Enhanced E-theses Project

Deliverable 6

A candidate semantic representation for enhanced e-theses: Guidelines for modeling enhanced e-theses

Document Description

Project	
Title:	Enhanced E-theses
Start date:	1st December 2007
Funding Agency	Knowledge Exchange
Document	
Deliverable number:	D6
Deliverable title:	A candidate semantic representation for enhanced e-theses: guidelines for modelling enhanced e-theses
Actual Date of Delivery:	January 2009
Author:	Peter Ruijgrok (Utrecht University Library) Martin Slabbertje (Utrecht University Library) Martin van Luijt (Utrecht University Library)
Editor(s):	Chris Awre (University of Hull) Ene Rammer Nielsen (Roskilde University Library)
Workpackage:	WP6
Workpackage title:	Creation of a possible generic semantic representation of an ETDs
Workpackage leader:	University of Utrecht
Version/Revision:	1.0
Draft/Final:	Final
Total number of pages: (including cover)	12
File name:	Enhanced Etheses project_deliverable 6_v1.0.doc

This work is made available under a Creative Commons attribution 3.0 licence. For details please see http://creativecommons.org/licenses/by/3.0/

Introduction

Background of this document

Within the framework of the Knowledge Exchange project "Enhanced E-theses", a candidate semantic representation has been created for enhanced e-theses. Because of the international character of the project, we were able to recognise the internal differences around the concept and the various components of (enhanced) e-theses between our countries (Denmark, the Netherlands and the United Kingdom). This made it possible to make a generic semantic representation that is not specific to one university or one country which can be internationally supported. The model makes it easier to offer enhanced e-theses services that transcend institutional borders to academic scientists.
One of the conclusions of our research is that there should be no distinction between an e-thesis and an enhanced e-thesis. This is simply because an e-thesis during its lifecycle can evolve into an enhanced e-thesis by adding some resources.

The input for the model consisted of several modeled enhanced e-theses, which were delivered by the project partners in the UK, Netherlands and Denmark. Based on intensive discussion between the different project partners, a candidate semantic representation has been created.

Purpose of this document

The most important goal of this document is to be a first step in the realisation of a model that can be used for a shared method of modeling of enhanced e-theses. The advantage of a shared method of modeling of e-theses is that this will increase the interoperability of enhanced e-theses. The desired end result of this is an optimisation of access to the enhanced e-theses and a better guarantee for the sustainable archiving of enhanced e-theses.

This document serves as the basis for further refining based on discussion and best practices. In order to make this possible, the decision has been made to keep the representation very generic. This offers the possibility to create a roadmap whereby newer versions will have a higher level of refinement and less freedom.
This provides organisations that would like to work with the guidelines the opportunity to grow along with the model as it evolves.
In this document we use visualisation to illustrate the described guidelines within the context of an enhanced e-thesis. We hope this makes the model ready for implementation and not just a conceptual functional model.

Definitions

Aggregation
In this document we define an Aggregation using the definition in the OAI-ORE 1.0 vocabulary: "A set of related resources (Aggregated Resources), grouped together such that the set can be treated as a single resource. This is the entity described within the ORE interoperability framework by a Resource Map." (Lagoze, van de Sompel et al., 2008)

Aggregated Resource
In this document we define an Aggregated Resource as stated in the OAI-ORE 1.0 vocabulary: "A resource which is included in an Aggregation. Note that asserting that a resource is a member of the class of Aggregated Resources does not imply anything other than that it is aggregated by at least one Aggregation. As such, this class is mostly informative and there is no need to assert that aggregated resources are instances of the ore:AggregatedResource class." (Lagoze, van de Sompel et al., 2008)

E -thesis
In this document we define an e-thesis as a publication in electronic form published in order to receive a doctoral degree.

Enhanced e-thesis
In this document we define an enhanced e-thesis as an e-thesis with zero or more resources added (the enhancements). Conceptually we do not prohibit any kind of resource to be included except that the resources should be identified by a URI. This means that even non-digital, real-world objects could be part of an enhanced e-theses as long as they could be identified or described by a URI.

Expression
In this document we define an Expression as stated in Functional Requirements for Bibliographic Records (IFLA, 1998): "the specific intellectual or artistic form that a work takes each time it is 'realized.'" "the intellectual or artistic realization of a *work* in the form of alpha-numeric, musical, or choreographic notation, sound, image, object, movement, etc., or any combination of such forms."

Manifestation
In this document we define a Manifestation as stated in Functional Requirements for Bibliographic Records (IFLA, 1998): "the physical embodiment of an *expression* of a *work*." "As an entity, *manifestation* represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form."

Resource Map

In this document we define a Resource Map as stated in the OAI-ORE 1.0 vocabulary: “A description of an Aggregation according to the OAI-ORE data model. Resource Maps are RDF graphs, and are serialised to a machine readable format according to the implementation guidelines.” (Lagoze, van de Sompel et al., 2008)

Restrictions

Versions

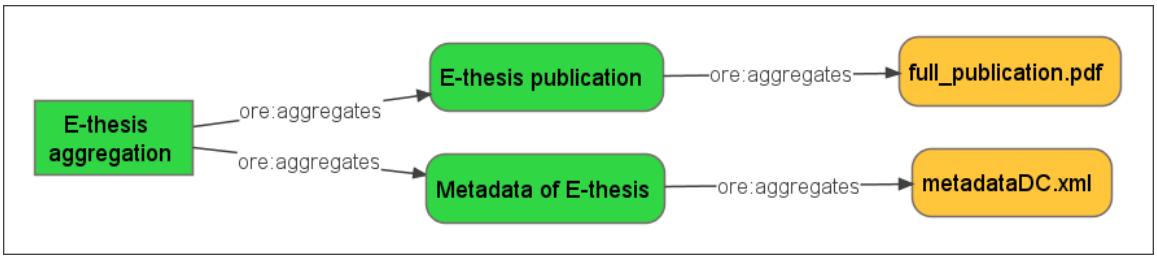
In this document we have simplified reality by omitting versioning of the e-theses or components thereof.

Guidelines for modeling enhanced e-theses

1. Level of breakdown

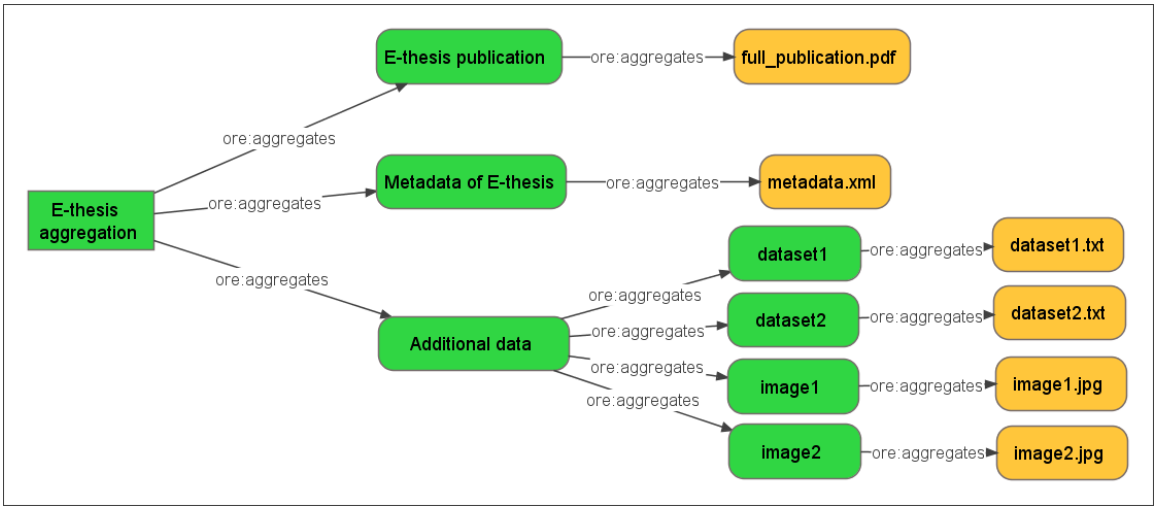
Different levels of modeling enhanced e-theses can be distinguished.

First level, required: The minimum level required for an enhanced e-thesis consists of one file for the text of the publication and one file for its metadata in DC format. Please note that at this level there is no difference between an e-thesis and an enhanced e-thesis!

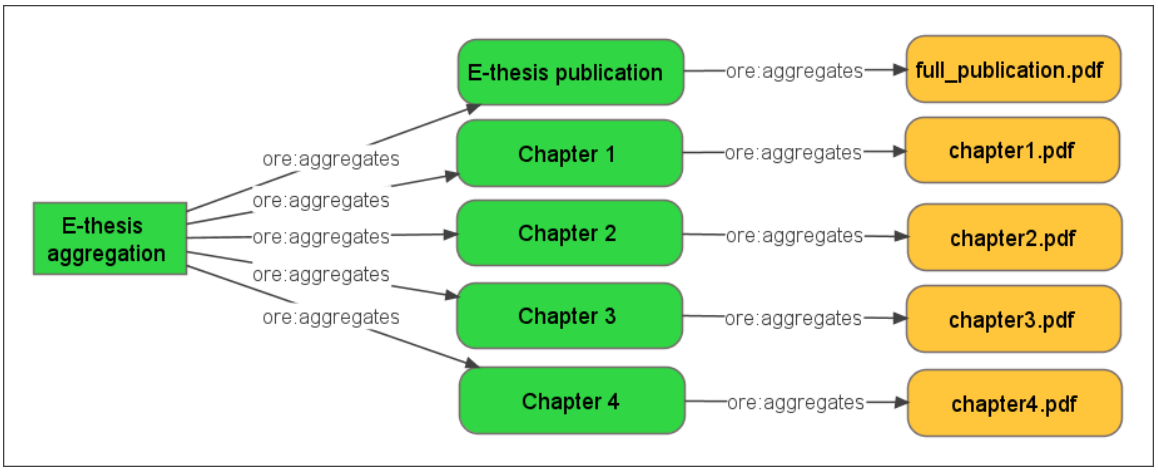


A specific detail in the model is the mandatory aggregation-resource for each file. This creates flexibility towards the future. We anticipate that over time other manifestations will be added. By keeping the model open for additions like adding the original MS/Word file for full_publication.pdf or a future situation where the 'pdf' format becomes obsolete and the need arises to add a file with the successor of Abobe PDF format. This must be possible without changing the identifier of the resource, in this case the aggregation "E-thesis publication".

Next level, recommended: We recommend that additional data is part of the enhanced e-thesis as well. Therefore an enhanced e-thesis consists of one file for the text of the publication (in one specific compound), one file for its metadata and one or more files for additional data (in one specific compound).



Higher levels, optionally: If needed this model can be expanded. Currently we have requirements for the following compounds:
Different chapters/parts of the text of the publication in addition to the mandatory full publication resource

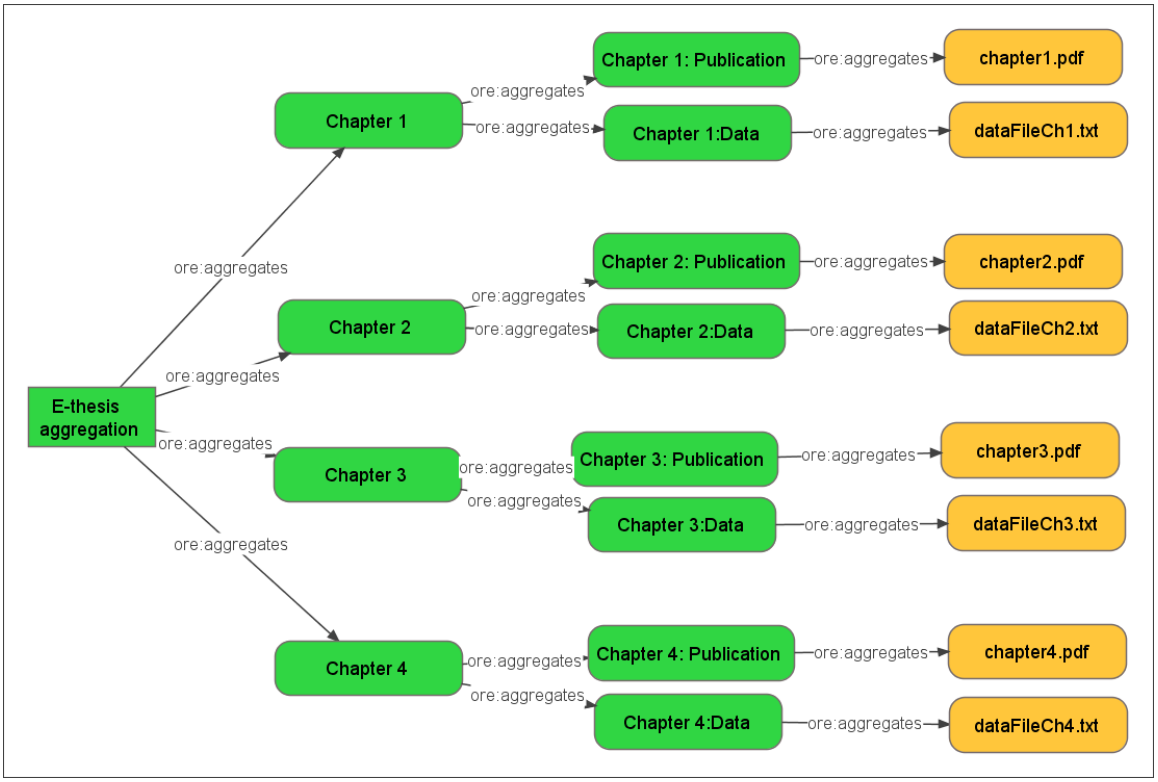


In this situation the content of an aggregation must be recognisable for its consumers. Especially in the metadata for the E-thesis aggregation it should be clear which aggregated resource contains the full-text.

Adding aggregated datasets to the aggregation

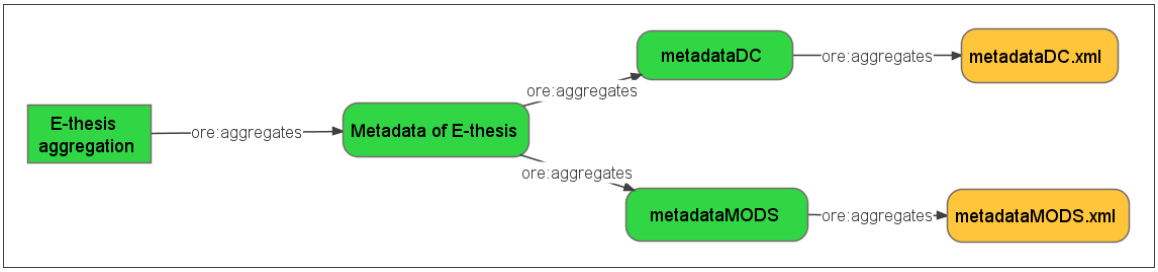
In this example the E-theses aggregation includes external datasets. We show an aggregation ('Additional data') that aggregates all datasets, but omitting that aggregation and aggregating the datasets ('Dataset 1' and 'Dataset 2') into the E-thesis aggregation is valid too.

Adding data-parts in the context of chapters of the publication



In this case each chapter had associated data that is relevant to that chapter only.

Multiple metadata formats



Some repositories offer metadata in a number of different formats, for instance to maximize interoperability while maintaining a high level of detail for clients that can handle this level of detail.

Pulling it all together the following model could be the end result:

It is advised that in deciding which aggregations should be created for an enhanced e-thesis one should focus on the justification of specific aggregations.

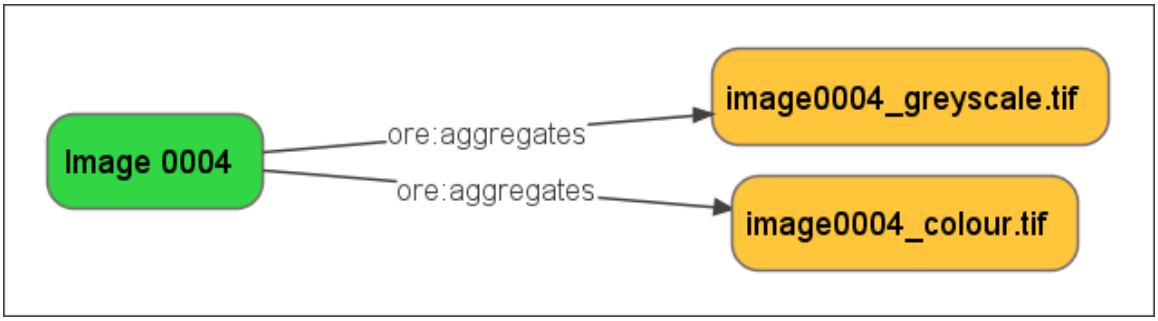
Our motto is: simpler is better!

2. Characterization of Aggregations and Aggregated Resources

We strongly recommend that each Expression is identified as an Aggregation and each Manifestation of an Expression is accessed as an Aggregated Resource using (the persistent identifier of) it's Aggregation.

The reason for this is to offer flexibility towards the future. We anticipate that other manifestations will be added, for instance adding the image in JPEG2000 format for digital preservation purposes. This must be possible without changing the identifier of the resource, in this case the aggregation "Image 0004"

Example:



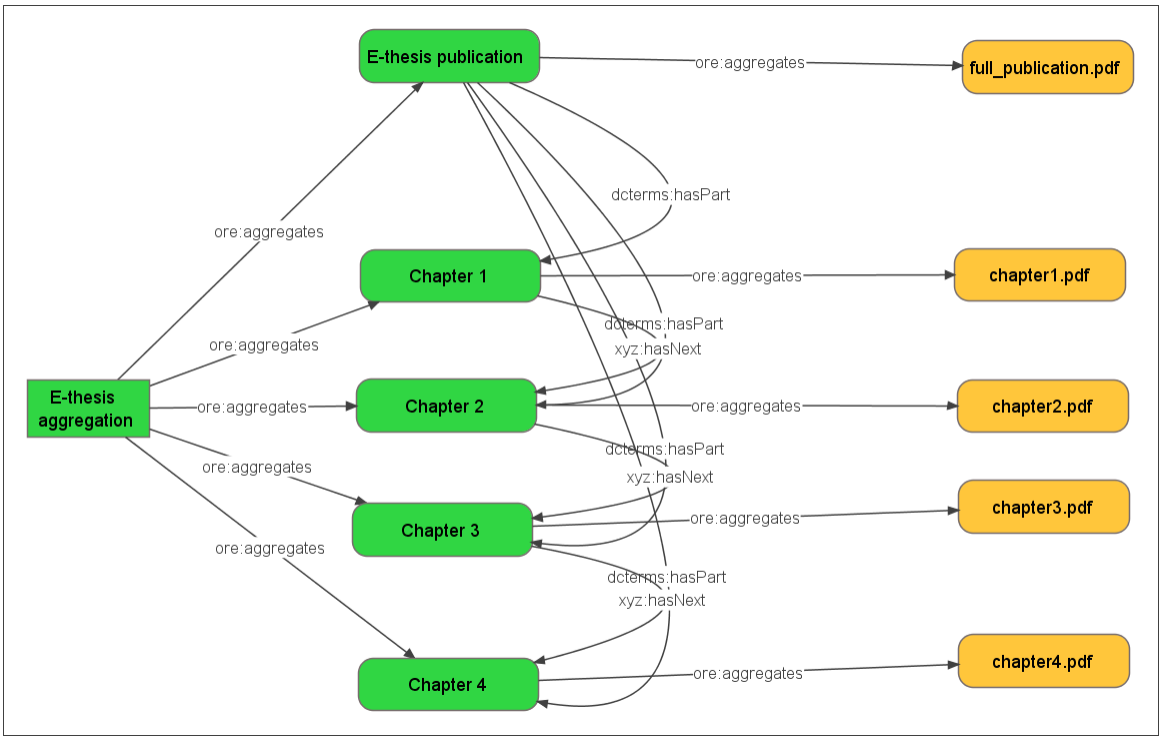
3. Identifiers

Since relationships exist between aggregations, each aggregation must have an identifier that guarantees the association with the aggregated resource over time to ensure that relationships 'stay alive'. Please note that ORE does not require the use of future-proof (persistent) identifiers!
Aggregated Resources (which are the actual bitstreams, according to Guideline 2) may have a persistent identifier. This is helpful in tracing objects whose URIs have become invalid, something which is not addressed by the ORE standard (and rightly so!).

4. Semantic relationships

Building on Guideline 2, we recommend that semantic relationships between resources should always be on Aggregation-level (between Aggregations or between Proxies) and not on the level of aggregated resources (bitstreams). This keeps the relationships valid as newer manifestations are added to the aggregation. Only semantic relations regarding technical dependencies between resources are allowed on Aggregated Resource-level (e.g. isDerivedFrom or hasDerivate).

Example:



5. Mandatory metadata of an Aggregated Resource in a Resource Map

Each web resource must have minimal metadata recorded in its Resource Map. It is recommended that each Aggregation records the following minimal metadata in its Resource Map regarding a specific Aggregated Resource:

- title
- author
- dateModified
- semanticType

6. Metadata as Aggregated Resources

It is recommended that metadata for the publication contained in the e-thesis is available as an aggregated resource (for example, as an XML file containing a MODS or DC description). Should the metadata be available in more than one format, use one aggregated resource per metadata format.
It is recommended that the namespace of the format of the metadata is identifiable through a URI.
It is required that the Aggregated Resource which contains the metadata has a specific metadata field to indicate its semantic type.

7. Use of vocabularies

If vocabularies are used to describe metadata or a semantic relationship it is required to state which vocabularies are being used. It is recommended that the vocabularies being used are identifiable through a URI. It is recommended to commit to the DRIVER-specified vocabularies.

Further research:

OAI-ORE offers a perspective of an immense network of linked information. Some information from within the text of a traditional publication, like references and lists of publications, can be used to extend this network. We would like to express this information in the aggregation, for example as a set of aggregated resources or as an aggregated XML-file with dcterms:references.