a project's data, knowing that TEI-A exists relieves one – at least partly – from the task of considering their needs when creating the data.

3.3 Use Case: OAI to RDF

The goal of the use case is to download and install the OAI2RDF script on a desktop computer locally and run it on the OAI-PMH interface from the BBAW (Berlin-Brandenburgische Akademie der Wissenschaften). Afterwards the RDF output should be enriched exemplarily and published as Linked Open Data on a server.

3.3.1 Walkthrough

1. The source of OAI2RDF is downloadable from an SVN server by using an SVN client. On Linux machines, an SVN client is most likely preinstalled. The URL for the code repository is http://simile.mit.edu/repository/RDFizers/oai2rdf/;
2. For installation, the RDFizer requires the following things:
   - The Java Development Kit (note that the Runtime Environment is not sufficient) in version 1.4 or higher. The command 'java -version' on the shell provides information about your Java version;
   - The building tool Apache Maven in version 2.0 or higher. The command 'mvn -version' on the shell provides information about your Java version;
   - An internet connection for the download of required libraries during the installation process;
   - The environment variable 'JAVA_HOME' must be set explicitly to the path of the java installation being used. On Linux machines, the installation path normally begins with /usr/lib/jvm/;
3. After the preconditions are fulfilled, 'mvn package' within the downloaded folder will start the building and installation process;
4. The script will be put to work by using the following scheme: `oai2rdf options URL output_folder`. So to start the grabbing of the metadata from the OAI-PMH interface, we need the base URL of the interface as a parameter. In the present use case this is http://edoc.bbaw.de/oai2/oai2.php. The output folder is created automatically, so it is not necessary to create it first. Because every OAI-PMH interface can set its default metadata schema by its own and the script only handles OAI_DC and MODS it is recommendable to use the -m option together with the parameter oai_dc which means that the script will explicitly request the OAI_DC metadata schema. All together the command line should look like this: `./oai2rdf.sh -m oai_dc http://edoc.bbaw.de/oai2/oai2.php edocBBAW/`;
5. The script now starts to download and transform the metadata of the OAI-PMH interface into the folder chosen before. By doing this it creates a new folder structure within this folder.
6. To link the produced RDF data with other Linked Open Data we take one of the RDF Files as an example and open it with the text editor of choice – it is important to use an editor which saves as plain text;

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:ow="http://www.ontoweb.org/ontology/1#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
      <ow:Publication rdf:about="oai:kobv.de-opus-bbaw:905">
            <dc:title>Eine Analyse des Kontextes wäre
                      hilfreich</dc:title>
            <dc:creator>Riedmüller, Barbara</dc:creator>
            <dc:subject>Wissenschaftsfreiheit</dc:subject>
            <dc:subject>Akademische Freiheit</dc:subject>
            <dc:subject>Forschungsfreiheit</dc:subject>
            <dc:subject>Genforschung</dc:subject>
            <dc:subject>General serials and their indexes</dc:subject>
            <dc:publisher>Ber1in-Brandenburgische Akademie der
                         Wissenschaften</dc:publisher>
            <dc:publisher>BBAW. Interdisziplinäre Arbeitsgruppe
                         Gegenworte - Hefte für den Disput über Wissen
            </dc:publisher>
            <dc:date>1998</dc:date>
            <dc:type>Article</dc:type>
            <dc:format>application/pdf</dc:format>
      </ow:Publication>
</rdf:RDF>
```

Listing 2: Dublin Core

7. Because the scenario is to link to the GND subject catalogue we pick one of the subjects (dc:subject) like 'Wissenschaftsfreiheit' and look it up in the online cata logue of the DNB (unfortunately the DNB does not provide a SPARQL endpoint to search automatically through a script). There is an URI in the information table which identifies the subject and provide some additional information by referring to it. The URI should be copied and entered as value of a new to write 'rdf:resource' attribute in the dc:subject element instead of the text node saying 'Wissenschaftsfreiheit'. The new Linked Data compliant statement should look like this:

```
<dc:subject rdf:resource="http://d-nb.info/gnd/4121933-8"/>
<dc:subject>Akademische Freiheit</dc:subject>
<dc:subject>Forschungsfreiheit</dc:subject>
<dc:subject>Genforschung</dc:subject>
```

Listing 3: Dublin Core Subject Description

8. This process can be repeated for other subjects as well as for persons registered in the GND. Of course doing this manually is a big effort. There are also tools like SILK and LIMES to automatically find reasonable links or one could write little scripts but this is not the issue of this scenario which mainly focuses on the OAI2RDF script and the demonstration of a possibility where to go with it.
9. To publish the edited file as Linked Open Data it is also required to change the value of rdf:about in each publication described into dereferenceable URIs and store the file under an URL which complies with the URI scheme chosen on a

server (for more details on Linked Open Data principles, see Heath & Bizer 2011). After that links to the file or entities within the file from outside of it should be generated to become part of the Linked Open Data cloud.

3.3.2 What did work

Although multiple steps were needed, the building and installation was without any errors. The main script is customizable for different purposes and schemas. It is also extendible by adding own 'transformers' – transformation scripts for specific schemas – which makes it very usable. The RDF output seems consistent and namespaces are auto matically added to the file.

3.3.3 What did not work

Unfortunately, the system of the folder structure and the different files into which the metadata are separated is not transparent for the user and is also not explained on the homepage of OAI2RDF. The effort which is produced by this separation and complex hierarchy is unnecessary in this situation.

The complexity of the software and the preconditions seem exaggerated considering that it is just doing a simple transformation. Users on an entry level might be scared by this. On the other hand, users with a certain degree of technical knowledge would maybe write a piece of code adapted to their particular context on their own.

3.4 Marc21 XML to SKOS/RDF

This use case describes the process of transforming a thesaurus encoded in Marc21 into a SKOS thesaurus in a way that does not involve (much) human interaction. The workflow relies upon an OAI-PMH interface, the Stellar Console and an AllegroGraph triple store where the SKOS/RDF thesaurus is stored. This use case shows how the task of trans forming legacy data into a more semantic format becomes easier when standard APIs to access the data and open source tools to manipulated it are available.

3.4.1 Walkthrough

The data used here come from Zenon, the OPAC of the German Archaeological Institute, and specifically from its thesaurus. This thesaurus is stored as Marc21 XML and can be fetched automatically as it is made available via an open OAI-PMH interface. The thesaurus is an essential tool for browsing the content of the library catalogue: each entry is assigned one or more subject terms that are drawn from the thesaurus. The image below shows the thesaurus visualized as bibliography tree: Zenon users, and prob ably many archaeologists, find this and similar Information Retrieval tools of extreme importance for their daily work.

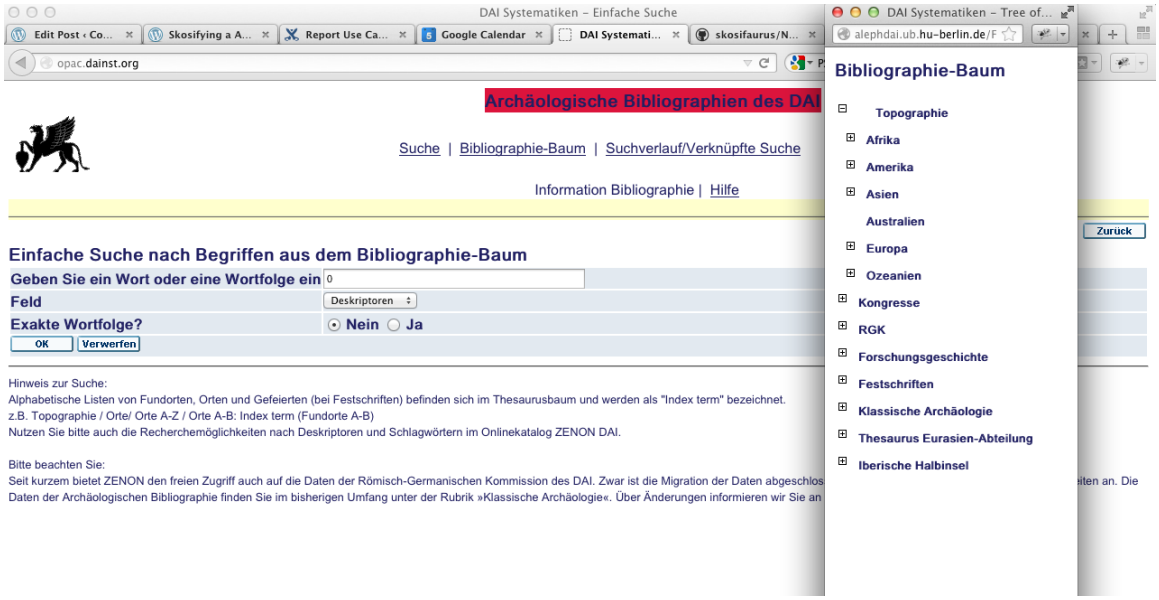

Figure 2: Zenon – the OPAC from the German Archaeological Institute

The main piece of software that was used to produce the SKOS/RDF result is the Stellar Console, a freely available and open source tool (Binding 2014) developed by Ceri Binding and Doug Tudhope in the framework of the AHRC-funded project "Semantic Technologies Enhancing Links and Linked Data for Archaeological Resources" (STELLAR). What the StellarConsole does is produce a more structured and semantic output, such as SKOS/RDF or CIDOC-CRM/RDF, by applying a set of (customizable) templates to the CSV file received as input.

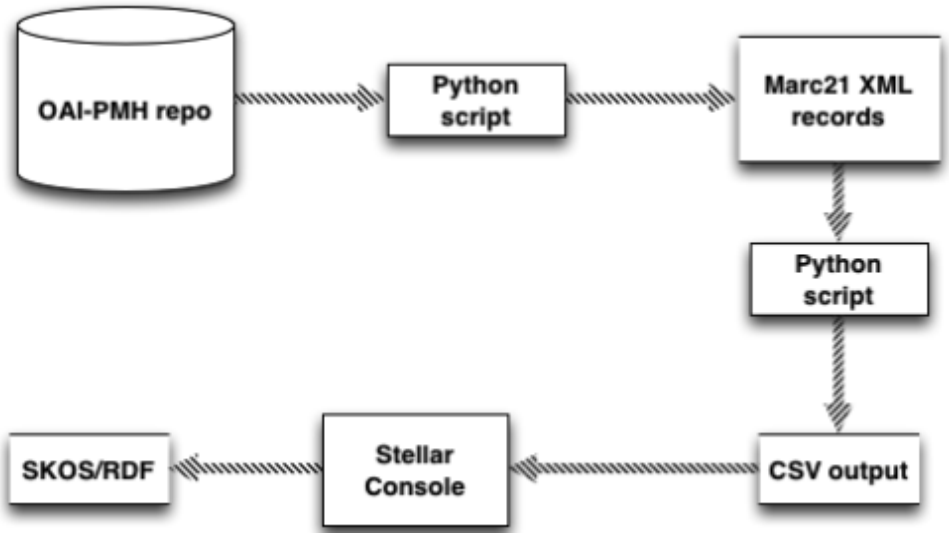

Figure 3: From Marc21 to SKOS/RDF

All that remained to do at this point was to write a short script – approximately a hundred lines of Python – in order (a) to harvest the OAI-PMH repository and fetch the ~80k records of the thesaurus and (b) to produce a CSV output to be fed into the Stellar Console (Romanello 2014).

3.4.2 What did work

The SKOS/RDF thesaurus was successfully produced by running through the StellarCon sole a CSV file that was created out of the harvested Marc21 XML records. The resulting RDF triples – slightly less than a million in total – were loaded onto an instance of the Allegro Graph triple store: figure 4 below shows how the thesaurus data is visualized by using the outline view of Gruff, a client for the Allegro Graph store.

3.4.3 What did not work

We experienced only one problem related to the text-encoding format. To understand the problem, it is important to mention that the Python script was run on a Mac OS plat form whereas the StellarConsole on a Windows one, as currently it works only on such a platform.

The problem resulted precisely from the way the CSV file was read by the Stellar Console. In the first version of the script, the lines that write the CSV file to memory looked like this:

```
file = codecs.open("thesaurus_temp.csv","w","utf-8")
file.write("\n".join(output))
```

This works in most cases. But if the file that one is writing is to be processed in a Windows environment – for whatever reason one may want (or have) to do so – one should use the following code instead, just to be on the safe side:

```
file = codecs.open("thesaurus_temp.csv","w","utf-8-sig")
file.write("\n".join(output))
```

The reason is that Microsoft uses a special sequence of bytes, a sort of Byte Order Mark (BOM) that is prepended to an UTF-8 encoded file, to let the software understand in which format the file is encoded. Without that character sequence, the StellarConsole, as well as other software such as MS Excel, will not be able to read correctly the file, thus resulting in the encoding of the SKOS/RDF output being corrupted.

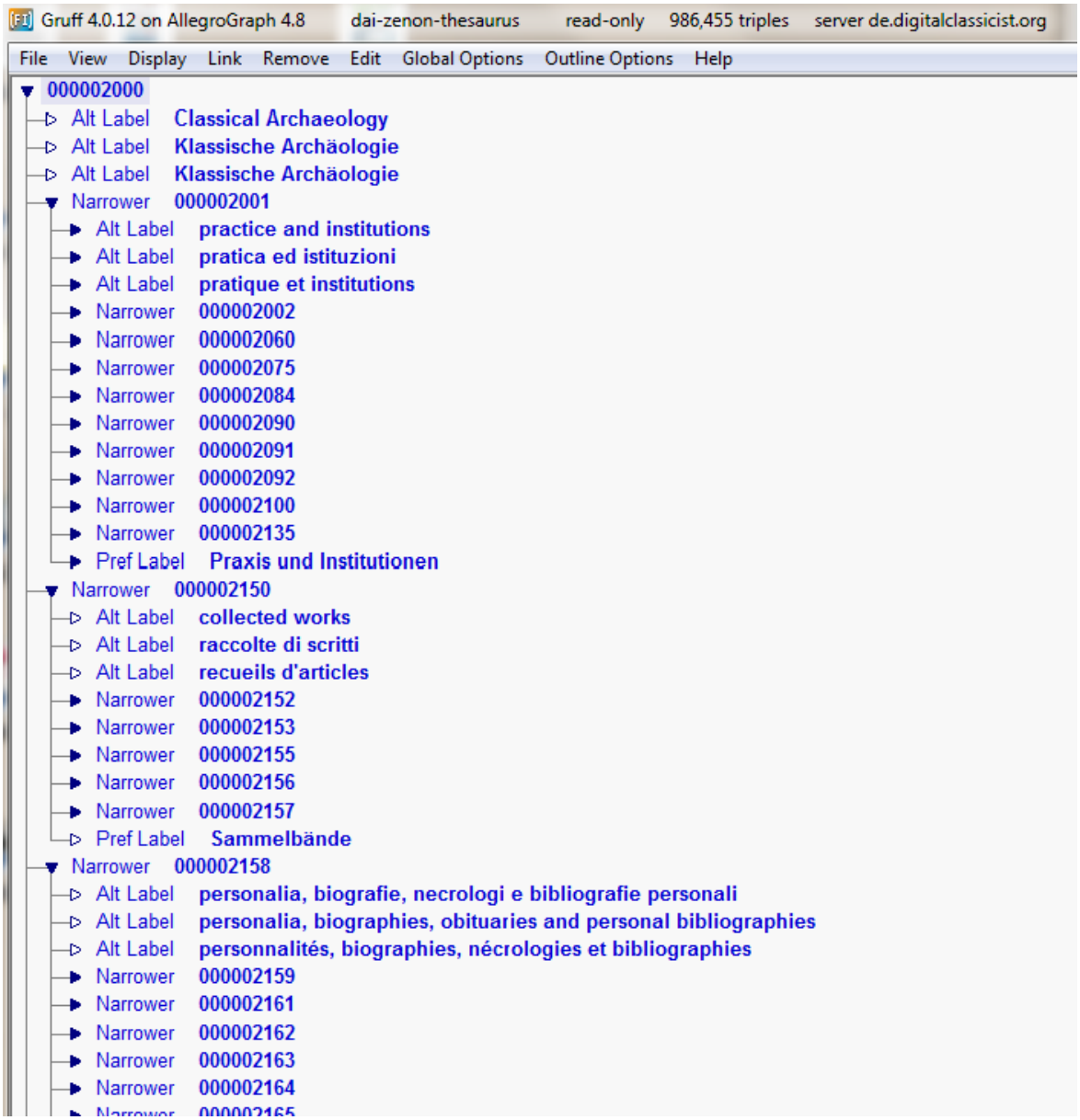

Figure 4: Zenon Thesaurus

3.5 Recommendations

General points:

- Interdisciplinary Interoperability has at least 3 modeling dimensions one should think of: the scheme, the serialization of the scheme and the adoption of the scheme;
- Interoperability often has a conflicting relation with the expressivity of the data, especially in an interdisciplinary perspective;
- Therefore, the issue is not about being most interoperable for any reason while minimizing the value of the data. It is about finding the right place in a matrix of aspects;
- For the data scheme, this right place may lie in between finding the right granu larity with which a scheme is applied reflecting the creators' knowledge of the usage of the scheme and the audience (and its discursive practices) they aim at;

- Assign metadata to your data. When your data is particularly special and lacks interoperability, metadata still offer the possibility to find and integrate your data;
- There are nevertheless approaches to reach interoperability from the data or the metadata level (see comparison of Dublin Core and CIDOC-CRM);
- Choose a scalable approach to interoperability where possible (see DC Refinement example or Linked Open Data for explanation);
- Be aware that the scalability approach only works where semantic differences are a question of hierarchy;
- More often interoperability is also handled through data integration ("from data bases to dataspaces") which leverages the data producer to take responsibility for the issue of interoperability. Keep that in mind when things get too complicated.

Scheme:

- Make use of attributes describing the background (authority file, encoding of the value for a data or metadata field when possible;
- Dublin Core is the most widely used metadata scheme for digital resources. Everyone understands it, so use it as the minimal level to describe your data;
- Dublin Core provide qualifiers and refinements for elements reflecting the scalab ility perspective to interoperability. Design your description in a way that it could be processed/understood without them ("dumb-down" principle);
- Dublin Core approaches interoperability through simplicity and abstraction: this can lead to inconsistent situations when putting together data from different sources;
- Design the Dublin Core Metadata in a way that it could be adequately understood by humans and machines ("one-to-one" principle);
- There are alternatives to Dublin Core like MODS, consider them in case DC does not reflect your needs;
- Choose a bottom-up description when you fear that substantial meaning is lost if you designed your data in an interdisciplinary interoperability perspective. A common way to do so in the humanities is CIDOC-CRM;
- CIDOC-CRM leaves your data as you want; it achieves interoperability at query time (read-only-integration);
- CIDOC-CRM is a good interoperability approach for heterogeneous data;
- While CIDOC-CRM may seem the perfect approach, the effort to implement it is big and the data with which it should be integrated needs also a CIDOC-CRM description.

Serialization:

- XML is the most widely used serialization of data in the humanities. Use it as the first way to expose your data;

- The web of data is a growing space for data designed for community and discip line agnostic data share. Consider to publish your data as Linked Open Data for outreach.

Adoption:

- Often a scheme can be applied in many ways. Gather information about the general use, existing best practices and points of discussion for the scheme you choose;
- If the scheme supports this (for example in the TEI Header), document the decisions you made when you applied the scheme, so anyone may get an under standing;
- Often there are subsets or defined best practices for schemes (like TEI-Lite, TEI-A). Look out for such approaches as they often are sufficient;
- There are also subsets or extensions for schemes (like CEI) which extend schemes while preserving interoperability. Look out for such approaches if a widely used scheme does not fulfill your need totally.

Aspects which also should influence your decisions:

- Look at your projects resources and choose a strategy which is manageable. Resources could consist in capacities, infrastructure and time;
- Look also for the infrastructure and tools of the communities you are aiming at and what scheme and data this infrastructure is capable to consume.

4 Interoperability and Identifiers

4.1 General Aspects of Identifiers

Identifiers are used in every aspect of our daily life, both in the real and the digital world, to refer to objects in an unambiguous manner. By means of identifiers objects become addressable and traceable also over time, provided that they remain valid (i.e. persistent).

In the following section, where the focus is specifically on identifiers in a digital environ ment, two main aspects are covered: on the one hand, the persistence over time of the reference contained within digital identifiers and, on the other hand, the use of common, shared sets of identifiers in order to achieve greater interoperability at the semantic level. In fact, these two aspects are intertwined: the persistence of identifiers is the *conditio sine qua non* for users and content producers to be able to refer to the objects they are dealing with. This can be seen in practice in what happens, for example, in the field of Classics to citations of publications available both in printed and electronic form: authors tend to avoid providing links to the online version of cited publications because of the fragility of URLs: although URLs are identifiers of resources on the Web, they are not persistent identifiers.

The role of identifiers in relation to interoperability lies mainly on the semantic level. Using an identifier for a place name, for instance, allows us to express in an unambiguous way, understandable also to software agents, which specific place is being referred to.

4.2 Persistent Identifiers

"Persistent identifiers are simply maintainable identifiers that allow us to refer to a digital object - a file or set of files, such as an e-print (article, paper or report), an image or an installation file for a piece of software [...]; persistent identifiers are supposed to continue to provide access to the resource, even when it moves to other servers or even to other organisations" (Tonkin 2008). Digital resources should always be provided with Persistent Identifiers.

Let us consider now in detail some of the available approaches to making identifiers persistent. More exhaustive overviews on PID solutions can be found in Hilse and Kothe 2006, Tonkin 2008 and Hakala 2010.

4.2.1 PURL - Persistent Uniform Resource Locators

Purls (PURL Administration 2014) are "Web addresses that act as permanent identifiers. The creation and management of PURLs is made easier by the existence of a REST API for which clients in several programming languages can be easily implemented. PURL, developed by OCLC, a non-profit consortium of library organizations in the United States, never became an IETF standard such as for example URN.

4.2.2 URN - Uniform Resource Name

The URN specification (Moats 1997) defines the syntax of names that can be assigned to resources on the Internet. URNs "are intended to serve as persistent, location-independent, resource identifiers".

The basic structure common to all URNs is `urn:<NID>:<NSS>` where NID indicates a Namespace Identifier and NSS is a Namespace Specific String. The string "`urn:nbn:de:bvb:19-146642`" is a valid URN in the National Bibliography Number namespace. This means that the syntax of the part following the second colon ":" is described in the NBN specifications.

One of the key aspects of URNs is the separation between the string acting as a persistent identifier and the technical services that are able to resolve that identifier. The main consequence of this heavily decentralised approach is that a single, global service aware of namespace-specific resolution services does not exist.

Handle System

"The Handle System includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store identifiers, known as handles, of arbitrary resources and resolve those handles into the information necessary to locate, access, contact, authenticate, or otherwise make use

of the resources." ("Handle System" 2014) It "offers currently the most robust and performant PID resolution system" (CLARIN 2008).

The handle defined by the European Research Consortium have the following syntax:

```
prefix/flag-institution-num1-num2-num3-checksum
e.g. 11858/00-XXXX-0000-0000-0000-C
```

DOI - Digital Object Identifier

According to ISO 26324 a "DOI name is permanently assigned to an object to provide a resolvable persistent network link to current information about that object, including where the object, or information about it, can be found on the Internet. While informa tion about an object can change over time, its DOI name will not change. A DOI name can be resolved within the DOI system to values of one or more types of data relating to the object identified by that DOI name, such as a URL, an e-mail address, other identifiers and descriptive metadata."(International DOI Foundation 2014a)

Syntax: "The DOI name syntax specifies the construction of an opaque string with naming authority and delegation. It provides an identifier "container" which can accommodate any existing identifier. The DOI name has two components, the prefix and the suffix, which together form the DOI name, separated by the "/" character. The portion following the "/" separator character, the suffix, may be an existing identifier, or any unique string chosen by the registrant. The portion preceding the "/" character (the prefix) denotes a unique naming authority" (International DOI Foundation 2014b).

4.3 Common Identifiers

Common identifiers are suitable to unambiguously denoting concepts, places and persons. Controlled vocabularies for place names, person names and subjects/concepts are in particular appropriate for (Digital) Humanities.

4.3.1 Identifiers for Places and Place Names

TGN (Getty Research Institute 2014) - The Getty Thesaurus of Geographic Names provides each place record [...] by a unique numeric ID". This ID can also be used for vari ants (historic place names, place name in different languages (e.g. "Wien", "Vienna"). DARIAH-DE provides as part of its technical infrastructure a RESTful interface to the TGN.

> Example: Augusta Vangionum (roman), Borbetomagus (celtic) or וורמסיא (hebrew) are only a few of the historic names denoting the place currently known as "Worms" (Latitude: 49.6356 Longitude: 8.3597). By using the TGN-ID 7005108, the occurrence of different names used in different sources could nevertheless be clearly identified as referring to the same place.
>
> http://ref.dariah.eu/tgnsearch/tgnquery.xql?id=7005108 will provide all information on Worms stored in the TGN.

GeoNames (Geonames 2014a) is a geographical database containing over 10 million geographical names thoroughly categorized in nine classes and feature codes both accessible by several web services (Geonames 2014b).

Pleiades ("Pleiades" 2014) is a community-build gazetteer and graph which provides IDs and addressable, stable URIs for 34.372 ancient places. Pleiades URIs are used within the previously mentioned Pelagios project (see section 2.2.7, p. 14) to provide a shared vocabulary for expressing annotations that involve geographical place names. The approach to semantic interoperability of Pelagios relies heavily on the use of a common set of stable URIs in order to express unambiguously the semantics of annotations.

The PND Personennamendatei (included in GND Gemeinsame Normdatei since 2012) contains 2,600,000 entries with unique IDs called PND-Nummer. PND is addressable via a DARIAH-DE REST service.

VIAF – the Virtual International Authority File – links several national authority files.

4.3.2 Identifiers for Subjects

DDC and LCSH are classification systems both originating from libraries and in the core an abstraction of the library shelves.

DDC - Deweys Decimal Classification is a hierarchical classification system tending to cover all aspects of human knowledge. It is divided in 10 classes each of them again divided in subclasses. DDC is maintained by the Online Computer Library Center (OCLC), which implemented dewey.info as an (experimental) Linked Data platform, wherein every class is identified with a URI.

A 'rival' classification system is the LCSH - Library of Congress Subject Headings maintained by the Library of Congress. A Linked Data Service to Authorities and Vocabularies is set up at id.loc.gov.

The DARIAH-DE Service "Interoperability through Standard Data" ("Interoperabilität durch Normdaten)" will refer to TGN, PND, DDC, ISO 8601 and ISO 3166. It is built on the reference Data Service Cone (Max Planck-Digital Library 2014).

4.4 Recommendations

We strongly recommend the use of identifiers as a means to achieve interoperability between data and digital resources with different provenance, from different data pools and often different disciplinary backgrounds and points of view the use of identifiers is strongly recommended. The persistence of such identifiers is a key aspect.

We recommend that, whenever possible, common identifiers are used in order to provide controlled vocabularies to refer to "things" in an unambiguous fashion. Since the availability of such identifiers varies from discipline to discipline (Digital Classicist 2014), we encourage individuals but particularly institutions that have an active role in producing and publishing electronic resources to work towards 1) providing such sets of identifiers and 2) guaranteeing that resolution services for such identifiers are made persistent in the long-term.

DARIAH-DE provides a PID service for research data which is based on the HANDLE system and is being developed within workpackage 1.2. This service is part of the European Persistent Identifier Consortium (EPIC), on which also CLARIN – another

European project for digital infrastructure – relies in terms of infrastructure for PIDs (CLARIN 2008). We recommend that DARIAH partners apply for an EPIC account to be able to issue PIDs for the resources they develop and publish online.

However, the persistence of the identifier has nothing to do with the persistence of the resource identified by that identifier, as we have seen above. Therefore, institutions have to take care not only of assigning PIDs to resources but also to devise workflows to guar antee that assigned PIDs are updated whenever the resource location is changed.

5 Licensing Data in the Digital Humanities

The following discussion should raise awareness for the importance of data licensing and help researchers to get an overview on how to apply a licence, which licences should be suitable for their research data and what should be taken into consideration for choosing a licence. The information given focuses on the legal context, in particular with regard to data and resources used in the Digital Humanities. These recommendations do not replace the need for competent legal advice by a lawyer in case a researcher wants to licence data, but it is meant to give an introduction to this topic (Ball 2012).

5.1 Bespoke versus Standard Licence

There are two options for licences – bespoke and standard licences. Bespoke licences are individually defined and customized licences. A drawback of a bespoke licence is that it always requires a human to read it before accessing data. Data owners should therefore consider whether it is possible to choose a standard licence. In the last decades, several initiatives have worked on defining open standard licences. Those which are especially interesting for licensing data in the Digital Humanities are introduced in the following section.

5.2 Approaches for a Solution – Open Standard Licences

There are several reasons for researchers to open up their data to others. Projects are required to open data if they request public funding. Funders intend to reduce costs by making data available for reuse. But opening data enforces the need to decide under which licence research data should be published. The current situation of missing licences for research data is unsatisfactory as it does not allow researchers willing to reuse data to estimate the risk for infringement.

If data owners wish to open their data, they should therefore licence it preferably under an open standard licence. Which licences are available?

5.2.1 Creative Commons

The Creative Commons licensing framework is probably the best known open licensing standard (Science Commons 2014). It offers a choice of six unported licences:

- Attribution (CC BY)*:* This licence lets others distribute, remix, and build upon the published data. Credit for the original creation is required. This is the most accom

modating of licences offered. Recommended for maximum dissemination and use of licensed materials.

- Attribution Share Alike (CC BY-SA)*:* This licence lets others remix, and build upon data even for commercial purposes, as long as they credit the original creation and license their new creations under identical terms. This licence is often compared to "copyleft" free and open source software licences.
- Attribution No Derivatives (CC BY-ND): This licence allows for redistribution, commercial and non-commercial, as long as the content is passed along unchanged and in its entirety, with credit to the original creation.
- Attribution Non-Commercial (CC BY-NC)*:* This licence lets others remix, and build upon data non-commercially, and although their new works must also acknowledge the original creation and be non-commercial, they do not have to license their derivative works on the same terms.
- Attribution Non-Commercial Share Alike (CC BY-NC-SA)*:* This licence lets others remix and build upon data non-commercially, as long as they credit the original creation and license their new creations under the identical terms.
- Attribution Non-Commercial No Derivatives (CC BY-NC-ND)*:* This licence is the most restrictive of the six main licences, only allowing others to download the data and share it with others as long as they credit the original creation, but they cannot change it in any way or use it commercially.

There is also a movement in the Creative Commons community where people have started to port the Creative Commons licences to the law of their countries. Currently there are about 550 ported licences available. Not all ported licences are compatible with each other. Because of this it is hardly possible to keep track of the many changes in ported licences. Therefore, data owners should whenever possible favor unported CC licences (de Rosnay 2009).

5.2.2 Europeana Rights Statements

In case of bespoke licences it would be helpful to support machine-readability of legal information by choosing an additional Europeana rights statement. Rights are still reserved in all cases but there is a need to inform users of differing levels of access to the data online and to point to the site where more information about rights can be found. The following Europeana rights statements are available (Europeana 2010):

- Europeana: Rights Reserved - Free Access: is applicable when users have free (as in gratis), direct and full access to the digitized object. Needing to register or other administrative procedures in order for users to gain access to the digitized object should not limit the user;
- Europeana: Rights Reserved - Paid Access: is applicable when users need to pay data providers to gain access to the data and therefore need to register;
- Europeana: Rights Reserved - Restricted Access: is applicable when users are limited in accessing data for reasons other than needing payment;

- Europeana: Unknown copyright status: applies to data where the data provider does not have conclusive information pertaining to the rights status. This value is only to be used when the copyright status of the work described is unknown.

5.2.3 Open Data Commons

The Open Data Commons ("Open Data Commons" 2014) licences were especially designed to fit the purpose of opening data and databases for reuse. ODC is an Open Knowledge Foundation project. There are currently three licences to choose from:

1. Open Data Commons Attribution Licence v1.0 (ODC-By): Users are free to copy, distribute and use the database; to produce works from the database; to modify, transform and build upon the database; as long as they attribute any public use of the database, or works produced from the database, in the manner specified in the licence. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database.
2. Open Data Commons Open Database Licence v1.0 (ODC-ODbL): Users are free to copy, distribute and use the database; to produce works from the data base; to modify, transform and build upon the database; as long as they attribute any public use of the database, or works produced from the database, in the manner specified in the ODbL. For any use or redistribution of the database, or works produced from it, they must make clear to others the licence of the data base and keep intact any notices on the original database. If users publicly use any adapted version of this database, or works produced from an adapted database, they must also offer that adapted database under the ODbL. If they redistribute the database, or an adapted version of it, then they may use technological meas ures that restrict the work (such as DRM) as long as they also redistribute a version without such measures.
3. Open Data Commons Database Contents Licence v1.0 (ODC-DbCL): To waive all rights in the individual contents of a database licensed under the ODbL above.

5.2.4 Public Domain

If data owners want to open their data and databases without restrictions they may do so declaring their data to be in the Public Domain. Creative Commons and Open Data Commons provide waivers for this:

- CC0: enables scientists, educators, artists and other creators and owners of copy right- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.
- Open Data Commons Public Domain Dedication and Licence (PDDL): Users are free to copy, distribute and use the database; to produce works from the data

base; to modify, transform and build upon the database. The PDDL imposes no restrictions on the use of the PDDL licensed database.

Please note that by German copyright law the copyright/author's right itself can neither be transferred to another person nor waived by the author herself, meaning that the above mentioned waivers are not legally valid (Kreutzer 2011, 15). But the Public Licence Fallback in sec. 3 CC0 serves as an alternative to the waiver in cases where a full waiver of some or all rights in the work is not possible under the respective applicable law (Creative Commons 2014a).

5.3 Machine-readability of Licences

One of the advantages of greater interoperability is that the (need for) human interven tion is reduced when information is exchanged between heterogeneous systems. However, in order for this to be realized fully, we need not only human-readable licenses but also machine-readable ones.

For the sake of example let us consider a dataset licensed under a Creative Commons licence. Human-readability is achieved when the licence is attached to the data as a text file which contains information about the adopted policy. Machine-readability, instead, means that information about which licence applies to the data is expressed in a language that can be understood by a software agent, such as a markup language.

At this point it can be also useful to reflect on what it actually means to be machine-readable and/or machine/understandable. The software agent, by reading licence information expressed as Dublin Core or RDF triples gets to know that licence X applies to the dataset Y. However, this does not mean that the same agent knows which actions (e.g. copy of files) are allowed and not allowed by that licence unless it is somehow instructed to do so.

Since OAI-PMH was recommended as minimum machine-interface to adopt when publishing data online, it is worth mentioning that it is possible, within an OAI-PMH repository, to include licence-related information and also, should it be necessary, to specify the granularity of such licences. In fact, in some cases one single licence applies to the entire dataset whereas in other cases one may want, or have to, attach different licences to different subsets (Open Archives Initiative 2005). OAI-PMH uses the rights property from Dublin Core to refer to any applying licence; when publishing data on the Web specifically under a CC licence, one can and should use Creative Commons Rights Expression Language (Creative Commons 2014b), an RDF specification for Copyright licences (Creative Commons 2014c).

5.4 Recommendations for Best Practices

To help improving interoperability owners of research data should consider the following recommendations for licensing:

- Integrate the license decision into the data publishing workflow of your institution;

- In case research data is collected in a funded project, the declaration under which licence this data will get published should be implemented in the proposal process of a project;
- Three essential pieces of license information should always be provided: the name of the rights holder, the year of publication of the data collection (i.e. the year in which the rights began to be exercised) and the type of licence applied to it;
- The use of open standard licences to improve interoperability is preferred;
- Make sure you have all the rights in connection with the data you wish to publish;
- Decide whether you consider permitting commercial or non-commercial use of your data;
- Creative Commons licences are all non-exclusive, meaning a Creative Commons licence and a bespoke non-exclusive licence parallel in use for the same data is allowed, but this causes legal conflicts and should be avoided;
- A human-readable, machine-readable and lawyer-readable version of your licence would be best practice;
- Be aware that different licences may apply to different parts of your data. There fore, select a licence separately for metadata, vocabularies, digital resources (Image, full text, audio file etc.), databases, and data from third parties included.

6 References

Note: All internet resources cited have been accessed on January 13, 2014.

Ball, Alex. 2012. *How to License Research Data*. DCC How-to Guide. Edinburgh. http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf.

Bauman, Syd. 2011. "Interchange vs. Interoperability." In *Proceedings of Balisage: The Markup Conference 2011*. Vol. 7. Balisage Series on Markup Technologies. Montréal, Canada. doi:10.4242/BalisageVol7.Bauman01. http://www.balisage.net/Proceedings/vol7/html/Bauman01/BalisageVol7-Bauman01.html.

Binding, Ceri. 2014. *Stellar · GitHub*. https://github.com/cbinding/stellar.CC:DA (ALCTS/CCS/Committee on Cataloging: Description and Access). 2000. "Task Force on Metadata. Final Report". Association for Library Collections & Technical Services. http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html.

CLARIN (Common Language Resources and Technology Infrastructure). 2008. "Persistent and Unique Identifiers." http://www.clarin.eu/sites/default/files/wg2-2-pid-doc-v4.pdf.

Creative Commons. 2014a. "CC0 1.0 Universal." http://creativecommons.org/publicdomain/zero/1.0/legalcode.

———. 2014b. "CC REL by Example." http://labs.creativecommons.org/2011/ccrel-guide/.

———. 2014c. "Describing Copyright in RDF - Creative Commons Rights Expression Language." http://creativecommons.org/ns.

Crockford, D. 2006. "RFC 4627 - The Application/json Media Type for JavaScript Object Notation (JSON)". Network Working Group. https://tools.ietf.org/html/rfc4627.

De Rosnay, M. D. 2009. "Creative Commons Licenses Legal Pitfalls: Incompatibilities and Solutions." http://halshs.archives-ouvertes.fr/halshs-00671622/.

Digital Classicist. 2014. "Very Clean URIs - The Digital Classicist Wiki." http://wiki.digitalclassicist.org/Very_clean_URIs.

Doerr, Martin. 2003. "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata." *AI Magazine* 24 (3): 75. doi:10.1609/aimag.v24i3.1720. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1720.

Dublin Core Metadata Initiative. 2014a. "DCMI History." http://dublincore.org/about/history/.

———. 2014b. "Using Dublin Core." http://dublincore.org/documents/usageguide/.

Europeana. 2010. "Guidelines for the Europeana:rights Metadata Element." http://pro.europeana.eu/c/document_library/get_file?uuid=06e63d96-0358-4be8-9422-d63df3218510&groupId=10602.

Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and Tim Berners-Lee. 1999. "RFC 2616 - Hypertext Transfer Protocol - HTTP/1.1". Network Working Group. https://tools.ietf.org/html/rfc2616.

Franklin, Michael, Alon Halevy, and David Maier. 2005. "From Databases to Dataspaces: A New Abstraction for Information Management." *ACM Sigmod Record* 34 (4): 27–33. http://portal.acm.org/citation.cfm?id=1107499.1107502.

Freed, N., and N. Borenstein. 1996. "RFC 2046 - Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types". Network Working Group. https://tools.ietf.org/html/rfc2046.

Geonames. 2014a. "About GeoNames." http://www.geonames.org/about.html.

———. 2014b. "GeoNames Webservice and Data Download." http://www.geonames.org/export/#ws.

Getty Research Institute. 2014. "About the TGN - Getty Thesaurus of Geographic Names Online." http://www.getty.edu/research/tools/vocabularies/tgn/about.html.

Hakala, Juha. 2010. "Persistent Identifiers - an Overview." http://metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview/.

"Handle System." 2014. http://www.handle.net/.

Haslhofer, Bernhard, and Wolfgang Klas. 2010. "A Survey of Techniques for Achieving Metadata Interoperability." *ACM Computing Surveys* 42 (2): 1–37. doi:10.1145/1667062.1667064. http://dl.acm.org/citation.cfm?id=1667064.

Haslhofer, Bernhard, and Erich J. Neuhold. 2011. "A Retrospective on Semantics and Interoperability Research." In *Foundations for the Web of Information and Services*, edited by Dieter Fensel, 3–27. Berlin, Heidelberg: Springer Berlin Heidelberg. http://cs.univie.ac.at/research/research-groups/multimedia-information-systems/publikation/infpub/2921/.

Hilse, Hans-Werner, and Jochen Kothe. 2006. "Implementing Persistent Identifiers". Consortium of European Research Libraries, European Commission on Preservation and Access.

"Homer Multitext Project: Documentation." 2014. http://www.homermultitext.org/hmt-doc/.

International DOI Foundation, ed. 2014a. "Einleitung." In *DOI Handbook Introduction*. http://www.doi.org/doi_handbook/1_Introduction.html#1.5.

———. 2014b. "DOI Name Syntax." In *DOI Handbook Introduction*. http://www.doi.org/doi_handbook/1_Introduction.html#1.6.3.

"Item 97-L-19(549)". 2014. *Open Context. Web-based research data publication*. Formats: ArchaeoML: http://opencontext.org/subjects/E54B6571-265E-45C1-054D-C272E8515E8D.xm, Atom: http://opencontext.org/subjects/E54B6571-265E-45C1-054D-C272E8515E8D.atom, and HTML: http://opencontext.org/subjects/E54B6571-265E-45C1-054D-C272E8515E8D.

Kansa, Eric C., Tom Elliott, Sebastian Heath, and Sean Gillies. 2010. "Atom Feeds and Incremental Semantic Annotation of Archaeological Collections." In *CAA 2010 Fusion of Cultures: Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology*. Granada.

Kreutzer, Till. 2011. "Validity of the Creative Commons Zero 1.0 Universal Public Domain Dedication and Its Usability for Bibliographic Metadata from the Perspective of German Copyright Law". Berlin. http://pro.europeana.eu/c/document_library/get_file?uuid=29552022-0c9f-4b19-b6f3-84aef2c3d1de&groupId=10602.

Lagoze, Carl, and Herbert Van de Sompel. 2014. "Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0." http://www.openarchives.org/OAI/openarchivesprotocol.html.

Lagoze, Carl, Herbert Van de Sompel, Pete Johnston, Michael Nelson, Robert Sanderson, and Simeon Warner. 2008. "ORE User Guide - Resource Map Implementation in Atom." http://www.openarchives.org/ore/1.0/atom.

Marshall, Catherine C., and Frank M. Shipman. 2003. "Which Semantic Web?" In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, 57–66. Nottingham, UK: ACM. doi:10.1145/900051.900063. http://portal.acm.org/citation.cfm?id=900063.

Max Planck-Digital Library. 2014. "Service for Control of Named Entities - MPDLMediaWiki." http://colab.mpdl.mpg.de/mediawiki/Service_for_Control_of_Named_Entities.

Moats, R. 1997. "RFC 2141: URN SYNTAX." http://www.ietf.org/rfc/rfc2141.txt.

NISO, ed. 2004. *Understanding Metadata*. http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

Nottingham, M., and R. Sayre, ed. 2005. "RFC 4287 - The Atom Syndication Format". Network Working Group. http://www.ietf.org/rfc/rfc4287.txt.

"OAI-PMH Static Repository Gateway." 2014. http://srepod.sourceforge.net/.

Open Archives Initiative. 2005. "Conveying Rights Expressions about Metadata in the OAI-PMH Framework. Version 2.0. http://www.openarchives.org/OAI/2.0/guidelines-rights.htm.

Open Archives Initiative. 2014. "Protocol for Metadata Harvesting. http://www.openarchives.org/pmh/.

Open Data Commons. 2014. http://opendatacommons.org/.

Pleiades. 2014. http://pleiades.stoa.org/.

PURL Administration. 2014. http://purl.org/docs/purl.html.

Pytlik Zillig, Brian L. 2009. "TEI Analytics: Converting Documents into a TEI Format for Cross-Collection Text Analysis." *Literary and Linguistic Computing* 24 (2) : 187–192. doi:10.1093/llc/fqp005. http://llc.oxfordjournals.org/content/24/2/187.abstract.

Romanello, Matteo. 2014. *Skosifaurus · GitHub*. https://github.com/mromanello/skosifaurus.

Sahle, Patrick. 2002. "Sinnsuche in der Badewanne. Tagungsbericht: Standards und Methoden der Volltextdigitalisierung." *Jahrbuch für Computerphilologie* 4. http://computerphilologie.uni-muenchen.de/jg02/sahle.html.

Science Commons. 2014. "Science Commons » About Science Commons." http://sciencecommons.org/about/.

"TEI: EpiDoc." 2014. http://www.tei-c.org/Activities/Projects/ep01.xml.

"TEI: TEI Lite." 2014. http://www.tei-c.org/Guidelines/Customization/Lite/.

"The Apache HTTP Server Project." 2014. https://httpd.apache.org/.

Tonkin, Emma. 2008. "Persistent Identifiers: Considering the Options." *Ariadne* (56) . http://www.ariadne.ac.uk/issue56/tonkin/.

Unsworth, John. 2011. "Computational Work with Very Large Text Collections." Edited by Kevin Hawkins, Malte Rehbein, and Syd Bauman. *Journal of the Text Encoding Initiative* (1). Selected Papers from the 2008 and 2009 TEI Conferences (June 8). http://jtei.revues.org/215.

Van de Sompel, Herbert, and Carl Lagoze. 2014. "Static Repository Example." In *Specification for an OAI Static Repository and an OAI Static Repository Gateway*. http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm#SR_example.

Van de Sompel, Herbert, Carl Lagoze, Michael Nelson, and Simeon Warner. 2004. "OAI-PMH Implementation Guidelines - Specification for an OAI Static Repository and an OAI Static Repository Gateway." http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm.

Wikipedia. 2014. "Hypertext Transfer Protocol - Request Methods." https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol#Request_methods.

7 List of Abbreviations

- API - application programming interface
- CC - Creative Commons
- CIDOC-CRM International Committee for Documentation - Conceptual Reference Model
- CTS - Canonical Text Services protocol
- DC - Dublin Core
- DDC - Dewey Decimal Classification
- DDI - Data Documentation Initiative
- DOI - Digital Object Identifier
- EDM - Europeana Data Model
- EPIC - European Persistent Identifier Consortium
- FRBR - Functional Requirements for Bibliographic Records
- GND - Gemeinsame Normdatei
- IETF - Internet Engineering Task Force
- LCSH - Library of Congress Subject Headings
- LD - Linked Data
- LOD - Linked Open Data
- METS - Metadata Encoding and Transmission Standard
- MODS - Metadata Object Description Schema
- OAI-ORE - Open Archives Initiative Object Reuse and Exchange format
- OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting
- OCLC - Online Computer Library Center
- ODC - Open Data Commons
- OWL - Web Ontology Language
- PID - Persistent Identifier
- PND - Personennamendatei
- PURL - Persistent Uniform Resource Locator
- RDF - Resource Description Framework
- RFC - Request for Comments

- SKOS - Simple Knowledge Organization System
- TEI - Text Encoding Initiative
- TGN - The Getty Thesaurus of Geographic Names
- URI - Uniform Resource Identifier
- URL - Uniform Resource Locator
- URN - Uniform Resource Names
- VIAF - Virtual International Authority File

8 Appendix: OAI Example

```
<?xml version="1.0" encoding="UTF-8"?>
<Repository xmlns:oai="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.openarchives.org/OAI/2.0/static-repository"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/static-repository
http://www.openarchives.org/OAI/2.0/static-repository.xsd">
    <Identify>
        <oai:repositoryName>Kalonymos Contributions</oai:repositoryName>
        <oai:baseURL>http://gateway.institution.org/oai/
                    www.steinheim-institut.de/edocs/oai/kalonymos-
                    contributions.xml</oai:baseURL>
        <oai:protocolVersion>2.0</oai:protocolVersion>
        <oai:adminEmail>oai@steinheim-institut.org</oai:adminEmail>
        <oai:earliestDatestamp>2002-09-19</oai:earliestDatestamp>
        <oai:deletedRecord>no</oai:deletedRecord>
        <oai:granularity>YYYY-MM-DD</oai:granularity>
    </Identify>
    <ListMetadataFormats>
        <oai:metadataFormat>
            <oai:metadataPrefix>oai_dc</oai:metadataPrefix>
            <oai:schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd
            </oai:schema>
            <oai:metadataNamespace>http://www.openarchives.org/OAI/2.0/
                                   oai_dc/</oai:metadataNamespace>
        </oai:metadataFormat>
        <oai:metadataFormat>
            <oai:metadataPrefix>oai_rfc1807</oai:metadataPrefix>
            <oai:schema>http://www.openarchives.org/OAI/1.1/rfc1807.xsd
            </oai:schema>
            <oai:metadataNamespace>http://info.internet.isi.edu:80/
                   in-notes/rfc/files/rfc1807.txt</oai:metadataNamespace>
        </oai:metadataFormat>
    </ListMetadataFormats>
    <ListRecords metadataPrefix="oai_dc">
        <oai:record>
            <oai:header>
                <oai:identifier>oai:www.steinheim-institut.de:kalonymos:
                      contributions:adde82d6-a988-11e1-9c05-002719b0d498
                </oai:identifier>
                <oai:datestamp>2012-04-01</oai:datestamp>
            </oai:header>
            <oai:metadata>
                <oaidc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
                xmlns:dcterms="http://purl.org/dc/terms/"
                xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai_dc/"
          xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
                    <dc:title>Andante, attacca: Der
                              jüdisch-polnisch-russische Komponist
                              Mieczyslaw Weinberg
                    </dc:title>
```

```
                    <dc:creator>Michael Brocke und Annette Sommer
                    </dc:creator>
                    <dc:subject/>
                    <dc:description/>
                    <dc:publisher>Salomon Ludwig Steinheim-Institut für
                         deutsch-jüdische Geschichte an der Universität
                         Duisburg-Essen
                    </dc:publisher>
                    <dc:contributor/>
                    <dc:date>2010</dc:date>
                    <dc:type>Text</dc:type>
                    <dc:format>PDF</dc:format>
                    <dc:identifier>http://www.steinheim-institut.de
                             /edocs/kalonymos/kalonymos_2010_4.pdf#page=1
                    </dc:identifier>
                    <dc:identifier>urn:nbn:de:0230-20090805284
                    </dc:identifier>
                    <dc:source/>
                    <dc:language>de</dc:language>
                    <dc:relation>Kalonymos. Beiträge zur deutsch-
                              jüdischen Geschichte aus dem Salomon Ludwig
                              Steinheim-Institut an der Universität
                              Duisburg-Essen, 13 (2010), Nr. 4, S. 1-5
                    </dc:relation>
                    <dc:coverage/>
                    <dc:rights/>
                </oaidc:dc>
            </oai:metadata>
        </oai:record>
    </ListRecords>
</Repository>
```

Listing 4: OAI Example