Report on Enhancing Interoperability between existing Open Access Publication Infrastructures

---

Anouar Boulal , Martin Iordanidis, Andres Quast, Jochen Schirrwagen

hbz – Library Service Center NRW, Cologne
Bielefeld University Library, Bielefeld

September 2st, 2010

status: draft

Content

Introduction

Within the last decade, the concept of enhanced publications has become more and more prominent. It promises the prospect of more comprehensive communication among scientists. By using electronic publication infrastructures, scientists can provide their colleagues with online access to their published manuscripts - along with various kinds of associated materials.

Among the latter we find research data, visualizations or instructive materials. Ideally, those enhanced publications provide more complete research results, thoroughly improve accessibility and facilitate the verification and re-use of scientific information.

A website presenting the publication and all its associated materials as hyperlinks seems to be appropriate and adequate from a human's point of view.

Mere technology and machines cannot interpret information given in a web page the same way human scientists do. For example, search engines cannot distinguish between raw data and an instructive video that is to be found at the end of a conventional hyperlink without additional semantic information. Hence, in order to give scientists the most sufficient and easy access to the information provided by these new forms of publications, the automated processing of enhanced publications becomes an essential desideratum. New mechanisms are required allowing software systems to interpret inherent relations.

Expressing the relations within all parts of an enhanced publication is a strong requirement. Descriptive and technical metadata are common tools that provide this data to information systems. The OAI-ORE standard provides a useful framework to describe relations and manage the content of enhanced publications. However, practical implementations within existing infrastructures are not well established yet. This is why practical approaches are addressed by the eco4r project.

The aim of the eco4r project is to use state-of-the-art-techniques for exposure and re-use of enhanced publications stored in living repository systems. Existing publications derived from the project partner's repositories will be examined in consideration of their practical interchangeability and interoperability. Re-use and interoperability, in this context, also applies to designated operation purposes such as long term preservation. The inevitable need of curating digital data in the long term adds a whole chain of requirements to the design of compound objects. For example, the dependency on several distinct format migrations within one

compound object demands that its constituent parts are technically described on an atomic level.

The repositories involved in the project context are quite different in terms of the deployed software systems as well as the materials stored within them.

The Bielefeld University Library uses the German software system OPUS for its repositories. It stores primarily dissertations as well as post-prints of published articles and others. The Library Service Center in Cologne (hbz) operates and manages the so-called Digital Peer Publishing platform (DiPP). DiPP is a Fedora-driven infrastructure for Open Access-journals which currently hosts 17 peer reviewed journals[1]. DiPP strongly promotes the idea of enhanced publications to publishing scientists.

The practical approach of eco4r is likely to be a living example for interoperability between existing Open Access publication infrastructures that fits into the Semantic Web. One important starting point of the eco4r project is the quantitative and qualitative analysis of materials residing in the source repositories.

As a result of the analysis, this draft report will suggest various requirements and recommendations concerning the creation and post-processing of the compound objects. They will be put up for discussion to the designated community.

State of the Art / Related Initiatives

Since Open Access publications have been widely distributed in uncounted systems all over the world, the need for virtual aggregations and easy-to-use gateways has become inevitable. With the OAI-PMH, the Open Archives Initiative (OAI) introduced a light-weight protocol which makes metadata visible and exchangeable among information systems. Its underlying principles of data interchange and services acting upon it provide a usable approach for the exposure and visibility of distributed publications (DRIVER). Nevertheless, the limits of a mere metadata exchange were discovered soon after and are becoming more conspicuous with evolving enhanced publications. This is due to the fact that internal structures and relations of these publications remain hidden within the OAI-PMH.

Driven by these conclusions, the Open Archives Initiative developed the OAI-ORE (OAI Object Re-use and Exchange) standard. Released in 2008, the standard adopts state-of-the-art semantic web techniques in order to provide a useful and extensible tool for the exchange and re-use of compound objects such as enhanced publications. Different projects and initiatives have begun to evaluate the capabilities of the standard since then. Because of their relevance for the eco4r project, the outcomes of these projects and initiatives are outlined in the following.

[1] a list of all e-journals published by DiPP can be found under the address: http://www.dipp.nrw.de/journals/

DRIVER-II

During this time, the DRIVER-II project[2] aimed at improving and strengthening the existing European repository infrastructure encouraged by DRIVER-I. The project's objectives are especially taking into account evolving enhanced publications and the OAI-ORE standard which technically handles them.[3]

One outcome of the DRIVER-II project thoroughly defines enhanced publications as publications enriched with three categories of information:

- research data (evidence of research)
- extra materials (to illustrate or clarify)
- post publication data (commentaries, ranking)

This definition distinguishes enhanced publications from compound objects in general and thereby focuses clearly on academic needs and scientific workflows. In this sense, an enhanced publication is a compound object that combines a textual resource (human readable) with at least one of supplementary information. This could be one or more data resources , various kinds of metadata records as well as a combination of different supplementary materials.

Within DRIVER-II, requirements and recommendations for enhanced publications were identified. As an example, they comprise their modeling, localization, properties, preservation and description of semantics.

ESCAPE: Enhanced Scientific Communication by Aggregated Publications Environments

The ESCAPE project[4] was a collaboration of three participants, namely the University of Groningen, the University of Twente and the Royal Netherlands Academy of Arts and Sciences funded by SURF. The project aims at the semi-automated / semi-manual enhancement of existing publications with information from distributed repositories and other web resources.

With the development of so-called aggregated publications environments (APE), ESCAPE's goal was to enable its users to enhance existing publications by adding supplementary materials and/or create a network of related information objects. The project outcome was a tool for the creation and editing of such aggregations. They will be stored as OAI-ORE resource maps in a resource repository. The latter is a

[2] http://www.driver-repository.eu/Driver-About/About-DRIVER.html
[3] The demonstrator of Enhanced Publications is one of the results: http://driver2.dans.knaw.nl/
[4] http://escapesurf.wordpress.com/

Fedora-based system which allows interoperability with various systems by using standards such as FOAF, DCMI TERMS, SKOS, SWAN and OAI-ORE.

Although the ESCAPE APEs are an exciting way to enhance the publication's value by providing supplementary information, human interaction is still required to drive this system. In order to ensure the quality of aggregations, two principles are followed: Firstly, peer group knowledge is required. Secondly, the enrichment tools have to be restricted to a known group of persons like institutional staff or scientific workgroups. Although the stored resource maps are publicly available, the creation of new aggregations will remain in a few person's responsibility.

eTheses Project

The eTheses project[5] of the Knowledge Exchange initiative is a joined effort of Denmark's Electronic Research Library (DEFF), German Research Foundation (DFG), Joint Information Systems Committee (JISC) and SURF foundation in the Netherlands. The project focuses on electronic theses and dissertations (ETD) rather than enhanced publications in general. Nevertheless, it also addresses enhanced publications and provides useful theoretical and practical approaches to the eco4r project. Within the project, experiences and concepts regarding the management of compound objects stored in a Fedora repository are discussed.

By using the concepts of atomistic, compound and complex objects provided by the ARROWS project, we can gain two insights regarding the eco4r project: the concept used in the DiPP system represents an approach between compound and complex objects whereas the publications found at the Bielefeld OPUS Repository follow a nearly-atomistic concept.

Although conceptually not comparable, the information provided by both systems should be accessible and interoperable from the scientist’s perspective. This means that, in an ideal sense, any kind of object has the same technical interchangeability to other systems regardless of its origin. Therefore, the eco4r project is likely to enhance the knowledge and practices acquired by the Knowledge Exchange project in various living information systems.

ICE-Theorem

The Integrated Content Environment – Theses with ORE Metadata (ICE-TheOREM) [Sefton-2010] used OAI-ORE by modeling chemical theses as an aggregate of chapters and supporting information. They provided authoring tools and a workflow for scholarly communication. For that purpose they combined OAI-ORE with the SWORD-App protocol in order to transfer theses and data-rich compound objects across systems for content management , thesis management and repositories.

[5] http://www.knowledge-exchange.info/Default.aspx?ID=334

Long term preservation

Long term preservation of enhanced publications - and compound objects in general - is an upcoming challenge. Since compound publications have become a common and widespread means of scientific communication, concepts and practical approaches retaining the aggregated information units become essential. Different file formats, authenticity and integrity seem to be manageable similarly to other materials. Whenever aggregated materials are considered on a resource level, the completeness of ownership-information , copyrights and provenance becomes an important requirement for retaining valuable information in the future [Doorenbosch and Sierman 2010]. The overlay journal, as it is planned as a demonstrator in eco4r, itself is a compound object that should be preserved. At the same time, the collection is expected to undergo continuous development. Therefore the eco4r project aims to clarify the requirements and preconditions for the preservation of compound objects – both those derived from repository systems and those assembled in an overlay journal. In a similar way, OAI-ORE is expected to be a useful standard for exposure and exchange of compound objects; METS seems to provide a flexible technology to describe compound objects sufficiently for long term preservation.

Standards for the representation of Compound Objects

What is a Compound Object?

In our consideration we use the term "Compound Object" as a generic concept of digital objects with dedicated properties. That is digital objects are composites of distributed web-resources, which form an aggregated object. The components of a Compound Object do not necessarily reside in the same repository. They can rather form some kind of federated information by hyperlinks and semantic relations. Furthermore, it is possible to construct new 'superior' objects by aggregating a set of distributed objects according to logical correlations.

[Doorenbosch and Sierman (2010)] suggest the term compound publications for new kinds of aggregated information resources that can comprise for example:

- Enhanced Publications - scholarly publications, which are a combination of one or more metadata records, full-text publication and supplementary material like data sets.
- Articles in an Overlay Journal - aggregations of distributed web resources which are grouped together under specific aspects, e.g. a scientific topic.

Description of the standards

METS

The 'Metadata Encoding and Transmission Standard' (METS) was primarily created in the digital library and archiving environment. It provides a framework to wrap several metadata types and can be considered as a container format which may hold descriptive, administrative, structural and technical metadata related to a digital object.

In addition to the metadata, a METS document can include any kind of digital content. This characteristic makes METS compatible with the OAIS information package concepts and hence can be used for creating and managing packages for submission (SIP), archival storage (AIP) and dissemination (DIP) in an archive system.

Within a METS container, each metadata type is described in an individual section. Each content item in a METS section is referenced through an internal unique identifier, which is used for reference purposes. The metadata themselves may either be stored inside the METS document or held externally and referenced from the main METS file. This feature may decrease the size of the main METS document but does not reduce the complexity of the framework.

The METS framework is indeed tightly structured but allows for high flexibility in integrating any kind of metadata associated with digital objects and managing them as collections. This flexibility may solve some of the interoperability challenges existing between repositories using different tools and metadata types – however, it does not guarantee that metadata content is or will ever be standardized.

The ability of METS to deal with large and complex digital objects, be they local or distributed, makes it a suitable option to describe compound digital objects. For the eco4R project context, METS will be used as an information package including digital content and metadata associated with a compound object that is to be delivered to an archiving system.

OAI-ORE

The OAI-Object Reuse and Exchange (OAI-ORE) framework has been chosen by the eco4r project as an interoperability layer because it allows describing scholarly work items in a way that is compatible to the web architecture.

This aspect is important, since scientific publications in the digital world are web resources with a location, an identifier and a representation. Due to new ways of scholarly communication like virtual research environments, those resources may have links to research data sets or are constituents of multiple parts. OAI-ORE allows the definition of clear logical boundaries between those resources and allows structuring a group of related objects in a flexible way.

The OAI-ORE data model introduces the entities *Aggregated Resource*, *Aggregation*, *Resource Map (ReM)* and *Proxy* on top of the RDF model, whereas each entity is represented by URIs. An *Aggregated Resource* is a constituent of an *Aggregation*. An *Aggregation* is a set of other resources and can be described by multiple *ReMs*. A *ReM* is a description of a single *Aggregation* and retains provenance information of the *Aggregated Resources*. Since an *Aggregated Resource* can belong to one or more *Aggregations*, a *proxy* resource is used to identify an *Aggregated Resource* in the context of a specific *Aggregation* with its own URI.

Furthermore, the data model supports nested Aggregations, which means Aggregated Resources can be Aggregations of other Aggregated resources. This is a typical use case in publication infrastructures for e-journals.

In order to describe specific publication types like e-journals, dissertations etc. the use of appropriate ontologies has to be considered. OAI-ORE allows extending the data model with further vocabularies and ontologies.

Thus the OAI-ORE framework turns publication items residing in repositories into reusable and exchangeable web resources. Hereby, digital items rely on a standardized and flexible RDF-based data model and at the same time become portable and exchangeable across systems.

Comparing the standards OAI-ORE and METS

Table 1 Comparative table of OAI-ORE and METS

Criteria	OAI-ORE METS
Structure	Graph structure based on RDF Tree structure based on XML
Data Model	Abstract data model as a basis for defined standards for the description and exchange of compound objects A flexible but also tightly structured XML container holding all metadata as separate but linked sections (descriptive, administrative and structural) relating to a single (atomic or compound) digital object
Accessibility and Reuse	• Each resource in the Resource Maps (Named Graphs) has its own protocol-based URI and can be directly dereferenced • Ability to dereference aggregation URIs and to reuse them through the “Nesting Aggregations” mechanism • The specification applies the principles of Linked Data as the basis for HTTP implementation • METS provides only on a document schema and leaves the question of transfer and discovery quite open • Each document (and section) is uniquely identified by an ID-attribute (not necessarily protocol-based) which allows the document to be referenced from another element or document
Extensibility	OAI-ORE does not restrict the use of complementary or domain-

	specific vocabularies. Instead, it points to the guiding principle that consists of reusing vocabularies from terms. These are not specific and fundamental to the ORE model in order to enhance the interoperability
	METS defines placeholder metadata sections to allow the embodiment or reference of standardized complementary XML metadata schemas. There are no restrictions in selecting specific metadata schemas. Hence, some well-known ones have been recommended (like PREMIS as a schema for digital preservation)
Serialization	OAI-ORE objects can be serialized in several data formats including: RDF/XML , RDFa, Atom etc.
	METS is based on the concept of XML packages to express the structure of complex digital objects (content files, metadata and relationships constituting the object)
Usage	Serves as a standard to model and exchange aggregations of web resources between different communities
	• when the aim is to capture the internal structure of a digital object and enrich the resulting document with metadata (e.g. book digitalization) • management and exchange of digital library objects within and between different repository systems • The ability of METS to comprise the content files, descriptive, structural and administrative metadata makes a METS document compatible with the OAIS information package concept and hence can be used as a SIP, AIP and DIP object
Interoperability	• OAI-ORE specification is based on semantic web principles and recommends a number of vocabulary terms to promote interoperability and semantic web technologies • OAI-ORE provides an abstract data model to unify and standardize the expression, reuse and exchange of compound digital objects • OAI-ORE explicitly specifies a serialization based on the Atom Syndication Format (it matches the concepts of the ORE model well and is a widely used-format). Thereby, it promotes interoperability and can avoid format conversion
	METS provides a framework for integrating various metadata schemas and vocabularies, which is indeed an important contribution to the interoperability issue but it cannot guarantee it. It serves only as a metadata container and there is no warranty that while the deployed metadata schemas are standardized
File Size	The ability of OAI-ORE to create nested aggregations by just nesting links to other ORE aggregations could reduce the resulting files to a reasonable size
	METS files could be huge since METS documents comprise information about all components making up a compound object (like a digitized book) in a single file. The various metadata can be either wrapped inside the document itself or stored externally but managed from within the root document

Similarities

A METS file and OAI-ORE Resource Map serve similar purposes. Both have facilities to describe compound objects and identify the various related resources that constitute the objects. In addition, a variety of relationships between those resources can be expressed.

Furthermore, the hierarchy of the METS structural map enables the navigation or browsing through files embedded in, or referenced by, the METS object. A similar effect can be accomplished by making use of the OAI-ORE Resource Maps. Navigating through the component parts of a compound object can also be realized by using ORE Proxies. They are useful whenever relationships between aggregated resources in a manner specific to the Aggregation context are to be achieved. An example is the navigation through a digital book, where the Aggregation represents the book concept and the aggregated resources represent pages.

Current Observations

An analysis of the current state of repositories used and the materials contained within them is crucial to eco4r's practical approach (see chapter 1). The following questions are of interest: what are the key features to describe compound objects in the relevant repositories? Are there any compound objects available yet? How are they represented internally and externally?

Characteristics of the repositories investigated

Fedora

Fedora -Flexible Extensible Digital Object Repository Architecture- is a robust, free, open-source digital content repository that provides modules and interfaces for creation, management and dissemination in addition to a wide variety of uses concerning digital content. Furthermore, Fedora was designed with the concept of compound digital objects in mind.

The Fedora architecture includes a generic RDF-based relationship model that not only represents hierarchical relationships among objects and their constituents like the part/whole relationship (e.g. articles in an electronic journals), but also more expressive graph-like ones.

Based on the Fedora Content Model Architecture and the Fedora Digital Object Model, compound objects can be described on three different levels:

- Each Fedora digital object forms an aggregation of content items. The Fedora Object Model includes one or more elements known as 'datastream'[6] to represent these items. There is no restriction to the physical location of such

[6] https://wiki.duraspace.org/display/FCR30/Fedora+Digital+Object+Model#FedoraDigitalObjectModel-Datastreams

content bitstreams. They can be either stored locally within the Fedora repository, or externally by inserting just a reference within the datastream.

- Each Fedora digital object can have relationships to other objects to build a semantic object network. Generic object-to-object relationships are defined in the Fedora relationship ontology[7]. If required, these set of relationships can be enhanced and readjusted to specific needs.

- Since Fedora version 3.0, the Content Model Architecture introduces a new component known as Content Model Object[8]. It represents a formal model that describes structural, functional and semantic information of a set of digital objects. It can also be seen as a class holding information about all of its constituent objects. Therefore, it can be used as a validation reference when creating, ingesting or modifying Fedora digital objects.

A complete discussion of the Fedora Object Model is beyond the scope of this document.[9]

OPUS v2 and 3

The “Online Publikationsverbund der Universität Stuttgart” (OPUS) is a repository software application originally developed by the University Library Stuttgart[10]. With about 100 installations this system is well established and the most frequently used repository system at German institutions.

In the following, versions 2 and 3 of the software are considered jointly, since the versions do not differ fundamentally from a conceptual and architectural point of view. Opus v4[11], which is a complete rewrite of the software, is not considered in this report because no final release is available at the time of writing.

OPUS is written in PHP and stores its bibliographic metadata in a database system like MySQL or mSQL. Bitstreams are stored in a local file system but are not referenced in the database.

At minimum, OPUS repositories expose Dublin Core metadata and support the XEpicur data transfer format for persistent identifiers. Additional format support exists for ProPrint - a format for print on demand services, XMetaDiss[12] to describe

[7] http://www.fedora-commons.org/definitions/1/0/fedora-relsext-ontology.rdfs
[8] https://wiki.duraspace.org/display/FCR30/Fedora+Digital+Object+Model#FedoraDigitalObjectModel-ContentModelObject
[9] More details can be found on https://wiki.duraspace.org/display/FCR30/Fedora+Repository+3.3+Documentation . The Fedora Digital Object Model is described in XML Schema and can be found at http://fedora-commons.org/definitions/1/0/foxml1-0.xsd
[10] http://elib.uni-stuttgart.de/opus/doku/about.php
[11] http://www.ub.uni-stuttgart.de/wirueberuns/projekte/opus4/
[12] http://www.d-nb.de/standards/xmetadiss/xmetadiss.htm

dissertations. XMetaDissPlus[13] supports further publication types by a controlled vocabulary.

Furthermore by assigning DDC numbers[14] to publications the scientific topics are roughly classified. OPUS supports the Subject Headings Authority File (SWD) providing a normed, terminologically controlled vocabulary as well.

The internal database scheme comprises a number of tables to store information about a record. With regard to the eco4r demonstrator the following tables are of particular interest:

- table_opus – identifies a record with an opus-Ident number and stores bibliographic metadata
- table_format – defines file extensions and mime-types permitted in an OPUS instance
- table_bereich_de – declares the path to the full-text document in the local file-system and stores the URL of the full-text document
- table_opus_diss – stores metadata specific to the dissertation publication type
- table_opus_coll - holds relations of n documents to m collections

Bitstreams are stored following the scheme <file-system-path>/<year>/<opus-id>/<file-extension>/<bitstream>. This means it is possible to store any number of bitstreams associated with a record item. They can be distinguished by their representation format (html, pdf) but the database scheme does not allow to model compound objects. Nor is it possible to describe the inner relations of those bitstreams, thus a description format for compound objects is lacking.

How are publication items represented in the internal data model of the partner repositories?

DiPP is a software system for publishing Open Access e-journals based on two different open source components. Whereas the Plone content management systems provides editors, authors and researchers with suitable tools to manage, submit and access journal content, the fedora system is used for storage and information management.

[13] http://www.d-nb.de/standards/xmetadissplus/xmetadissplus.htm
[14] only very few DDC numbers are used based on a mapping from a formerly used DNB-Sachgruppen

Even though DiPP does not run on the latest fedora version supporting the CMA, content models are introduced on a conceptual level. The content model thereby reflects the structures typically found in any kind of journal.

- dipp:eJournal: *defines* a pure logical object representing an eJournal The objects hold all metadata relevant for a whole eJournal, i.e. the ISSN or other identifiers. It also declares all relations to the underlying structural objects representing the journal's structure or hierarchy.
  Examples are:
    - BuR – Business Research: http://www.business-research.org/
    - brains, minds and media: http://www.brains-minds-media.org
    - zeitenblicke: http://www.zeitenblicke.de/
- dipp:container: defines a logical object representing structural parts of the journal, a volume or an issue. A container can either point to other containers or allocate a subset of *articles* to an aggregation . Here are some examples:
    - A volume (05.2008) of the jvbr eJournal: http://www.jvrb.org/archiv/5.2008
    - A volume (06.2009) of the jvbr eJournal: http://www.jvrb.org/archiv/6.2009
    - VRIC 2008 Special Issue of the jvbr eJournal: http://www.jvrb.org/archiv/6.2009/vric2008
- dipp:article: Defines the logical unit of the publication object. This holds all relevant metadata and points to all the parts representing an article. An article object itself provides no datastream representing the article content.

Here are some examples:

- The journal article 'Fitting 3D morphable models using implicit representations': http://www.jvrb.org/archiv/1279
- The journal article 'The art to keep in touch: The good use of Lagrange multipliers': http://www.jvrb.org/archiv/1276
- The journal article 'The Monotonicity Puzzle: An Experimental Investigation of Incentive Structures'. Contains supplementary materials, that can be observed on the splash-page: http://www.business-research.org/2010/1/accounting/2498

    - The Journal article 'Multi-Mode Tensor Representation of Motion Data'. Contains supplementary materials, that can be observed on the splash-page: http://www.jvrb.org/archiv/1419

- dipp:data: defines objects containing all the data and bytestreams needed to represent the article content. It holds the articles text sources as well as embedded images, and all kind of supplementary materials.
- oaiSet: a model mainly required for the declaration of sets within OAI-PMH. It defines assemblages of article objects according their journal membership and their subject heading (DDC based). In theory, oaiSets are applicable to all kinds of assemblages in which articles share the same behavior expressed in metadata.

The following chart provides an overall view of the content models described above:

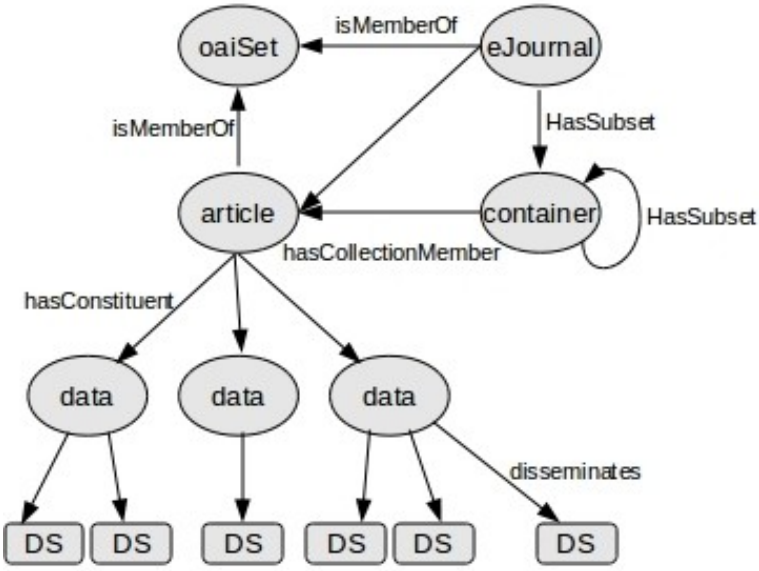


Figure 1 internal content model used in DiPP

The following chart illustrates a DiPP Journal article according to the OAI-ORE Data Model. The article is described by an ORE Resource Map and represents an Aggregation containing simple Aggregated Resources (Datastreams).

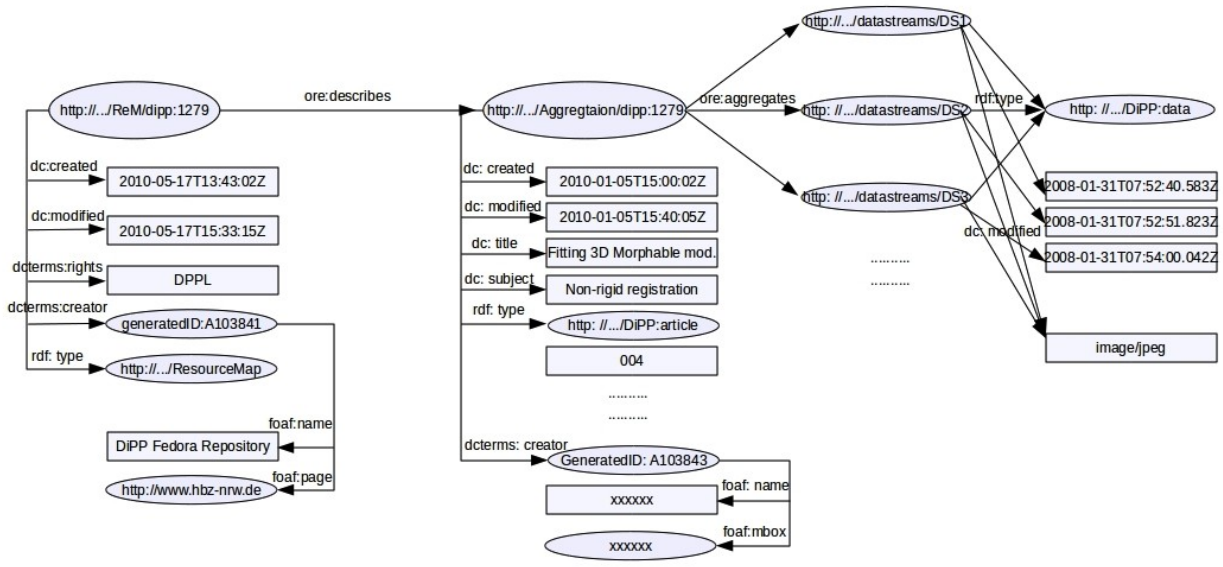


Figure 2 ORE representation of an article in DiPP

Repositories of Bielefeld University Library

Bielefeld University Library operates several instances of the OPUS software versions 2 and 3 for its publication management, namely BiPrints[15] for post-prints, BieSOn[16] for dissertations and BieColl[17] for collections like proceedings and periodicals. Other publication types are managed as well, e.g. e-journals and digital imprints, but they are not considered in this report.

A number of additions and customizations to the OPUS software were made to meet specific requirements, e.g. for collection management. The OAI-PMH module of each repository is configured to expose Dublin Core and XEpicur metadata. Items contained in the OPUS repositories are grouped by DDC and their publication type, the latter using a vocabulary defined by DINI.

Only very few samples could be found, which are compounds of multiple parts:

Dissertations in BieSOn

1. Faculty of Clinical Linguistics – the record with bieson-opus-id ‘1503’ and URN ‘urn:nbn:de:hbz:361-15037’ resolves to the splash-page at the URL ‘http://bieson.ub.uni-bielefeld.de/volltexte/2009/1503/’ . The record has three pdf files assigned:

[15] BiPrints - publication server for primary publications and additional deposit: http://repositories.ub.uni-bielefeld.de/biprints/ , containing 3899 records (31st May 2010)
[16] BieSOn - Bielefelder Server für Online-Publikationen - http://repositories.ub.uni-bielefeld.de/bieson/ , containing 1250 records (31st May 2010)
[17] BieColl - Bielefeld eCollections - http://repositories.ub.uni-bielefeld.de/biecoll/ , containing 414 records (31st May 2010)

- the full-text
- the index of annex
- scans

2. Bielefeld Graduate School of Economics and Management – the record with bison-opus-id ‘1518’ and URN ‘urn:nbn:de:hbz:361-15189’ resolves to the splash-page at the URL ‘http://bieson.ub.uni-bielefeld.de/volltexte/2009/1518/’. The splash-page contains a link referring to another HTML-page, which contains links and descriptions of the bitstreams :

    - dissertation full-text as pdf file
    - annex with source-code in a zip file

Post-Prints in BiPrints

1. journal article originally published at Springer – the record with biprints-opus-id ‘4229’ has two locations

    a. at the journal ‘Marine Biotechnology’ website with DOI ‘10.1007/s10126-009-9208-z’

    b. at the repository with URN ‘urn:nbn:de:0070-bipr-42297’
    two files are assigned to the record:

        i. the full-text in a pdf-file

        ii. supplementary material in a doc-file

    The splash-page refers to the location at Springer using a DOI, however the DOI is not contained in the OAI-DC metadata.

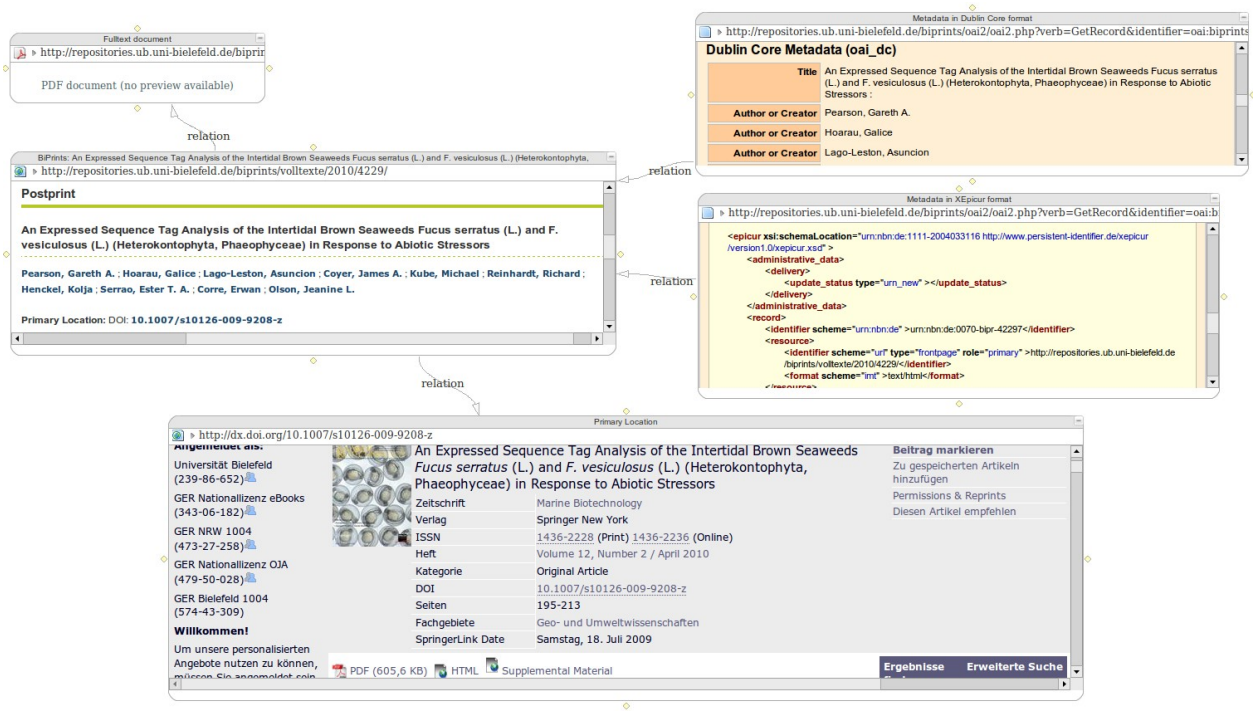


Figure 3 relations between different components of a post-print record in BiPrints

Collection in BieColl

BieColl contains collections from departments of Bielefeld University, which are

- conference proceedings of the faculty of technology
- periodicals archive
- proceedings of the “Bielefeld Conference”

Grouping of collections is done by the OAI setSpec information. [18] Unfortunately it is not possible to address publications of a specific conference or journal etc.

- book section
- conference or workshop item
- journal article
- and by DDC number

Depending on the publication type the items inhere in different structures. E.g. a conference item from the collection ‘Bielefeld Conference’ may consist of

- an abstract

[18] based on specifications defined by DINI

- presentation slides or a poster
- audio or video recording
- a full-text paper
- a link to a publisher's conference proceeding http://www.emeraldinsight.com/journals.htm?issn=0737-8831&volume=27&issue=4&PHPSESSID=l3i12k5r0ldf2v8enb9hoamp50

In fact that granular and complete information is not contained in the repository but could be modeled in OAI-ORE.

Modeling Approach

Creating a compound object as a collection of conference papers:

Each Collection is addressable by an HTTP-Get query, e.g. 'Bielefeld Conference' with http://biecoll.ub.uni-bielefeld.de/ergebnis.php?suchart=teil&Lines_Displayed=11&sort=o.date_year+DESC%2C+o.title&suchfeld1=oc.coll_id&suchwert1=50 , where the interesting parameters are *suchfeld1=oc.coll_id* and *suchwert1=50* .

Creating a compound object which is of type 'conferencePaper' and has a relation to one or more collections:

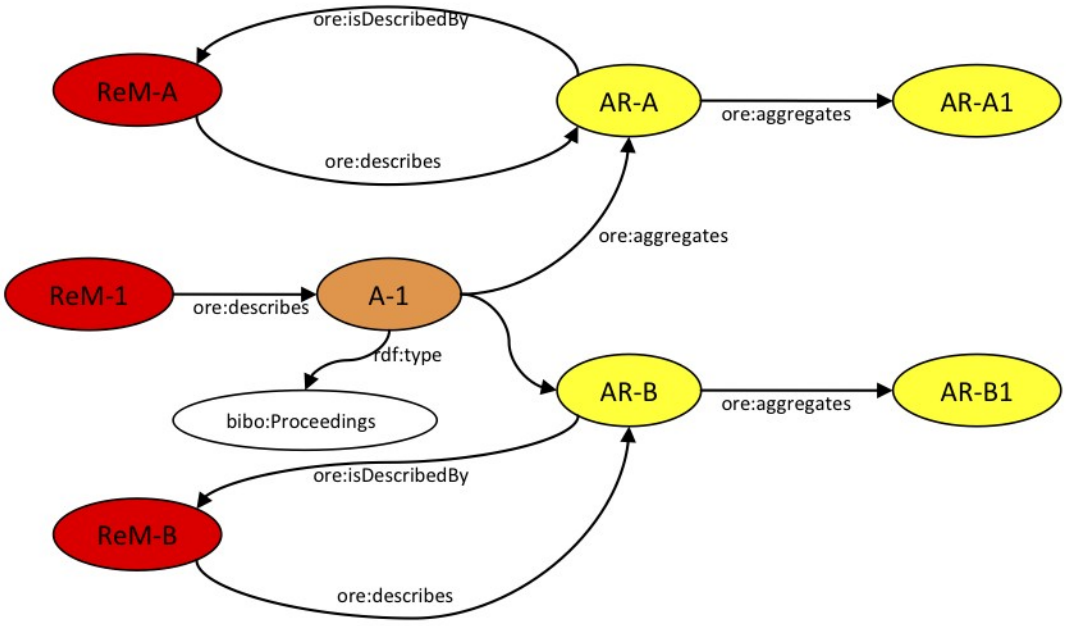


Figure 4 modeling of a conference paper as part of a proceeding

Table 2: entity description according to figure 4

ReM-1	Resource map of the collection
ReM-A	Resource map of a conference paper A, which is part of the collection
ReM-B	Resource map of a conference paper B, which is part of the collection
A-1	Aggregation of type bibo:Proceedings
AR-A	Aggregated Resource, which is also described by a resource map and has aggregated resources - the conference paper object AR-A1
AR-B	Aggregated Resource, which is also described by a resource map and has aggregated resources - the conference paper object AR-B1

Issues when creating compound object representations

Through the emergence of multi-part publications (as an electronic journal) consisting of primary data and supplementary materials, modeling publications as compound digital objects was already in the architectural design phase of the DiPP platform one of the important goals. This goal is also one of the major factors that affected the choice of Fedora as a repository system. However, at that time, even though Fedora enables the modeling of compound objects, no elaborate Content Model Architecture was available. For example, since Fedora3.3, a new entity known as Content Model Object was introduced to assign common behavior to object collections 'compound objects'. Thus, the flexibility of the DiPP data model was limited.

The OPUS data model was originally designed to handle single files (e.g. a PDF file representing a dissertation). Since multi-part scientific publications have emerged, requirements for a new architecture enabling the modeling of such new sorts of publications have been claimed. Until now, no appropriate adaptations to the initial OPUS data model have been performed.

In addition to the technical deficits of our live repositories related to compound objects, adaptations on the business logic level (ingest workflows, data management) have not been covered sufficiently.

According to the difficulties mentioned above, the process of constructing compound objects from our live systems led to some challenges which will be represented hereafter.

1- Incompleteness and Lack of Granular Metadata

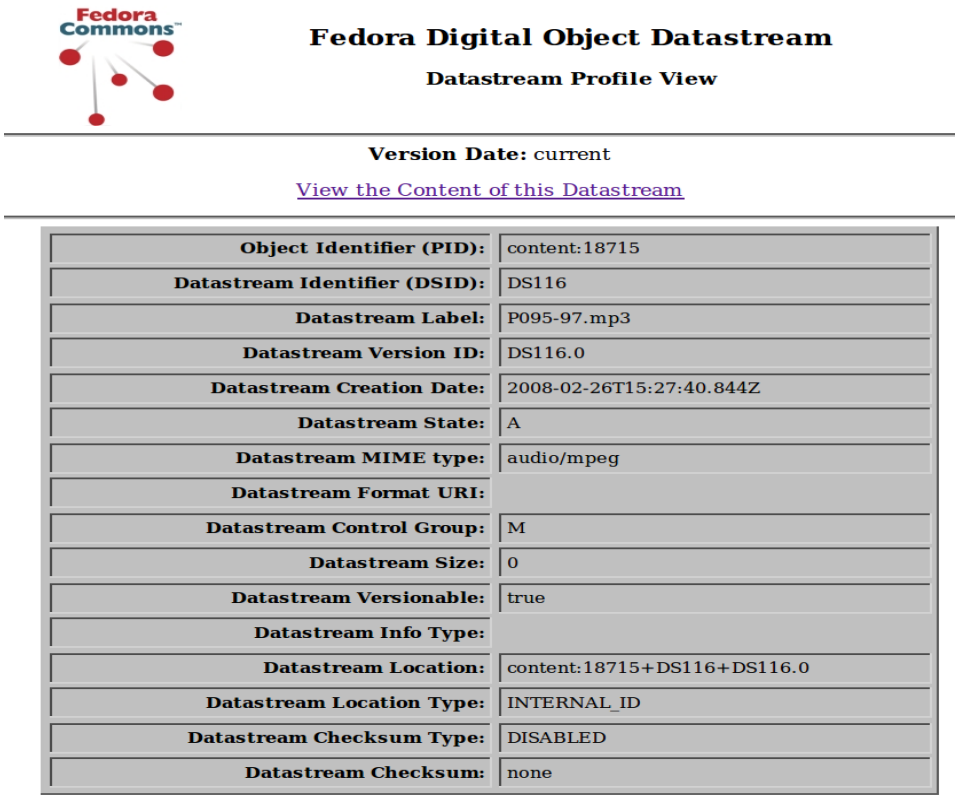
Publications derived from repository systems operated by *hbz* and *UniBi* offer metadata relying on approved standards like the German DINI-Certificate. Other widely known standards like OAI-PMH have been adopted. However, the exposure of metadata information concerning compound objects requires some special considerations.

In order to enable the reuse and long term preservation of compound objects, it is essential that their individual components and the semantics governing them are adequately described.

The analysis of the data stored in our live repositories has shown that there is a lack of granular descriptive and technical metadata. To provide a concrete example, let us consider a journal article stored in the DiPP repository.

The article with the title '*Beja Pedagogical Grammar*' (http://www.afrikanistik-online.de/archiv/2008/1283/section_d#N35F73) from the electronic journal '*Afrikanistik online*' contains many .mp3 files to complement a text about articulation particularities of a special African language. In terms of compound objects, the journal article can be considered an aggregation and the .mp3 files aggregated resources. Whereas the article (in this case the compound object) document is appropriately described with metadata (title, subject, author, contributor, URN, URL, etc.), only a limited set of metadata can be derived for the single .mp3 files (Aggregated Resources). The following snapshot from the Fedora web interface shows the metadata available for a datastream holding an .mp3 file.

There is no metadata explicitly associated with the .mp3 file, but some information can be derived from the enclosing datastream (label, id, mimeType).



This example is representative for single aggregated resources stored in the repository. The requirement that every entity within a compound object should be considered as a standalone resource which may be potentially reused in other environments was not taken into account during the modeling process of the underlying applications.

The fact that Aggregated Resources holding just a limited set of metadata information are tightly bounded to the enclosing Aggregation implies that they cannot be reused in other environments. This has an impact on the flexibility of building new compound objects from the starting point of objects existing already.

Another problem aligned with the incompleteness of metadata is the obscure 'role' of the granular components in respect of the pertaining compound objects. Accordingly, considering only the repository, it is not possible to differentiate between primary and supplementary materials of a journal article, nor to determine the main work of a multi-part publication. This may in contrary be clearly visible on the e-journal's splash-page. This may result in splash-pages (which in most cases provide a non-machine processable representation of just some selected information from the repository) containing important information that do not exist in the metadata.

The lack of granular metadata also implicates some challenges with respect to the preservation of compound objects. Especially, the problems become apparent when aggregating content published from different authors and/or stored in different repositories. As in the case of an overlay journal, in a worst case situation, each single component part may arise from a different author and may be located in a different repository. Therefore, each single entity may be covered by different policies, which may restrict its usability. A not so dramatic but similar situation has been observed by journal articles containing additional supplementary materials. In addition to the lack of provenance and authorship information, no copyright statement could be gathered fore those objects. The reuse of this individual supplementary materials is therefore strongly limited. Moreover, metadata like authorship, rights and provenance are vital for long term preservation, and compound objects consisting of component units holding no such information are considered as unready for delivery to preservation systems.

2- Lack of Semantics Describing the Composite Elements

Observations during the construction of compound objects from our live repositories have also shown that the publication type information (e.g. book, article, eJournal...) is not expressive enough to enable external machine-processing and inference. The value of the type-metadata consists of 'just' a URI, without any reference to elements responsible for the class- and relationship-description as we know them in semantic web ontologies. An ontology is an appropriate tool in the hands of an

author to provide a machine-processable meaning of its documents. It represents an authentic machine-processable description of the data and the relationships dominating this data. This information can be of big interest to an LTP-system, since not only the digital objects along with their structural and technical metadata are relevant but also an author-authentic meaning of the data.

3- Identifier

OAI-ORE specifies the use of protocol-based URIs to identify any entity in its data model as described above. The idea behind it is to make both the aggregations and their constituents accessible as parts of the web architecture and thus each object (compound or atomic) can be referenced uniquely and globally.

From the long-term preservation perspective, not only the integrity of digital objects is of interest when dealing with compound objects, but also the coherence between the constituent units and their persistent identification [Doorenbosch, P. & Sierman, B. (2010)]. A global persistent identification of the resources guarantees a durable validity of the relationships by providing independence against local changes (storage location or removal) in the different repositories.

In a Fedora repository and thus in the DiPP-system, each entity in the Content Model Architecture is identified by a locally unique identifier, which may be user-defined or automatically assigned by Fedora. In addition, every object has an implicit URI associated with it. This has many advantages such as guaranteeing the uniqueness of the references and the compatibility with semantic web technologies. Fedora provides also web services (which make use of this URIs) to access and manage the stored objects.

In terms of the persistence problem, objects of the content codel entities dipp:*eJournal* and dipp:*article* include a URN:NBN information, which is managed by a resolver service for persistent identifiers provided by the German National Library (DNB). These references are indeed persistent but point only to a representation (a website) and not to the compound object itself (or even to the article main file). Because compound objects can have more than one representation (PDF, html, image...) this identifier cannot be used as a reference for the aggregation as a whole. This can be observed in the following example of an article published in the Business Research e-journal. The URN 'urn:nbn:de:0009-20-24985' resolves to the website with URL ' http://www.business-research.org/2010/1/accounting/2498' which is just a representation of an article. The main text and the supplementary materials can in fact be seen on the website, but there is no reference to a machine-processable representation illustrating the article as a compound object.[19]

[19] A result of a study made by the Driver-II project [cite: inventory study into the present ….] concerning the use of persistent identifier in the EU digital repository-landscape, shows that 75% of the repositories assign PIDs to each document. 15% do not use such identifier and the rest answered this question with 'don't know'. It was also mentioned that the high percentage of repositories using persistent IDs may only be caused by a misinterpretation by

The BiPrints example highlights the need to fully describe identifiers in the metadata of related deposits of a publication, that of the repository and publisher location.

If a digital object comprises multiple files linked in a splash-page, it is necessary to identify the main item, which is usually the full-text. In those cases a single occurrence of the dc:format element has a very limited expressiveness about associated files.

Best Practices for the construction and processing of compound objects

Introduction

In this chapter we will provide recommendations and suggestions in the matter of constructing compound object representations from the starting point of yet existing content in repositories. These recommendations are the result of a content analysis performed during the first work package of the eco4r project. Thereby, the goal was to identify difficulties and challenges encountered during the process of constructing compound objects from existing publication systems based on Fedora and OPUS in the involved institutions. Furthermore, the results reached by two related projects (the eTheses project and the DRIVER-II project) were analyzed as well, in order to compare the experiences.

Recommendations

Identifier

Given that we aim to expose our compound objects in form of OAI-ORE Resource Maps and to promote the Semantic Web compatibility, it is advisable that both the constructed compound objects and their constituents are identified by a globally unique URI. Since the identifier of a compound object should reference the whole aggregation and not just a component part or a special representation, it should not coincide with the URL of the splash page or the bitstream which represents the main work.

In the digital long term preservation, the use of stable (against local changes), worldwide unique identifier is a vital condition as stated in the OAIS Model. Such an identifier enables an authentic referencing and assures a reliable assignment of metadata to the digital object. For this purpose, common internet addresses like an URL are not convenient since they usually change over time.

---

the asked persons of the term.

Furthermore the DINI initiative (http://www.dini.de/ueber-dini/) also recommends to generate one persistent identifier for each object version (manifestation) whenever the content of the work has changed.

In the context of compound objects, the use of persistent identifier enables an authentic preservation of the aggregation structure, since the individual resources are accessible over time and therefore the relationships among them remains valid.

Therefore, we recommend the use of standardized and well established persistent identification schemes for both the aggregation and every aggregated resource. A list of the most used persistent identification systems, specifications and standards can be found on nestor handbook Version 2.0 chapter 9.4 [Neuroth, Oßwald, Scheffel, Strathmann, Jehn].

Properties and Metadata

In the section 'Current Situation' we already mentioned the lack of granular metadata and demonstrated the rigidity aspect caused accordingly in a repository. We observed in many cases that in multi-part publications, only the resources on the top of the hierarchy (like the main work of a dissertation) are fairly described with metadata. In contrary, resources that are considered as complementary or supplementary are only marginally described. This complicates the task of reusing such resources in other environments. The component units are therefore strongly tied to the comprising aggregation.

The DRIVER guidelines [Verhaar.] referring to Enhanced Publications and the Knowledge Exchange Deliverable 6 [Ruijgrok, Slabbertje, van Luijt, Awre, Rammer Nielsen] as well as those concerning the modeling of enhanced e-theses, recommend a set of key metadata. They are listed in the table hereafter.

Table 3: recommended metadata information from DRIVER and KE

Key Information	Enhanced Publications (DRIVER)	Enhaced E-Theses (KE)
Title	E-Prints can have a title	For every aggregated resource
Author	On all levels of the enhanced publication	For every aggregated resource
Last modification date	For enhanced publications as a whole	For every aggregated resource

Semantic Type	For each aggregated resource	For every aggregated resource
Description	For atomic or compound datasets	------
Technical Format	For every aggregated resource	------
Mime Type	For every aggregated resource	------
Provenance	On all levels of the enhanced publication	------

In addition to structural information and technical metadata that can be generated automatically by using tools like JHOVE2 or DROID, we also recommend the following minimum set of metadata information to deal with the issues of re-usability and preservation:

authorship:
The compound object as a whole should hold a reference to the authorship-information. This is of big significance especially when nesting aggregations coming from different sources to build a 'super' compound object. An Overlay Journal is an appropriate example of such a 'super' collection, in which one or more resources (atomic or compound) are collected and represented according to specific features. Another example is a dissertation including references of other scientific publications. In these cases the authorship information augments the reliability of the resulting nested aggregations. This represents also mandatory information for OAI-ORE Resource Maps. In the case of aggregated resources, the explicit need of the authorship information differs from case to case. The authorship information may sometimes be implied automatically from the comprising aggregation. An example can be a one-author dissertation deposed in several parts according to its chapters in the repository. In this case, each chapter represents an aggregated resource and the authorship information can be deduced from the context information. In cases concerning multi-author publications, the need for the author-information is much more obvious.

last modification date:
Another property also stated as mandatory in the ORE data model, is the point of time at which the aggregation was most recently modified. This information is of big importance when aiming to trace the modifications of a resource for versioning purposes which is also considerable for long term preservation.

Classification as Minimum Indication of the main publication subject

The domain of compound objects does not only consists of the question of how to model and represent collections of logically related resources. The issue of how to find related resources, and which properties a retrievable resource should exhibit, is of big importance as well. Ones an appropriate data model and a reliable architecture have been established, an application representing a service like an Overlay Journal may start automatically retrieving resources from different locations to build a compound object according to some pre-specified criteria. A crucial condition in this scenario is, that a clear indication of specific features and properties of the individual resources should be exposed to a harvesting application (e.g. Overlay Journal application). An example of such a specific property is the main subject providing a short but precise information of the resource content. A widely adopted method for this purpose is the use of library classification schemes. They typically provide a well-defined hierarchical and logical system organized in classes (representing topics) and subdivisions of classes. The user are then requested to include manually this information in the publication metadata.

In addition, more complex methods originating from the 'Data Mining' and 'Machine Learning' domains, like the unsupervised clustering, can be adopted to categorize resources. But, despite the fact that such methods may provide a full-automatic categorization, they are coupled with difficulties in implementation, configuration and management, and do not always guarantee an error-free classification.

Therefore, every resource (compound or atomic) should exhibit information about the main subject of the underlying document(s). For this purpose, we recommend to provide the classification information as a minimum property. Furthermore, we recommend the adoption of widely used universal classification schemes to augment interoperability. Examples are the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC) and the classification scheme devised by the Library of Congress (LCC).

Semantic Type:

The publication type information about an aggregation and its aggregated resources can play an important role in identification and discovery purposes. Since the statement of an object-type as a simple string can lead to misinterpretation in external applications, we recommend that the type should be an URI pointing to an item in an appropriate and accessible ontology.

description:

In addition to the semantic type-information as described above, which is also a good practice when aiming to augment interoperability, providing a human readable description enables better documentation and is hence recommended. It provides generic information about the aggregation, e.g. what the aggregation is representing and which application methods are related with it.

Legal information

Information about copyright and policies is of crucial importance when dealing with compound data objects. The composite nature of this objects and the fact that the component parts of an aggregation may not necessarily have to be located in a single repository, may lead to some difficulties in legal issues. Since the author of the aggregation is not necessarily the author of the aggregated resources, agents aiming to build a new aggregation should adhere for the legal information of the single components before processing the aggregation.

Relationships

The content analysis phase has demonstrated the crucial importance of the use of relationships between the data objects stored in repositories. This fact has become obvious as we tried to construct compound objects from the live repositories Fedora3.3 and OPUS. Whereas Fedora takes advantage of a well-defined data model that enables the use of semantic relationships among the different data objects, OPUS suffers from such possibilities to explicitly express semantic relations. Rather publications and additional data resources are informally modeled in OPUS without semantics, which prohibits its automatic processing. This enables in some way to express the concept of an aggregation but disregards completely the relationships between the component parts of this aggregation. Furthermore, it is nearly impossible to model more expressive relationships as the simple containment one, like sequential, lineage, versioning, manifestation and bibliographic relationships as described in [Verhaar].

Hence, we strongly recommend not only to focus mostly on storing individual objects in a repository and enriching them with metadata, but also to recognize the existence of relationships between various resources. These relationships should be stated explicitly using typed relationships which describes the logical connection between the data objects.

We recommend also using RDF to describe the resources within a repository. Furthermore the use of semantic ontologies organizing this content and representing a set of concepts (classes) within a domain and the relationships between those domains should be taken into account. So, the data and relationships can be represented in a machine readable form and according to that can serve as input for applications using inference engines to extract new knowledge that has not been stated explicitly. In addition, the use of RDF can have a positive impact on LTP since automated systems are able to trace the structure of any items [Woutersen-Windhouwer, Brandsma, Hogenaar,Hoogerwerf, Doorenbosch, Dürr, Ludwig, Schmidt, Sierman (2009)].

A Feasibility Study - Concept of an Overlay Journal

The eco4r project aims to implement and test the recommendations for constructing and re-using compound objects. This will be done by prototyping an overlay journal of online dissertations.

An overlay journal does not publish any original articles. It rather selects and aggregates articles already published. As an example this selection can be done by specific subjects, scientific disciplines etc.

Concept

In case of the eco4r demonstrator the overlay journal will aggregate electronic dissertations with optional supplementary materials or references to related scientific web resources from live institutional repositories. Based on the subject classification new aggregations will be created. They link dissertations like issues in a magazine.

Following this approach nested aggregations need to be modeled using OAI-ORE. Furthermore the lineage or provenance of the dissertation items is important information to be described.

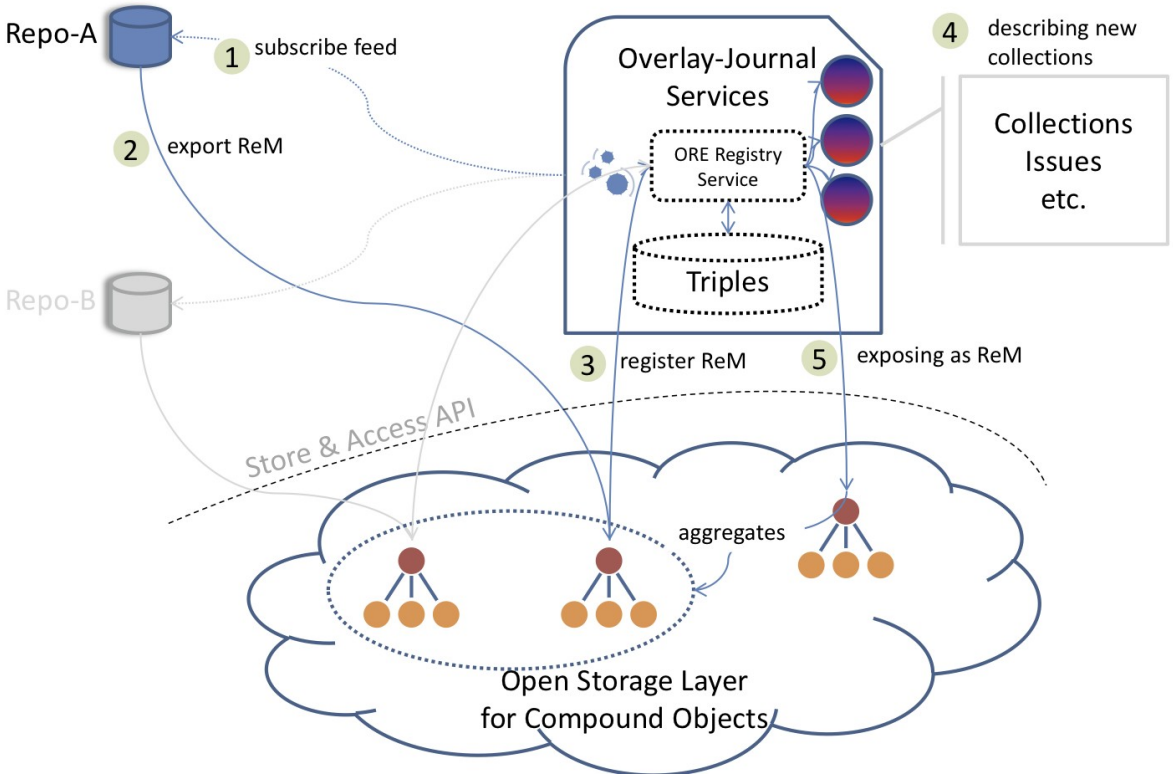


Figure 5 Overlay Journal concept based on Compound Objects

Workflow

1. The overlay journal subscribes to an atom feed provided by a repository.

2. The repository publishes a new dissertation, exposes the resource map that describes the logical boundary of the dissertation and updates the atom feed with a new entry.

3. The overlay journal recognizes the new Resource Map and retrieves, registers, analyzes it.

4. Now the overlay journal can either announce a new issue about a topic or add the newly published dissertation to already existing collections.

5. The newly created or updated aggregation is exposed as a ReM.

Architectural Approach

Tarrant et al. propose to adopt the OAI-ORE data model instead of relying on the repositories' internal data models . This implies also to move the storage layer of a repository to an external layer, e.g. a cloud. This allows using the potential of OAI-ORE most effectively, since the resource maps do not need to export from one repository to import into another repository but can be accessed and reused in a transparent manner.

Furthermore this strategy should also ease long term preservation strategies, since the data model of the compound objects is transparent and accessible in RDF as well.

The functionalities of a repository are then focused on specific services, like the selection of dissertations in an overlay journal.

References

Beedham, Missen, Palmer, Ruusalepp. Assessment of UKDA and TNA compliance with OAIS and METS standards. http://ie-repository.jisc.ac.uk/51/1/Report_oaismets.pdf

Brooking, Shouldice, Robin, Kobe, Martin, Hunter. Comparing METS and OAI-ORE for Encapsulating Scientific Data. http://www.itee.uq.edu.au/~eresearch/papers/2009/IEEEeScience2009.pdf

Dappert, A. & Enders, M (2008). Using METS, PREMIS and MODS for Archiving eJournals. D-Lib Magazine, 14/9-10, http://dlib.org/dlib/september/08/dappert/09/dappert.html

DINI (2010). DINI-Zertifikat Dokumenten- und Publikationsservice 2010. http://nbn-resolving.de/urn:nbn:de:kobv:11-100109986

DINI (2010). Gemeinsames Vokabular für Publikations- und Dokumenttypen. http://nbn-resolving.de/urn:nbn:de:kobv:11-100109998

Doorenbosch, P. & Sierman, B. (2010): Institutional Repositories, Long Term Preservation and the changing nature of Scholary Publications. - Open repositories Conference, Madrid, July, 2010, http://or2010.fecyt.es/Resources/documentos/GSabstracts/InstitutionalRepositories_LongTermPreservation_etc.pdf

DRIVER (2008). DRIVER Guidelines 2.0 – Guidelines for content providers – Exposing textual resources with OAI-PMH . http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf

McDonough, Jerome P.. Aligning METS with the OAI-ORE data model. http://www.ideals.illinois.edu/bitstream/handle/2142/10744/mcdonough.jcdlpreprint.doc.pdf?sequence=4

METS Primer Revised 2010-4. http://www.loc.gov/standards/mets/METSPrimerRevised.pdf

OAI-ORE (2008). ORE Specifications and User Guides . http://www.openarchives.org/ore/1.0/toc

Sefton, Downing (2010): ICE-Theorem – End to end semantically aware eResearch infrastructure for theses. http://journals.tdl.org/jodi/article/view/754

Tarrant, O'Steen, Brody, Hitchcock, Jefferies, Carr (2009). Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications. http://journal.code4lib.org/articles/1062

Verhaar. Enhanced Publications: Object Models and Functionalities. http://www.driver-repository.eu/component/option,com_jdownloads/Itemid,83/task,view.download/cid,54/

Woutersen-Windhouwer, Brandsma, Hogenaar,Hoogerwerf, Doorenbosch, Dürr, Ludwig, Schmidt, Sierman (2009). Enhanced Publications : Linking Publications and Research Data in Digital Repositories. http://dare.uva.nl/aup/nl/record/316849

Ruijgrok, Slabbertje, van Luijt, Awre, Rammer Nielsen (2009). A candidate Semantic Representation for enhance e-theses: Guidelines for modeling enhanced e-theses

Neuroth, Oßwald, Scheffel, Strathmann, Jehn (2009), nestor Handbuch Version 2.0 Chapter 9.4: Persistent Identifier: ein Überblick