

---

# A PORTAL FOR DOCTORAL E-THESES IN EUROPE;

## Lessons Learned from a Demonstrator Project

Drs. M.P.J.P. Vanderfeesten

{vanderfeesten@surf.nl}



### PREFACE

This report is based on the work carried out in a small pilot project to set up a European doctoral e-theses Demonstrator. The work was funded by the Joint Information Systems Committee (JISC) in the UK, the National Library of Sweden and SURFfoundation in the Netherlands. The project has been performed by SURFfoundation.

We wish to express our sincere gratitude to Paul Needham for his support and contributions, as well as Rita Voigt, Susanna Dobratz and others involved in the Knowledge Exchange Workshop on Interoperability of Repositories.

Maurice Vanderfeesten, Gerard van Westrienen

Utrecht, July 2007.

© Stichting Surf, June 2007

This report appears under the Creative Commons licence

[Attribution-Noncommercial-No Derivative Works 2.5 Netherlands](https://creativecommons.org/licenses/by-nc-nd/2.5/nl/).

## TABLE OF CONTENTS

<b>1 Management summary .....</b>	<b>5</b>
<b>2 Introduction .....</b>	<b>7</b>
2.1 Structure of the report .....	9
<b>3 Theory: The process of harvesting and interoperability of repositories .....</b>	<b>10</b>
3.1 Repository systems, from stored collection to interoperable output .....	12
3.2 Mapper of the Repository systems .....	13
3.3 Gate of the Repository systems .....	15
3.4 Collector from the Harvester of the Portal .....	16
3.5 Mapper from the Harvester of the Portal .....	16
3.6 Uploader from the Harvester of the Portal .....	18
3.7 Search engine of the Portal .....	18
3.8 Web interface of the Portal.....	20
<b>4 Theory: A model of interoperability .....</b>	<b>23</b>
4.1 Harvester.....	24
<b>5 Practice: The e-Theses demonstrator deployment.....</b>	<b>26</b>
5.1 Meresco metadata management (harvest, store, search) .....	26
5.2 Sahara harvester .....	29
<b>6 Practice: Interoperability issues and recommendations</b>	<b>32</b>
6.1 Collection of the Repository systems .....	32
6.1.1 Generic problem: the external record has a finer granularity then the internal record .....	32
6.2 Mapper of the Repository systems .....	33
6.2.1 Generic problem: the internal record has a finer granularity then the external record, as a result creating more ambiguity.....	33
6.2.2 Generic problem: simple Dublin Core lacks content guidelines for interoperable utilisation .....	34
6.2.3 e-Thesis related problem: simple Dublin Core lacks e-Thesis specific expressions.....	36
6.2.3.1 Ad 1. Recommendation to adapt simple Dublin Core metadata for e-theses	39

6.2.3.2	Ad 2. Generic metadata format for Academic Information Domain, with e-Theses elements incorporated.....	40
6.3	The Gate of Repository systems.....	41
6.3.1	Generic problem: the OAI protocol v.2.0 is not well supported by the repository.....	41
6.4	OAI-PMH.....	41
6.4.1	Generic problem: wrong XML encoding scheme.....	42
6.4.2	Generic problem: not well-formed XML structure – URL encoding.....	42
6.4.3	Generic problem: XML Validation, not recognised XML structures. ....	43
6.4.4	Generic problem: very short lifespan of resumption token. ....	44
6.4.5	Generic problem: BaseURL interpretation. ....	45
6.4.6	Generic Problem: Firewall blocking the harvester. ....	45
6.4.7	Generic problem: changing identifiers & updating timestamps ....	46
6.5	The Collector of the Harvester from the Portal .....	46
6.5.1	e-Thesis problem: Collected metadata features do not meet service features. 47	
6.6	Mapper of the Harvester from the Portal .....	47
6.6.1	Generic problem: Collecting metadata without normalisation results in ambiguous data, which makes it harder to create interoperability .....	48
6.7	Uploader of the Harvester from the Portal .....	49
6.8	Search engine of the Portal .....	49
6.9	E-Theses filter of the Web interface from the Portal.....	49
6.9.1	e-Thesis problem: Filter is not accurate enough .....	49
6.9.1.1	Ad Sub-problem 1: front-end filter.....	50
6.9.1.2	Ad Sub-problem 2: mapping and up-scaling.....	52
6.9.1.3	Ad Sub-problem 3: heterogeneous collections .....	53
6.10	Results of the Web interface from the Portal.....	54
6.10.1	e-Thesis problem: cultural differences create ambiguous metadata, ambiguous metadata creates ambiguous search results.....	54
6.10.2	Generic problem: Access to repository or full text document restricted by embargo 56	
6.10.3	Generic problem: Access to repository system only from University Campus 57	
6.10.4	Generic problem: Certificates and User experience .....	58

## **7 Summary and Conclusions ..... 59**

7.1	Generic recommendations related to repositories.....	59
7.1.1	Questions and answers in a wider IR perspective .....	59
7.2	E-theses specific recommendations.....	60
7.2.1	Questions and answers on harvesting e-theses .....	61
7.2.2	Questions and answers on metadata quality for e-theses .....	61
7.2.3	Questions and answers on Rich e-theses metadata formats.....	63
7.3	Recommendations related to interoperability between data and service providers 63	
7.3.1	Questions and answers on Interoperability .....	64
7.4	Cultural and educational recommendations .....	65
7.4.1	Questions and Answers on ETD's involving Cultural aspects.....	65
7.5	Conclusion .....	67

## **Annex ..... 68**

### **I. Basic repository information & Metadata analysis [WP:0+2] ..... 72**

### **II. Functional specifications for an e-Theses portal [WP:1] 96**

### **III. Harvester quickscan [WP:3] ..... 97**

### **IV. Issues inventory and recommendations for e-Theses [WP:5+6] ..... 104**

### **V. Screenshots of the Demonstrator [WP:7+8]..... 112**

### **VI. Example of Embargo handing ..... 125**

### **VII. List of Terms ..... 127**

# 1 MANAGEMENT SUMMARY

*If the facts don't fit the theory, change the facts.*

*- Albert Einstein*

For the first time various repositories with doctoral e-theses have been harvested on an international scale. This report describes a small pilot project which tested the interoperability of repositories for e-theses and has set up a freely accessible European portal with over 10,000 doctoral e-theses<sup>1</sup>.

Five repositories from five different countries in Europe were involved: Denmark, Germany, the Netherlands, Sweden and the UK. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) was the common protocol used to test the interoperability. Based upon earlier experiences and developed tools (harvester, search engine) of the national DAREnet service in the Netherlands, SURFfoundation could establish a prototype for this European e-theses Demonstrator relatively fast and simple.

Nevertheless, some critical issues and problems occurred. They can be categorised into the following topics:

- a) **Generic issues related to repositories:** the language used in the metadata fields differs per repository; sometimes all fields are in English; or fields are both in the local language and in English; or fields are only in the local language. It even differs per record. Furthermore, the quality of the data presented differs. The adagio 'garbage in is garbage out' is very much applicable for these kind of search services. Better validation of the data at the repository-side is needed<sup>2</sup>. A further issue is the semantic and syntactic differences in metadata between repositories, which means that the format and content of the information exchange requests are not unambiguously defined. For a fast and feasible setup of services, further standardisation is recommended and references are made to the Guideline developed by the European DRIVER project at [www.driver-support.eu](http://www.driver-support.eu).

An issue of a generic nature is also that the representation of a complex or compound structure of the thesis or enhanced (multimedia) publication. To create this structure we recommend to use the DIDL document meta-structure. This is a MPEG-21 standard that is flexible enough to support multiple purposes, does not rely on metadata formats and is self-descriptive<sup>3</sup>. DRIVER provides the specifications in the guidelines.

- b) **E-theses specific issues:** to be able to harvest *doctoral* theses, the service provider needs to be able to filter on this document type. Up to now there is no commonly agreed format, which makes semantic interoperability possible. It is recommended to distinguish between the various types of theses in the Dublin Core format "dc:type" and use the following qualifications: 'Bachelor thesis',

---

<sup>1</sup> See the demonstrator temporarily at <http://e-thesis.sharelab.cq2.org>

<sup>2</sup> The DRIVER project is developing a Validation Tool. See [www.driver-support.eu](http://www.driver-support.eu)

<sup>3</sup> A visualisation of a DIDL document example can be found at [http://www.surfgroepen.nl/sites/e-theses-demonstrator/Shared%20Documents/DIDL\\_document/xmlContainer-v2.2.2.xml](http://www.surfgroepen.nl/sites/e-theses-demonstrator/Shared%20Documents/DIDL_document/xmlContainer-v2.2.2.xml)

'Master thesis', 'Doctoral thesis'. Furthermore, there is a need to standardise on the date field, as various dates may be referred to (date of publication; date of graduation; starting date of the research etc). We recommend to use in the Dublin Core metadata field "dc:date" the date of publication of the doctoral e-theses. A last e-theses specific issue is related to the metadata field "contributor". For a doctoral thesis one could distinguish various 'contributors', like juror, committee member, referee, etc. We recommend to use the contributor field in Dublin Core for the person who supervised the thesis.

- c) **Issues related to data providers and service providers:** besides the use of the OAI-protocol for metadata harvesting and the use of Dublin Core it is recommended for data providers to further standardise on the semantic interoperability by using the DRIVER guidelines<sup>4</sup> with an addition of the e-Theses specific recommendations described above. To be able to offer more than basic services for e-Theses, one has to change the metadata format from simple Dublin Core to a richer and e-Theses specific one. To offer the same quality as the basic recommendation on syntactic interoperability, the e-Theses metadata format has to be unambiguously defined. Currently, it is recommended to make a further study to benchmark richer formats like ETD\_MS, UKETD\_DC and XMetaDiss on syntactic and semantic interoperability this can possibly be taken up in DRIVER II. In this project, we operated as a service provider. We needed to fix, normalise and crosswalk the differences between every repository to get a standard syntactic and semantic metadata structure. For five repositories this work is manageable. However, when the number of repositories increases, this will become harder. The scaling up is a big issue. To stimulate the broad take up of various services, data providers have to work on implementing standards that create interoperability on syntactic and semantic levels.
- d) **Cultural and educational differences:** In every country the educational processes are different. The Bologna declaration has standardised education in Europe up until the Master's degree. After this degree, there is no clear European or international definition on the post-graduate degree. Not only the graduation and publication process differs, but also the duration of the research process. Therefore the quality of the results in a cross-European search of doctoral theses may vary enormously.

Doctoral theses contain some of the most current and valuable research produced within universities, but are underused as research resources. Where electronic theses and dissertations (ETDs) are publicly available, they are used many times more often than paper theses that are available only via inter-library loan.

The current developments in Europe around digital repositories is encouraging for the visibility and retrievability of doctoral e-theses. They are an integrated part of the academic information domain and we therefore recommend to embed doctoral theses into the general academic repositories infrastructure.

This project proved that within this repository infrastructure, interoperability of doctoral theses on a European scale is possible. But we only have reached the first phase. Further work needs to be done to create qualitatively and quantitatively richer services, and thereby make the visibility, retrievability and (re)use of this valuable knowledge possible.

---

<sup>4</sup> For DRIVER guidelines look at <http://www.driver-support.eu>

## 2 INTRODUCTION

*Keep true to the dreams of thy youth.*

*- Friedrich von Schiller (1759 - 1805)*

Doctoral theses contain some of the most current and valuable research produced within universities, but are underused as research resources. Where electronic theses and dissertations (ETDs) are publicly available, they are used many times more often than paper theses that are available only via inter-library loan.

In many countries in Europe, institutional repositories have been set up, based upon the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH); many of them with doctoral e-theses. In some countries they have been harvested at a national level, e.g. in the Netherlands with the 'Promise of Science' portal<sup>5</sup>. Up until now, there hasn't been any initiative in practice to harvest various repositories with e-theses on an international scale and to set up a freely accessible European portal and test the interoperability in practice.

Therefore, a pilot project started in October 2006 with the following aims:

- to share current practices, relevant for interoperability of e-theses repositories on an international/European level;
- to get better insight into (critical) issues and potential solutions related to interoperability of e-theses repositories;
- to set up a demonstrator – an interoperable portal of European e-theses, based upon the experiences of the partners involved.

Five repositories from five different countries in Europe were involved: Denmark, Germany, the Netherlands, Sweden and the UK. The repositories needed to comply with the following points of departure:

- available at least until July 2007;
- support the OAI Protocol for Metadata Harvesting (OAI-PMH) for machine-machine communication;
- contain *Doctoral* Theses;
- support the OAI Dublin Core Metadata format;
- contain a full text object, or a set of objects that fully represents the Doctoral Thesis;
- The full text of the Doctoral Theses must be Open Accessible.

We made clear from the beginning that it was not our intention to sustain the service or demonstrator after the project.

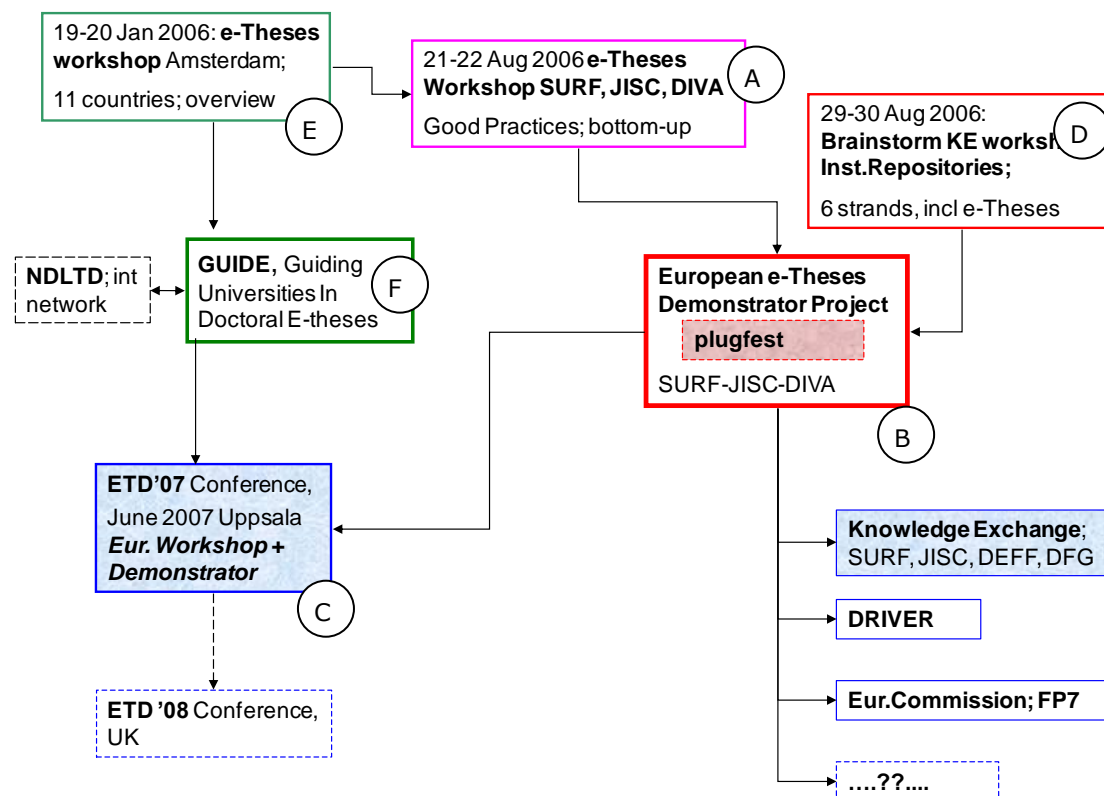
The project was related with various events in the last year and a half. To learn from the experiences and good practices, JISC and SURFfoundation organised a workshop in Amsterdam, in January 2006, attended by representatives from 11 countries in Europe<sup>6</sup>.

---

<sup>5</sup> See: <http://www.darenet.nl/promiseofscience>

<sup>6</sup> For a report, see: <http://www.ariadne.ac.uk/issue46/e-theses-rpt/>.

This workshop was the starting point for various activities relating to doctoral e-Theses in Europe (see (E) in Figure 1). Also, this spark set the GUIDE group (F) in motion again. GUIDE<sup>7</sup> is the European Working Group of the Networked Digital Library of Theses and Dissertations (NDLTD<sup>8</sup>). It is an open group, whose aim is to stimulate European doctoral Electronic Theses and Dissertations (ETD) developments. One European project, for example, is the DART-Europe project<sup>9</sup>.



**FIGURE 1: INFLUENTIAL E-THESES EVENTS IN THE PAST AND FUTURE**

In August 2006, SURF, JISC and DIVA (see (A) in Figure 1) organised a two day workshop, in Utrecht, with representatives from three specific doctoral e-theses projects in Europe: DIVA (Sweden), Ethos (JISC, UK) and Promise of Science (SURFfoundation, the Netherlands). We exchanged information, good practices and lessons learned, and decided to set up a demonstrator of e-theses in Europe with a few countries. A small amount of seed money had been made available from JISC, SURFfoundation and the National Library of Sweden for this demonstrator. The demonstrator was presented at the international ETD 2007 conference in Uppsala, Sweden in June 2007 (C). The Knowledge Exchange<sup>10</sup>

<sup>7</sup> See: <http://www.dartington.ac.uk/guide/index.asp>

<sup>8</sup> NDLTD: The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organisation dedicated to promoting the adoption, creation, use, dissemination and preservation of electronic analogues to the traditional paper-based theses and dissertations. [www.ndltd.org](http://www.ndltd.org)

<sup>9</sup> DART-Europe: an academic consortium of university partners, who are undertaking work on E-Theses. <http://www.dartington.ac.uk/dart/>

<sup>10</sup> Knowledge Exchange is a co-operative effort that intends to support the use and development of ICT infrastructure for higher education and research. <http://www.knowledge-exchange.info/>



workshop on 16 and 17 January in Utrecht can be seen as an important moment in this project (B), where the representatives of the 5 countries involved (besides the 4 KE countries, also Sweden) presented a first pilot of a working portal of European e-Theses<sup>11</sup>. At the same time, SURFfoundation, JISC, DEFF (Denmark) and the DFG (Germany) have been working closely together in "Knowledge Exchange" (D). They have organised a workshop on Interoperability of Institutional Repositories in these 4 countries in Europe. One of the six strands was: "a Plug fest on e-Theses". For this strand, Sweden (the DIVA project) was also invited.

This current e-theses demonstrator project (B) brings the various developments, initiatives and know-how together.

## 2.1 STRUCTURE OF THE REPORT

Chapter 1 contains the Management Summary and Chapter 2 this Introduction.

In Chapter 3 we will introduce the reader to the basics about the harvesting process in the Open Archives Architecture. In this chapter we will follow the metadata from a university Database to a visitor of an interoperable service.

Chapter 4 is focussing on general interoperability aspects. Chapter 5 is about the deployment of the Demonstrator, the software architecture that has been used and the project activities.

Chapter 6 is based on the metadata flow described in Chapter 3, where we encountered issues from our daily practice working with the demonstrator. Besides these issues, recommendations are also provided based upon international guidelines, best practices and expertise from the participating members, Knowledge Exchange and the DRIVER project. These recommendations for data- and service-providers can serve as guidelines, compatible with the DRIVER guidelines, to setup an interoperable Doctoral e-Theses service.

Chapter 7 contains the summary and the conclusions on the generic issues that concern all repositories, issues that are special to ETD's, issues that are relevant to data and service providers, and several cultural aspects.

In the Annexes the full workout of the work-packages of the project is presented.

- 
- Denmark's Electronic Research Library (DEFF) in Denmark
  - German Research Foundation (DFG) in Germany
  - Joint Information Systems Committee (JISC) in the United Kingdom
  - SURFfoundation (SURF) in the Netherlands

<sup>11</sup> See: <http://e-thesis.sharelab.cq2.org>

### 3 THEORY: THE PROCESS OF HARVESTING AND INTEROPERABILITY OF REPOSITORIES

*Any sufficiently advanced technology is indistinguishable from magic.*

*- Arthur C. Clarke (1917 - ),  
"Profiles of The Future", 1961 (Clarke's third law)*

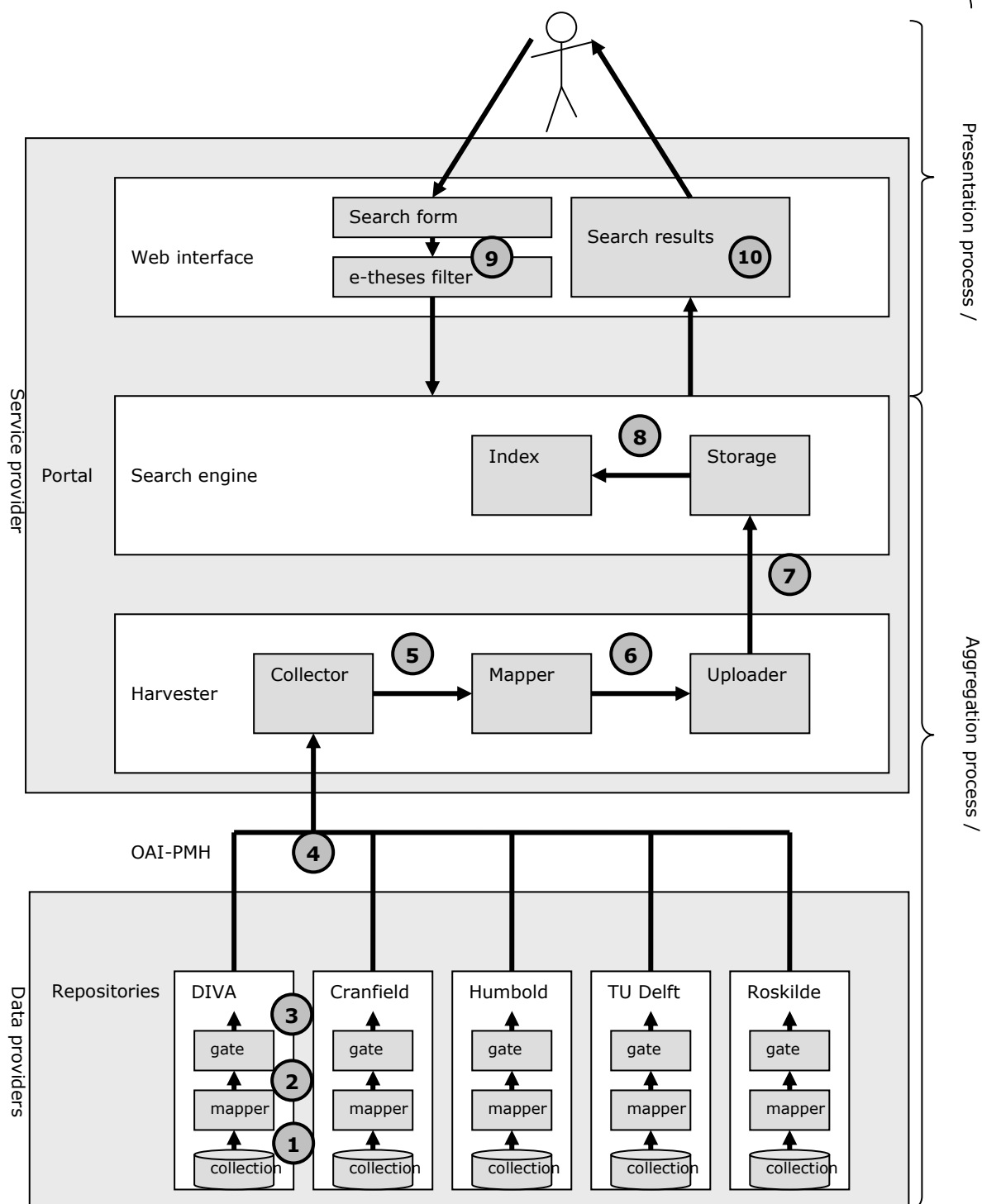
The purpose of this chapter is to familiarise the reader with the Aggregation and Presentation Process of e-Thesis publications. We will use this chapter to create a common ground for the following chapter. We will create this common ground by using two example records from the DIVA and Humboldt repositories.<sup>12</sup> In Figure 2, below, we have indicated the data transitions<sup>13</sup> from one component to another, using the numbers (1) through to (10). The different components (like 'gate' and 'collector') we have indicated by a dark-grey colour in Figure 2. These transitions and components are explained in this chapter; also by the aid of the two example records. This basic framework will be used to explain the issues involved in chapter 6.

In the Figure 2, below, we show the global framework of the Open Archives model. The model is based on two parties, the data provider and service provider. The data provider is the party who provides the content. Often the term 'repository' is used in stead of 'data provider'. The service provider is a party who delivers services on-top of the data providers' content. In this example, we offer a portal service that enables access to all repositories from one website. In the text, we also will use the term 'portal' for activities that are carried-out by service providers.

---

<sup>12</sup> The example records are modified from reality to capture all issues found in the metadata and concentrate them in these two examples.

<sup>13</sup> In the illustrations this number will be placed inside a circle, in the text between rounded brackets.



**FIGURE 2: GENERAL DOCUMENT WORKFLOW OF THE OAI ARCHITECTURE**

When we look at the left side of Figure 2, we can make a distinction between two major groups: the data providers and the service providers. Service providers depend totally on the content that data providers provide. These two groups communicate over the internet with a machine-machine language called OAI-PMH (see section 3.3 and 3.4). When we look at the right side of the illustration, we see the terms 'presentation process' and 'aggregation process'. Portal visitors are presented with a website where they can make search queries and results are returned. This is done in the presentation process. But, before this presentation can be made, a lot of preparation has to be done in the

aggregation process. In the Demonstrator, metadata is harvested from several repositories (DIVA, Cranfield, Humboldt, TU delft and Roskilde). This harvesting is carried out using the OAI-PMH communication protocol. The OAI-PMH ensures that a harvester makes a specific request and the repository can understand that request. For example a harvester requests the following: "Give me all your metadata records" by using exactly the following sequence of characters: "?verb=ListRecords". The repository gate understands this request and generates a response in XML. This XML response must be exactly structured in a way the harvester can 'understand'. The XML response can be shown in a plain browser window, for example click on the following link:

[http://repository.tudelft.nl/oai?verb=ListRecords&metadataPrefix=oai\\_dc](http://repository.tudelft.nl/oai?verb=ListRecords&metadataPrefix=oai_dc)

This shows the XML response of the TUDelft repository. The XML response is collected by the collector of the harvester. This XML response contains metadata records. These metadata records are mapped by the mapper into other records to enrich the metadata. Then this is uploaded by an uploader to a storage area in the search engine. The metadata collected from all repositories is stored in the storage area. From there this metadata is indexed by the search engine and ready for search requests from a portal visitor.

When the aggregation process is ready, the presentation process can start working. In this process we start at the top of Figure 2, with the visitor. The visitor puts a search term in the search form. To this term some extra information is added by the e-theses filter. This extra information makes sure that only Doctoral e-Theses are returned from the search engine. After the search terms and the filter information has been processed by the search engine, the search engine gives back the results and is transformed by the web interface into a human readable representation.

To describe how the data and service provider work together to serve the end-user, as best as possible, we use an example of two records that we are going to follow throughout the whole process from repository collection to the portal visitor.

### 3.1 REPOSITORY SYSTEMS, FROM STORED COLLECTION TO INTEROPERABLE OUTPUT

To fully understand the complexity of interoperability and interrelational consequences of an issue or recommendation we have to start at the level inside the repository. The story starts with a record in the collection of the Humboldt and DIVA repositories. DIVA uses a repository system called "The DIVA Publishing System"<sup>14</sup> Humboldt uses a repository system called "The e-doc Server". Figure 3 zooms in on the repository systems. In this example one can see that Humboldt and DIVA use different formatting of the "Internal Records" (1) in their total **Collection**. DIVA uses XML to store their metadata information, and Humboldt uses a Database Management System. In order for the outside world to read these different formats, the repository re-formats/transforms this metadata information to another format that can be read by others. This format that is readable by others is called an interoperable format. In this example, we use the term "External Record" (2) to indicate that the metadata information of a record is transformed into an interoperable format, for example, simple Dublin Core. This transformation is done by a component called the **Mapper**. After that, a component called the **Gate** will put the

---

<sup>14</sup> For more detailed information about the DIVA publication System, please see <http://epc.ub.uu.se/files/E-Poster.pdf>

transformed metadata records into an envelope and send it to a service provider who requested the contents of the envelope (3). This envelope must also be an interoperable standard; otherwise the receiving party cannot understand what to do with it. It will be as if one receives a box with a description in Arabic: أنا أنقله أتت ثقت آتف إيشات ج لله.<sup>15</sup> Using the Open Archive Initiative's communication protocol for metadata harvesting (OAI-PMH), the requests and responses are formalised. With this formalisation other machines can 'understand' each other. This understanding is based on trust. A harvester, sending a request to a repository, trusts this repository to get the message and expects the repository to respond in a predictable way. Both the repository and the harvester are expected to behave according to this formal communication protocol in order to guarantee interoperability.

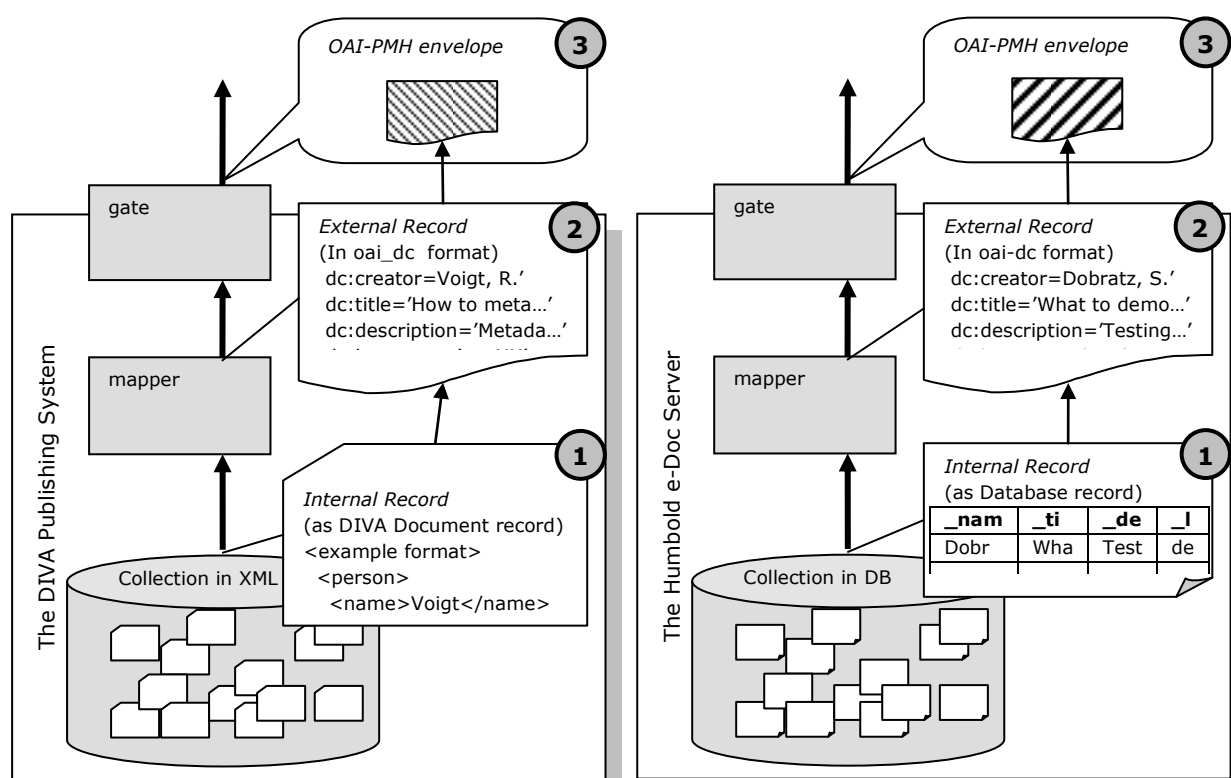


FIGURE 3: DETAIL OF THE REPOSITORY DOCUMENT FLOW




### 3.2 MAPPER OF THE REPOSITORY SYSTEMS



Each repository system can deliver the metadata in a format they support. For example Humboldt University supports many metadata formats<sup>16</sup>. A record can be represented in simple Dublin Core, but also in Xmetadiss or E-Prints' Application Profile<sup>17</sup>. However, the OAI-PMH standard demands that metadata is provided, at least, in the simple Dublin Core

<sup>15</sup> To read this, transform to a western type font and read in reverse order.

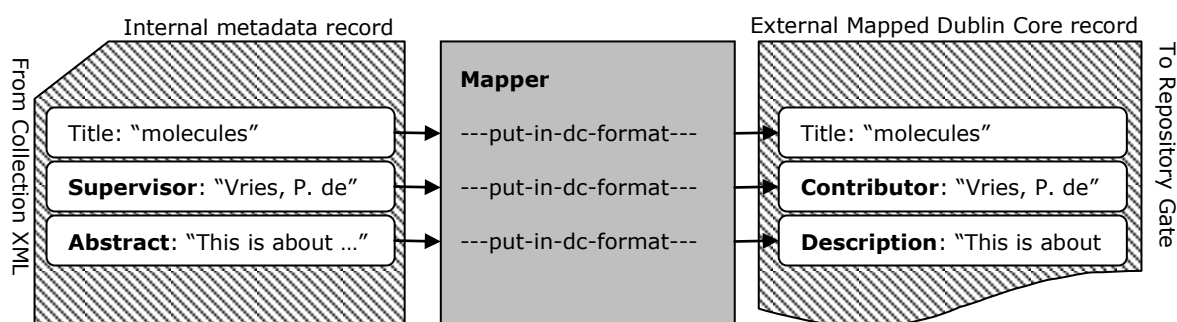
<sup>16</sup> <http://edoc.hu-berlin.de/OAI-2.0?verb=ListMetadataFormats>

<sup>17</sup> Click the following links to see the same record in different formats:  
[[simple Dublin Core](#)], [[Xmetadiss](#)], [[E-Prints' Application Profile](#)]

format (oai\_dc; see Figure 3, (2)). To deliver this format in simple Dublin Core (oai\_dc), the repository system transforms from the internal XML format  or Database fields  to a common simple Dublin Core format . (Look at the different shapes of the records.) This record then is transformed into a format for external use. At this stage, we call this record an 'External record'.

This transformation work is done by a Mapper, that places the content of one field of the internal format  into a field in the external format . Below, in Figure 4, we zoom in on the transformation process of the Mapper. We can see the Internal record changes in appearance, yet the content remains largely the same. In this particular example, we can see that the field names change from 'Supervisor' to 'Contributor' and from 'Abstract' to 'Description'.

**Transformation:** put content from one format to a different format



**FIGURE 4: MAPPER FROM THE DIVA REPOSITORY INVOLVED IN A TRANSFORMATION PROCESS**

The actual results of the External Records are shown in Table 1 and Table 2. These tables, below, are metadata examples from the DIVA and Humboldt repository systems. These examples will be used as a beacon throughout chapters 3 and 6.

```
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">

  <dc:title>Mixing Oil and Water : Studies of the Namibian Economy</dc:title>
  <dc:creator>Stage, Jesper</dc:creator>
  <dc:subject>Namibia</dc:subject>
  <dc:subject>energy use</dc:subject>
  <dc:subject>structural decomposition analysis</dc:subject>
  <dc:subject>hedonic pricing</dc:subject>
  <dc:subject>townships</dc:subject>
  <dc:subject>groundwater use</dc:subject>
  <dc:subject>fisheries</dc:subject>
  <dc:subject>bioeconomic modelling</dc:subject>
  <dc:description>This thesis consists of economic aspects of natural resource ... </dc:description>
  <dc:publisher>Umeå University, Sweden</dc:publisher>
  <dc:date>2003</dc:date>
  <dc:format>text/html</dc:format>
  <dc:format>application/pdf</dc:format>
  <dc:format>application/xml</dc:format>
  <dc:identifier>http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-117</dc:identifier>
  <dc:identifier>urn:nbn:se:umu:diva-117</dc:identifier>
  <dc:type>text.thesis.doctoral</dc:type>
  <dc:source>91-7305-508-5</dc:source>
  <dc:language>en_UK</dc:language>
  <dc:rights>Copyright Jesper Stage </dc:rights>
</oai_dc:dc>
```

**TABLE 1: EXAMPLE OF OAI\_DC METADATA RECORD FROM DIVA IN XML = **

```
<?xml version="1.0" encoding="UTF-8"?>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>Akute Enzephalitiden im Erwachsenenalter</dc:title>
  <dc:title>klinisches und ätiologisches Spektrum und Langzeitverlauf</dc:title>
  <dc:creator>Schielke, Eva</dc:creator>
  <dc:subject>Medizin</dc:subject>
  <dc:subject>Enzephalitis</dc:subject>
  <dc:subject>Langzeitverlauf</dc:subject>
  <dc:subject>Neuropsychologie</dc:subject>
  <dc:subject>Magnetresonanztomographie (MRT)</dc:subject>
  <dc:subject>encephalitis</dc:subject>
  <dc:subject>long-term outcome</dc:subject>
  <dc:subject>neuropsychology</dc:subject>
  <dc:subject>magnetic resonance imaging (MRI)</dc:subject>
  <dc:subject>Medizin</dc:subject>
  <dc:subject>YE 4500</dc:subject>
  <dc:description>Akute Enzephalitiden treten überwiegend .....</dc:description>
  <dc:description>Acute encephalitis occurs mainly sporadically .....</dc:description>
  <dc:publisher>Medizinische Fakultät - Universitätsklinikum Charité</dc:publisher>
  <dc:date>2001-11-06</dc:date>
  <dc:type>Text</dc:type>
  <dc:type>dissertation</dc:type>
  <dc:format>text/html</dc:format>
  <dc:identifier>http://edoc.hu-berlin.de/habilitationen/schielke-eva-2001-11-06/HTML/index.html</dc:identifier>
  <dc:format>application/pdf</dc:format>
  <dc:identifier>http://edoc.hu-berlin.de/habilitationen/schielke-eva-2001-11-06/PDF/Schielke.pdf</dc:identifier>
  <dc:language>eng</dc:language>
</oai_dc:dc>
```

TABLE 2: EXAMPLE OF OAI\_DC METADATA RECORD FROM HUMBOLDT IN XML = 

### 3.3 GATE OF THE REPOSITORY SYSTEMS

The OAI-PMH is an XML page that serves as an envelope which wraps the metadata describing a resource (see Table 3). This XML page to wrap metadata records we call the 'OAI-PMH' (see Figure 3, (3)). The OAI-PMH envelope is generated by a module called the repository Gate. The repository Gate responds to requests from the harvester of the Service provider (explained later-on).

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-
PMH.xsd">
  <responseDate>2007-05-08T13:21:47Z</responseDate>
  <request verb="GetRecord" identifier="oai:DiVA.org:umu-117" metadataPrefix="oai_dc">
http://www.diva-portal.org/oai/OAI</request>

  <GetRecord>
    <record>
      <header>
        <identifier>oai:DiVA.org:umu-117</identifier>
        <timestamp>2006-03-19</timestamp>

        <setSpec>umu</setSpec>
        <setSpec>postgraduateTheses</setSpec>
        <setSpec>comprehensiveTheses</setSpec>
        <setSpec/>
      </header>
      <metadata>... between this space the metadata of one record is placed ... </metadata>
    </record>
  </GetRecord>
</OAI-PMH>
```

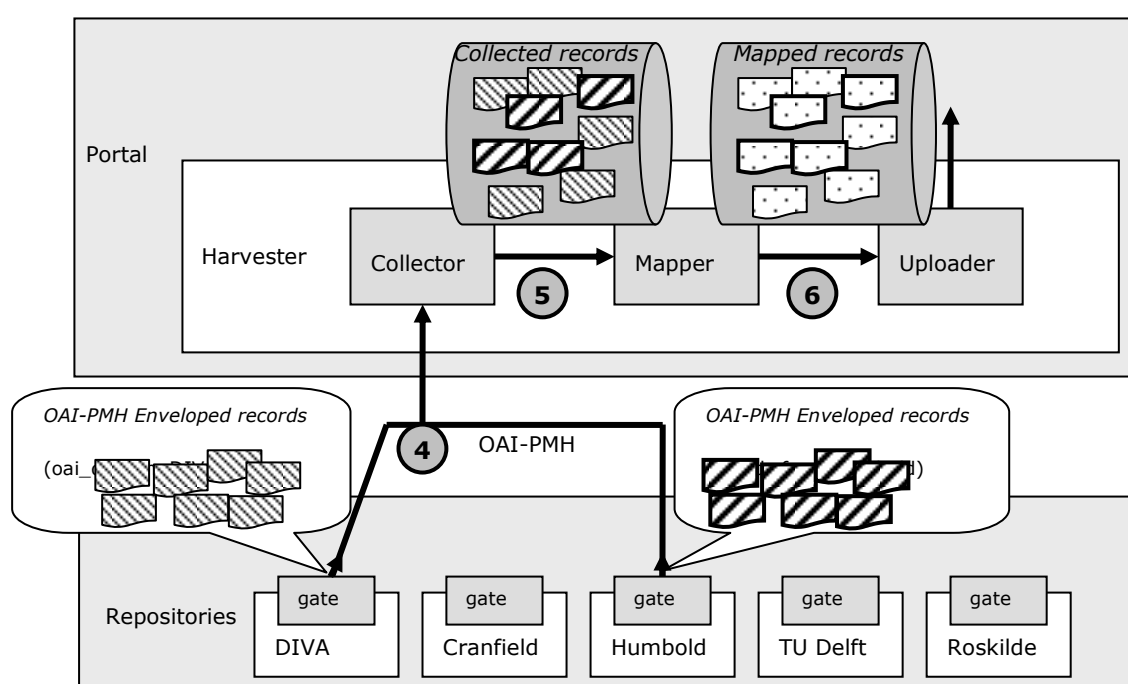
TABLE 3: EXAMPLE OF OAI-PMH ENVELOPE OF A RECORD FROM DIVA IN XML

### 3.4 COLLECTOR FROM THE HARVESTER OF THE PORTAL

We now make the step to the Service Provider. Before the service provider has the metadata records, it has to collect them from the repositories (see Figure 5, (4)). Fortunately the repository Gates understand and speak the OAI-PMH 'language'. The Collector is a module that is part of the Harvester. The Harvester aggregates metadata records from the repositories. These repositories provide metadata records in a useful format for the service providers.



In the example, as shown in Figure 5, the Collector asks the Gate of the DIVA and Humboldt repositories to deliver some metadata records. As a response, the Repositories Gates give a number of records the Collector requested.

When the Collector has received the metadata records in the OAI-PMH envelope, it immediately delivers them to the Mapper module at the Service provider side (see Figure 5, (5)).



**FIGURE 5: COLLECTOR ASKS SOME RECORDS FROM THE DIVA AND HUMBOLDT REPOSITORY**

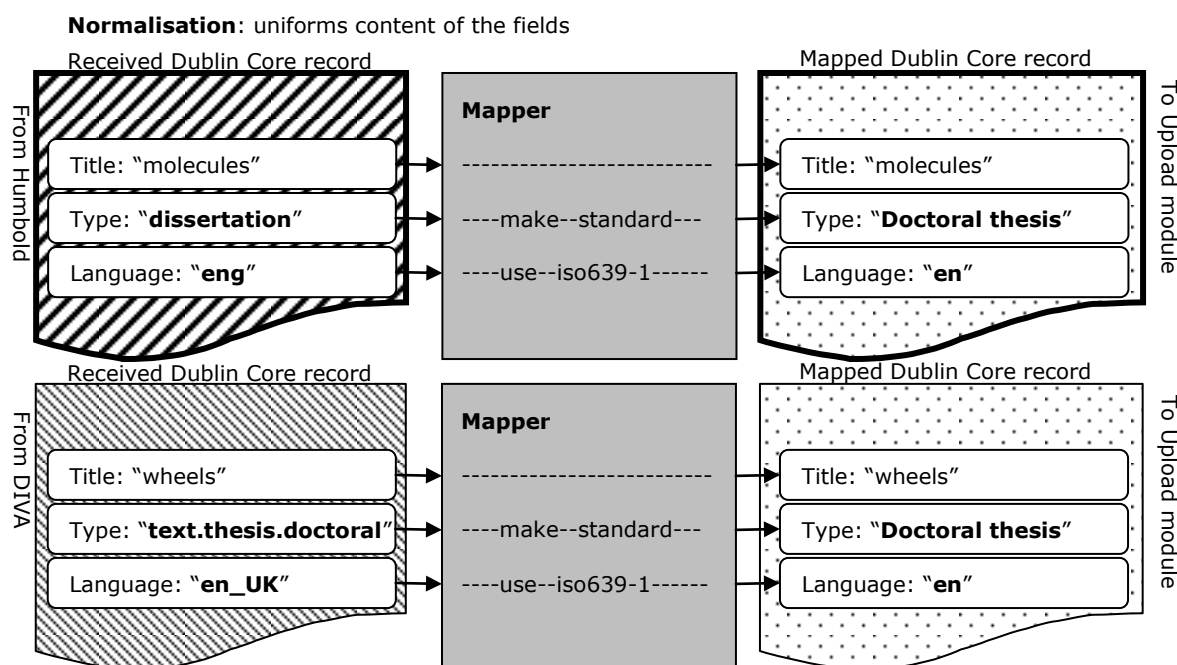
### 3.5 MAPPER FROM THE HARVESTER OF THE PORTAL

As you can see in the metadata examples in Table 1 and Table 2, the External records of DIVA  and Humboldt  use similar fields, but the type of content might differ a little, for example, if one looks at the dc:type field or the dc:language field. The type of both documents are Doctoral theses, however the two repositories use different terms to describe the same concept. Humboldt uses the term 'dissertation' and DIVA uses 'text.thesis.doctoral'. As a service provider you don't want your visitor to look first for 'dissertation' and next search for 'text.thesis.doctoral'. Also the content of the Language field used in the records from Humboldt and DIVA differs. Humboldt uses "eng", a three letter code, and DIVA uses "en\_UK", both to indicate the English language.

To get control over the many different interpretations of each term, one has to normalise the different terms to one term. Normalisation increases the quality of a service by correcting and standardising terms. For example the terms network, networking and networks mean, in essence, the same. Also Professional dissertation, Doctoral work, and Doctoral thesis are similar concepts. When a person is looking for Doctoral work, only 33%



of the records will return without normalisation. With normalisation 100% of the records belonging to the concept Doctoral thesis will return, because the terms for Doctoral Thesis are normalised by the service provider. The Mapper of the service provider is a mechanism that is able to correct these little flaws to increase the quality of the service. (See Figure 6.)



**FIGURE 6: NORMALISATION PROCES OF ONE DIVA AND ONE HUMBOLDT RECORD**

In the example shown in Figure 6, we can see that the Mapper processes a Humboldt record as follows: the content of the title field is copied into the title field of the mapped Humboldt record, the type field containing the text 'dissertation' is changed to 'Doctoral thesis', and the language field contains a three-letter abbreviation of a language, which is changed in a two-letter abbreviation. Also, this normalisation is done for the DIVA records. As a result the metadata in the portal service uses the same terms. Multi-interpretations are flattened out, and questions like: "what is exactly meant by 'dissertation', is it a Masters or Doctoral thesis?" are solved by the service provider's normalisation process. The user will be offered a high-quality cross-repository umbrella search engine. The end-user can, for example, search better to find a Doctoral thesis that is written in the English language<sup>18</sup>.

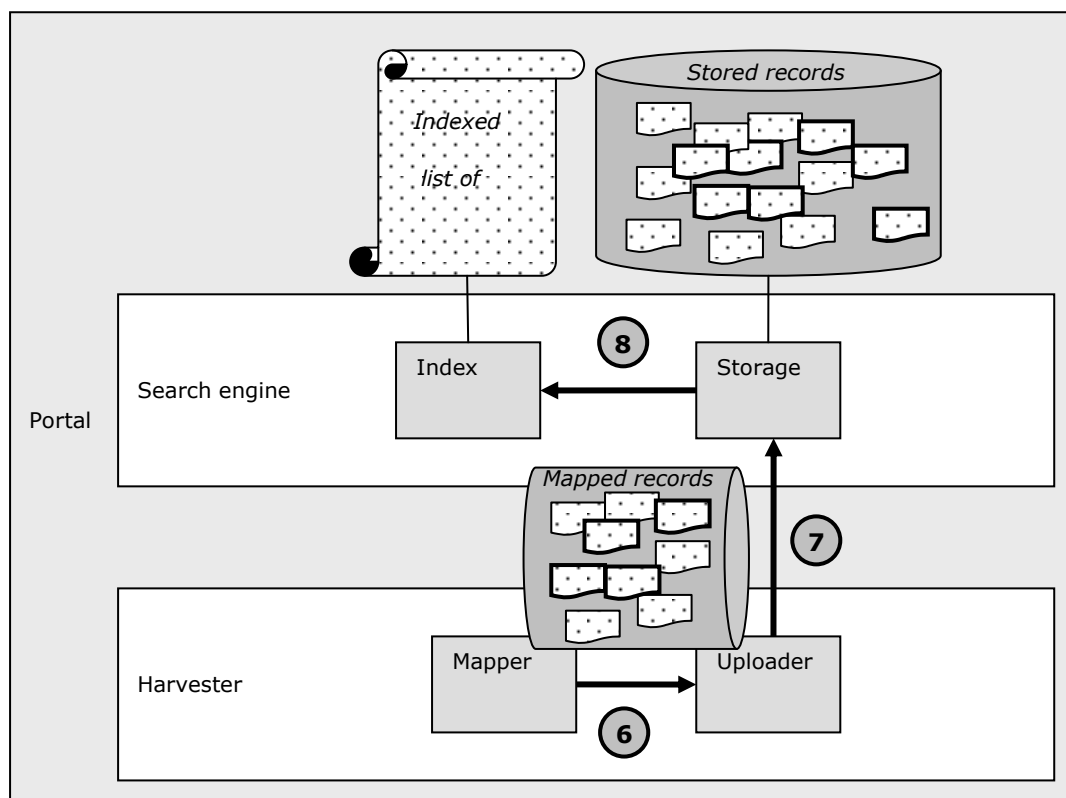
<sup>18</sup> This normalisation process has not been done in the demonstrator due to budget reasons. It is time consuming and expensive to make crosswalks for every repository system, therefore the content of the fields have been mapped directly without making changes.

Crosswalks are pieces of scripting software that perform automatic normalisation and transformation operations on each record. Below is an example of the mapping script that is been used by the demonstrator:

```
upload.fields['dccontributor'] = join(dc.contributor)
upload.fields['dcreator'] = join(dc.creator)
upload.fields['dcdate'] = dc.date
upload.fields['dcdescription'] = join(dc.description)
upload.fields['dcidentifier'] = join(dc.identifier)
upload.fields['dcsubject'] = join(dc.subject)
upload.fields['dctitle'] = join(dc.title)
upload.fields['dctype'] = join(dc.type)
```

### 3.6 UPLOADER FROM THE HARVESTER OF THE PORTAL

To continue: The Mapper processes the Mapped records to the Uploader module. (See Figure 7, (6).) The Uploader component is like a traffic agent and tells the stream of mapped metadata records where to go. In this case the Uploader component sends the metadata to the storage facility of the Search engine, where all the mapped metadata will be stored. (See Figure 7, (7).)



**FIGURE 7: THE UPLOADER SELECTS THE TARGET FOR THE METADATA RECORDS.**

### 3.7 SEARCH ENGINE OF THE PORTAL

The Stored records are Indexed by an indexing mechanism. (See Figure 7, (8)) Indexing means that all words of all records are put in a long list and sorted in alphabetical order. See the example of Table 4 and Table 5. This is not exactly how a search engine works, but gives a good idea of what we mean.

```
<record>
  <header>
    <identifier>oai:DiVA.org:umu-117</identifier>
    <timestamp>2006-03-19</timestamp>
  </header>
  <metadata>
    <oai_dc:dc>
      <dc:title>klinisches und ätiologisches Spektrum und Langzeitverlauf</dc:title>
      <dc:creator>Schielke, Eva</dc:creator>
      <dc:type>dissertation</dc:type>
    </oai_dc:dc>
  </metadata>
</record>
```

```

<record>
  <header>
    <identifier>oai:HUMBOLT:diss-1542</identifier>
    <datestamp>1997-05-25</datestamp>
  </header>
  <metadata>
    <oai_dc:dc>
      <dc:title>Spektrum of Oil and Water </dc:title>
      <dc:creator>Stage, Jesper</dc:creator>
      <dc:type>text.thesis.doctoral</dc:type>
    </oai_dc:dc>
  </metadata>
</record>

```

**TABLE 4: EXAMPLE OF METADATA RECORDS**

Title	record id	author	type	date modified
ätiologisches	oai:DiVA.org:umu-117	Eva	dissertation	2006-03-19
ätiologisches	oai:DiVA.org:umu-117	Schielke	dissertation	2006-03-19
klinisches	oai:DiVA.org:umu-117	Eva	dissertation	2006-03-19
klinisches	oai:DiVA.org:umu-117	Schielke	dissertation	2006-03-19
Langzeitverlauf	oai:DiVA.org:umu-117	Eva	dissertation	2006-03-19
Langzeitverlauf	oai:DiVA.org:umu-117	Schielke	dissertation	2006-03-19
Oil	oai:HUMBOLT:diss-1542	Jesper	text.thesis.doctoral	1997-05-25
Oil	oai:HUMBOLT:diss-1542	Stage	text.thesis.doctoral	1997-05-25
Spektrum	oai:DiVA.org:umu-117	Eva	dissertation	2006-03-19
Spektrum	oai:HUMBOLT:diss-1542	Jesper	text.thesis.doctoral	1997-05-25
Spektrum	oai:DiVA.org:umu-117	Schielke	dissertation	2006-03-19
Spektrum	oai:HUMBOLT:diss-1542	Stage	text.thesis.doctoral	1997-05-25
und	oai:DiVA.org:umu-117	Eva	dissertation	2006-03-19
und	oai:DiVA.org:umu-117	Schielke	dissertation	2006-03-19
Water	oai:HUMBOLT:diss-1542	Jesper	text.thesis.doctoral	1997-05-25
Water	oai:HUMBOLT:diss-1542	Stage	text.thesis.doctoral	1997-05-25

**TABLE 5: EXAMPLE OF METADATA RECORDS TRANSFORMED INTO A SEARCH ENGINE INDEX**

When a portal visitor looks for a word, for example 'Spektrum', the search engine looks for this word in the list and finds 4 entries. Next, the search engine looks to see what stored records belong to this word, oai:DiVA.org:umu-117 and oai:HUMBOLT:diss-1542. The additional information is collected from the metadata store and returned to the visitor. (See also Figure 9, (10).)

### 3.8 WEB INTERFACE OF THE PORTAL

Figure 9 shows the web interface of the demonstrator.<sup>19</sup>



**FIGURE 8: WEB INTERFACE OF THE EUROPEAN E-THESES PORTAL - DEMONSTRATOR -**

The Stored records are not all Doctoral theses because some data providers do not offer sets specific for e-Theses to separate the Doctoral theses from the rest. There is more about this in section 6.9.1. As a result, the service provider has to harvest all records and gets a heterogeneous collection of all sorts of records. After harvesting and storing the metadata records, the e-thesis records are recognised by the type field (for example `dc:type='Doctoral thesis'`). To show the visitor only the results that are Doctoral theses,

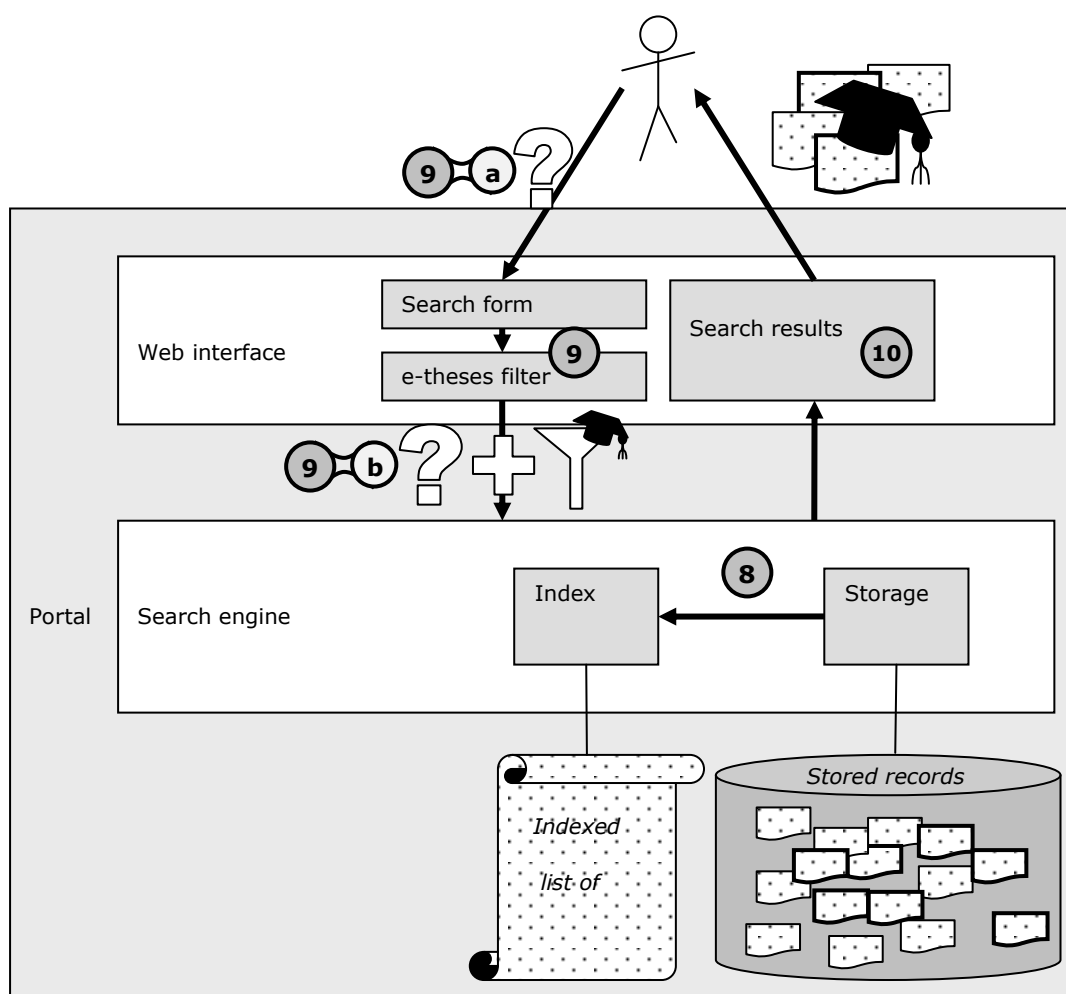
<sup>19</sup> The current link is <http://e-thesis.sharelab.cq2.org>

there must be some keywords added to the query of the visitor which make sure that only Doctoral theses are returned. This is what we call the e-theses filter (Figure 9, (9)).

In Figure 9 we see that the visitor is asking the Portal a search question (9-a) (in Figure 9 represented by the question mark) using the Search form. The Search form adds keywords that select only e-thesis metadata records (9-b) (in Figure 9 represented by the funnel with the square academic cap<sup>20</sup>) to the search question.

The search engine looks at what records are listed in the index which belong to the type of Doctoral thesis and contain the terms in the search question. The results are the metadata records that are doctoral theses which contain the requested terms in the search question. (See Figure 9, (10) and Figure 10.)

For example see Table 5: When the visitor enters the word "spektrum", the results will be only be of type 'dissertation' or 'text.dissertation.doctoral', as indicated in the index.



**FIGURE 9: THE E-THESIS FILTER IS SELECTING ONLY DOCTORAL THESES FOR THE VISITOR.**

<sup>20</sup> <http://en.wikipedia.org/wiki/Mortarboard>

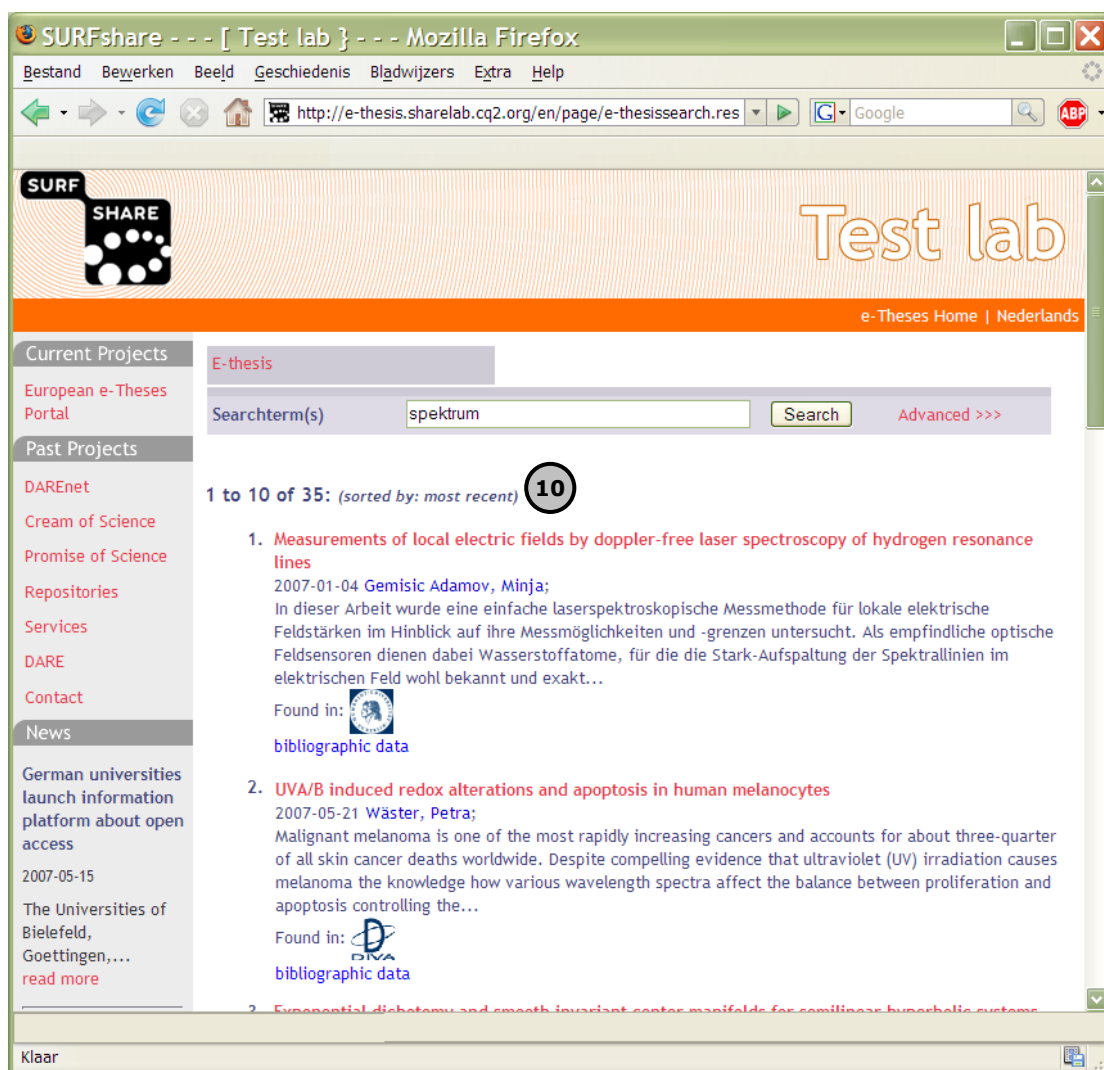


FIGURE 10: SEARCH RESULTS OF THE EUROPEAN E-THESES PORTAL - DEMONSTRATOR - (10)

## 4 THEORY: A MODEL OF INTEROPERABILITY

*"The lack of interoperability strongly implies that the described product or products were not designed with standardisation in mind."*<sup>21</sup>

The first step we made with the demonstrator shows that the interoperability basics work. These basics are harvesting repositories from different countries and presenting the data to a user. Yet, to be able go further than the interoperability basics, one has to agree upon and implement international standards on the use of metadata, the OAI-PMH protocol and a structure for representing complex documents.

Figure 11 shows a Model of Interoperability Levels to give an idea what levels of interoperability there are. At the moment the interoperability of repositories just reached level 2, we agreed on the names of the fields of Dublin Core. Now we have to work on Level 3; semantic interoperability. This paper provides some guidance to reach that level.

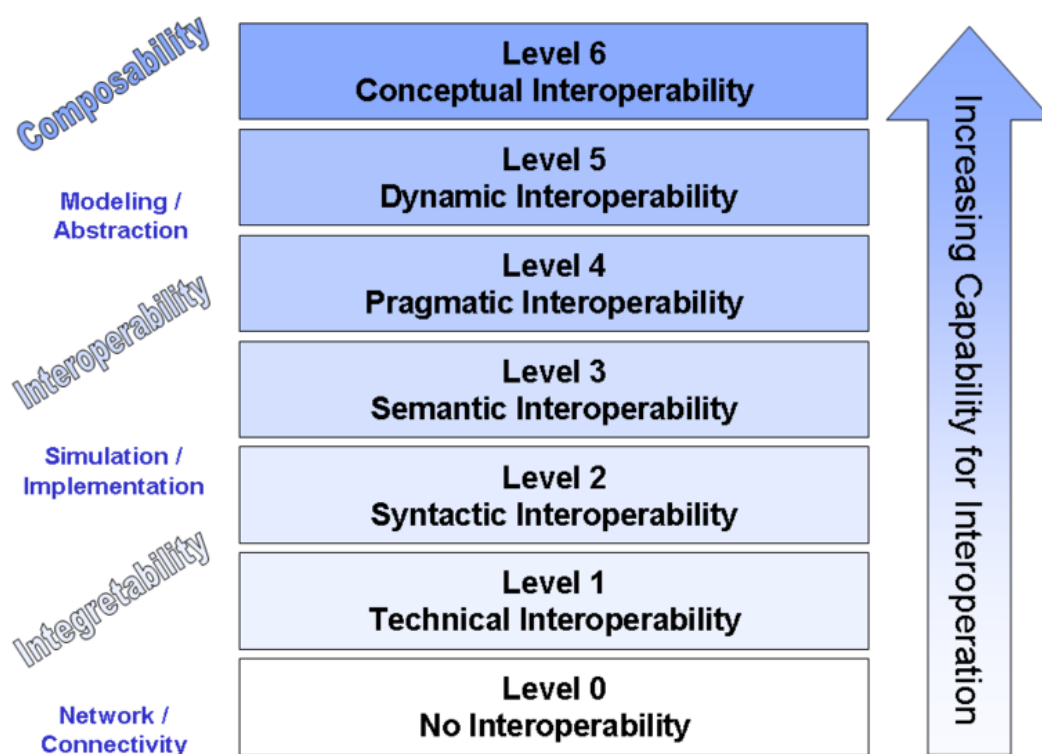


FIGURE 11: LEVELS OF CONCEPTUAL INTEROPERABILITY MODEL (LCIM)<sup>22</sup>

<sup>21</sup> <http://en.wikipedia.org/wiki/Interoperability>

Below, the descriptions of the first five levels of interoperability have been given:

**Level 0:** Stand-alone systems have No Interoperability.

**Level 1:** On the level of Technical Interoperability, a communication protocol exists for exchanging data between participating systems. On this level, a communication infrastructure is established allowing systems to exchange bits and bytes, and the underlying networks and protocols are unambiguously defined.

**Level 2:** The Syntactic Interoperability level introduces a common structure to exchange information; i.e., a common data format is applied. On this level, a common protocol to structure the data is used; the format of the information exchange is unambiguously defined.

**Level 3:** If a common information exchange reference model is used, the level of Semantic Interoperability is reached. On this level, the meaning of the data is shared; the content of the information exchange requests are unambiguously defined.

**Level 4:** Pragmatic Interoperability is reached when the interoperating systems are aware of the methods and procedures that each system is employing. In other words, the use of the data – or the context of its application – is understood by the participating systems; the context in which the information is exchanged is unambiguously defined.

With the HTTP and OAI-PMH protocols we can transfer metadata records (level 1). We even have agreed upon the metadata format (level 2), yet we haven't unambiguously defined the content (level 3). With this paper we try to provide recommendations to reach level 3.

The DRIVER guidelines are helping repositories to be interoperable in the levels 1, 2 and 3. Level 1 is about reducing issues concerning the HTTP and OAI-PMH protocol. Level 2 is about agreement on the OAI\_DC metadata format. And level 3 is about reducing issues with the content of the OAI\_DC format by making standardisation recommendations.

As an addition, making the OAI\_DC format interoperable for e-theses at level 3, this paper provides some recommendations in section 7.1.

To increase the interoperability to level 4, further study on e-theses specific formats have to be made.

## 4.1 HARVESTER

The harvester plays a crucial role in enabling technical interoperability. Many issues appear to happening during machine-machine communication (see chapter 6) and one might say that the underlying networks and protocols are not unambiguously defined. To ensure

---

<sup>22</sup> Conceptual Interoperability by Tolk A, Diallo SY, Turnitsa CD, Winters LS (2006) "Composable M&S Web Services for Net-centric Applications," Journal for Defense Modeling & Simulation (JDMS), Volume 3 Number 1, pp. 27-44, January 2006, see also :

[http://en.wikipedia.org/wiki/Levels\\_of\\_conceptual\\_interoperability](http://en.wikipedia.org/wiki/Levels_of_conceptual_interoperability)



interoperability, despite the issues, the harvester must be a little more tolerant on the strictly unambiguous definition. This is not recommended, but a practical necessity.

For service providers, a harvester quickscan has been made. In the annex section "harvester quickscan", a function list has been compiled about two harvesters SAHARA and PKP OAI Harvester<sup>23</sup>.

Both systems use different programming languages (Python and PHP) and have different architectures. Despite the differences, both systems are Open Source available, can handle different metadata formats and are tolerant of interoperability errors in repository XML output.

In PKP, a search engine is included and comes as a complete package. SAHARA is made for harvesting only, it can output to many targets (like a search engine of any choice, a database or a file system) and has an excellent administration web interface.

---

<sup>23</sup> More information about the PKP harvester: <http://pkp.sfu.ca/?q=harvester>

More information about SAHARA: <https://www.uitwisselplatform.nl/projects/sahara/> (see documentation)

## 5 PRACTICE: THE E-THESES DEMONSTRATOR DEPLOYMENT

*By far the best proof is experience.*

*- Sir Francis Bacon (1561 - 1626)*

When we started with the idea of the demonstrator, we wanted to test the interoperability and not system failures, etc. This was the reason we used a duplicate of DAREnet. The only difference was that we harvested other repositories. The DAREnet service has become a robust system that has been tested and improved for several years. The demonstrator was ready in a day. Collecting the baseURLs and testing was taking more time.

This system consists of separate pieces of software that work together - a front-end, a search engine, and a harvester. The front-end is custom-made for DAREnet and is mainly developed to communicate with the search engine. The search engine is released as an open source product under the name Meresco Core<sup>24</sup>, it has a high performance and is a good alternative to the commercial heavy-weight search engines. For the third part, a standalone harvester called Sahara<sup>25</sup> has been developed, also released as an open source product.

### 5.1 MERESCO METADATA MANAGEMENT (HARVEST, STORE, SEARCH)

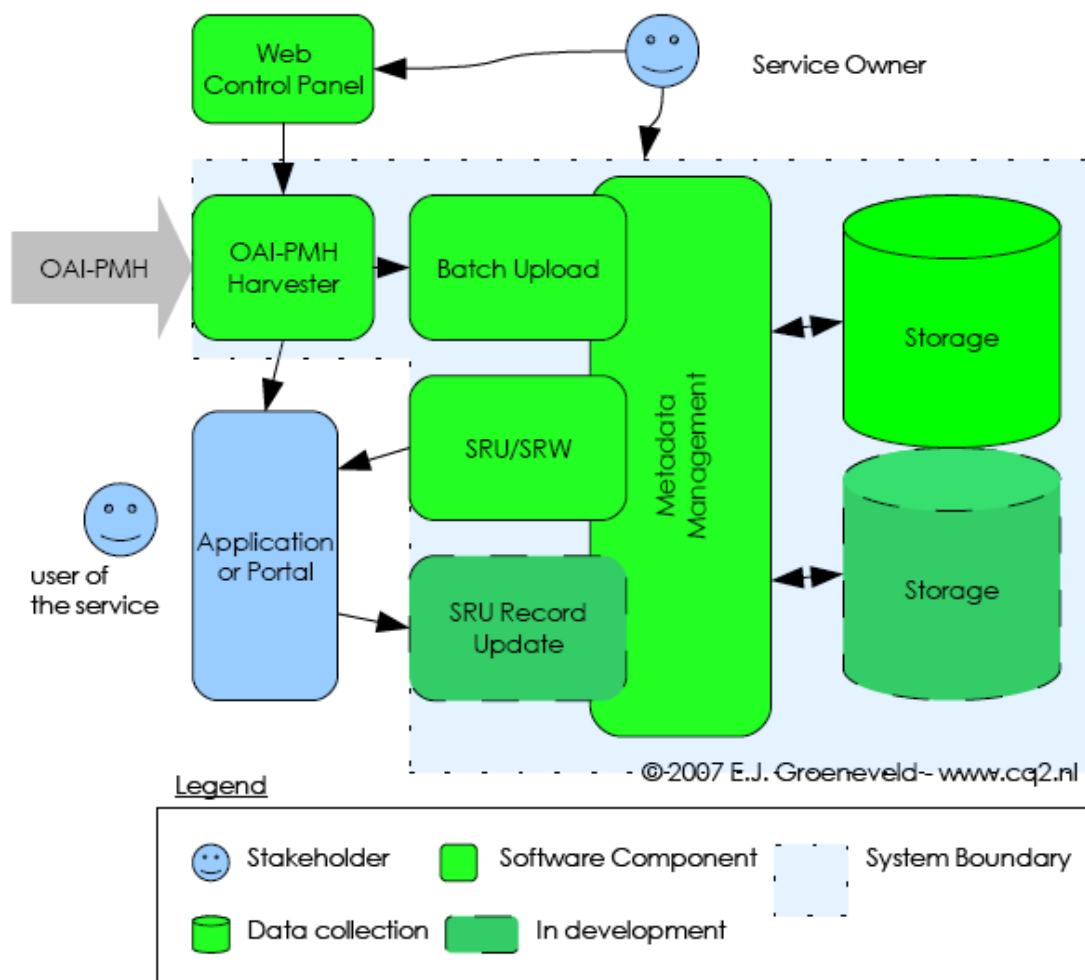
The Structural View outlines the software components that fulfil the processes of collecting the metadata, processing, combining, storing, selecting and serving. The Components for collecting (Harvester and its Control Panel) are separate applications, while others are plug-ins to the Metadata Management component. The Portal is assumed not to be part of the service, although it is very well possible that a Service Owner exploits a portal as well.

The following picture outlines the primary processes and functions of MERESCO. The green parts are components of MERESCO, the other parts are components of the context.

---

<sup>24</sup> Meresco Core is a stable, high performance, scalable Metadata Search Appliance based on Open Source technology. Meresco Core has low system requirements and yet is a good alternative to the commercial heavy-weight search engines. <https://www.uitwisselplatform.nl/projects/meresco/>

<sup>25</sup> Sahara is a standalone robust high performance OAI-harvester with web control interface <https://www.uitwisselplatform.nl/projects/sahara/>



**FIGURE 12: MERESCO STRUCTURAL VIEW<sup>26</sup> (GROENEVELD, MERESCO ARCHITECTURAL DESCRIPTION, 2007)**

The components are:

1. OAI-PMH Harvester (Sahara) – this component is a separate application that performs the collection of (meta)data by means of the OAI-PMH. It works on a set of grouped baseURLs. It pre-processes the data before it sends it to the Batch Upload.
2. Web Control Panel (Sahara) – a component of Sahara which administers the baseURLs along with additional meta-metadata needed for harvesting. All meta-metadata is also available via a REST<sup>27</sup> web service.
3. Batch Upload – a secure SOAP<sup>28</sup>-like streaming interface for uploading batches of documents into the Metadata Management component, of which it is a plug-in.

<sup>26</sup> See the document about the Meresco architectural description for more information.

<http://download.cq2.org/MERESCOArchitecturalDescription.pdf>

<sup>27</sup> Representational State Transfer (REST)

[http://en.wikipedia.org/wiki/Representational\\_State\\_Transfer](http://en.wikipedia.org/wiki/Representational_State_Transfer)

4. Metadata Management – processes and combines data from various sources such as the Batch Upload and the SRU Record Update<sup>29</sup> plug-ins into one or more collections. A document can be in one collection or spread over multiple collections as per the storage requirements of the Service Owner. Parts of documents may come from different sources at different times. Each document (or record) can be cross-walked, enriched, converted by functional rules (for the Crosswalk and a Normalisation plug-in). It maintains a flexible registry for supporting quick select functionality. All internal processing and data flow is configurable.
5. Store – provides scalable and reliable storage of documents in a way that allows for high speed retrieval and streaming (video).
6. SRW/SRU<sup>30</sup> – this is an example plug-in that standardises the communication between the portal and the metadata management. Other possibilities for plug-ins are RSS or OAI-PMH.
7. SRU Record Update (see note 29) – a real-time interface to add or update parts of a document. It can be used instead of the Batch Upload to build a complete document out of different parts (coalescing). It follows the Library of Congress SRU Record Update standard. The Service Owner builds his service by selecting and configuring the components mentioned above. Configuration of the Metadata Management system is done by connecting plug-ins into a suitable constellation. The result is a trigger driven data network.

---

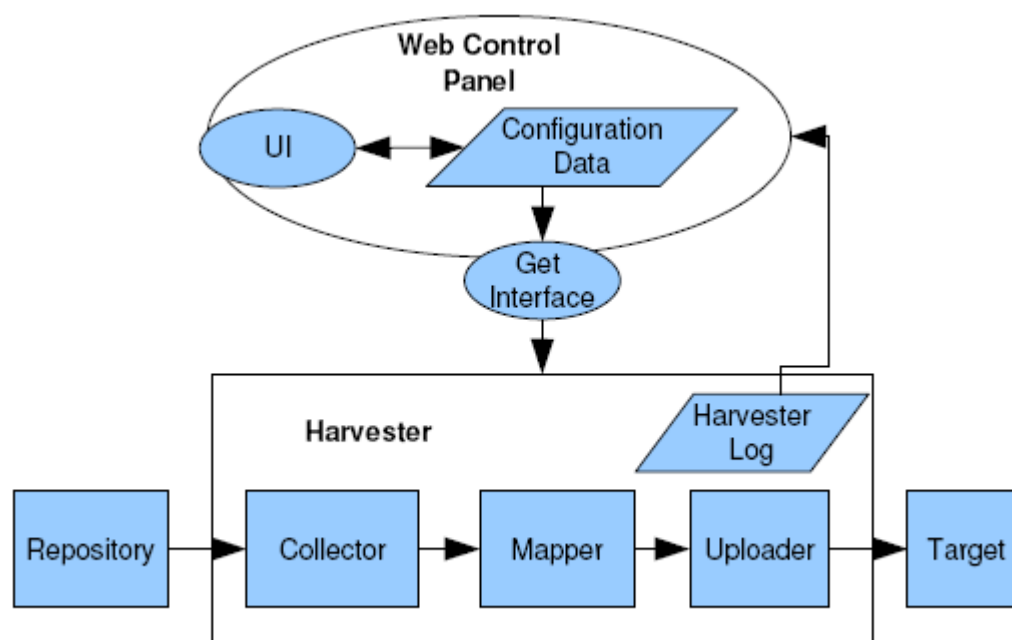
<sup>28</sup> Simple Object Access Protocol, or Service Oriented Architecture Protocol (SOAP)

<http://en.wikipedia.org/wiki/SOAP>

<sup>29</sup> Search/Retrieve via URL (SRU) record update <http://www.loc.gov/standards/sru/record-update/>

<sup>30</sup> Search/Retrieve Web service (SRW) [http://en.wikipedia.org/wiki/Search/Retrieve\\_Web\\_Service](http://en.wikipedia.org/wiki/Search/Retrieve_Web_Service)

## 5.2 SAHARA HARVESTER

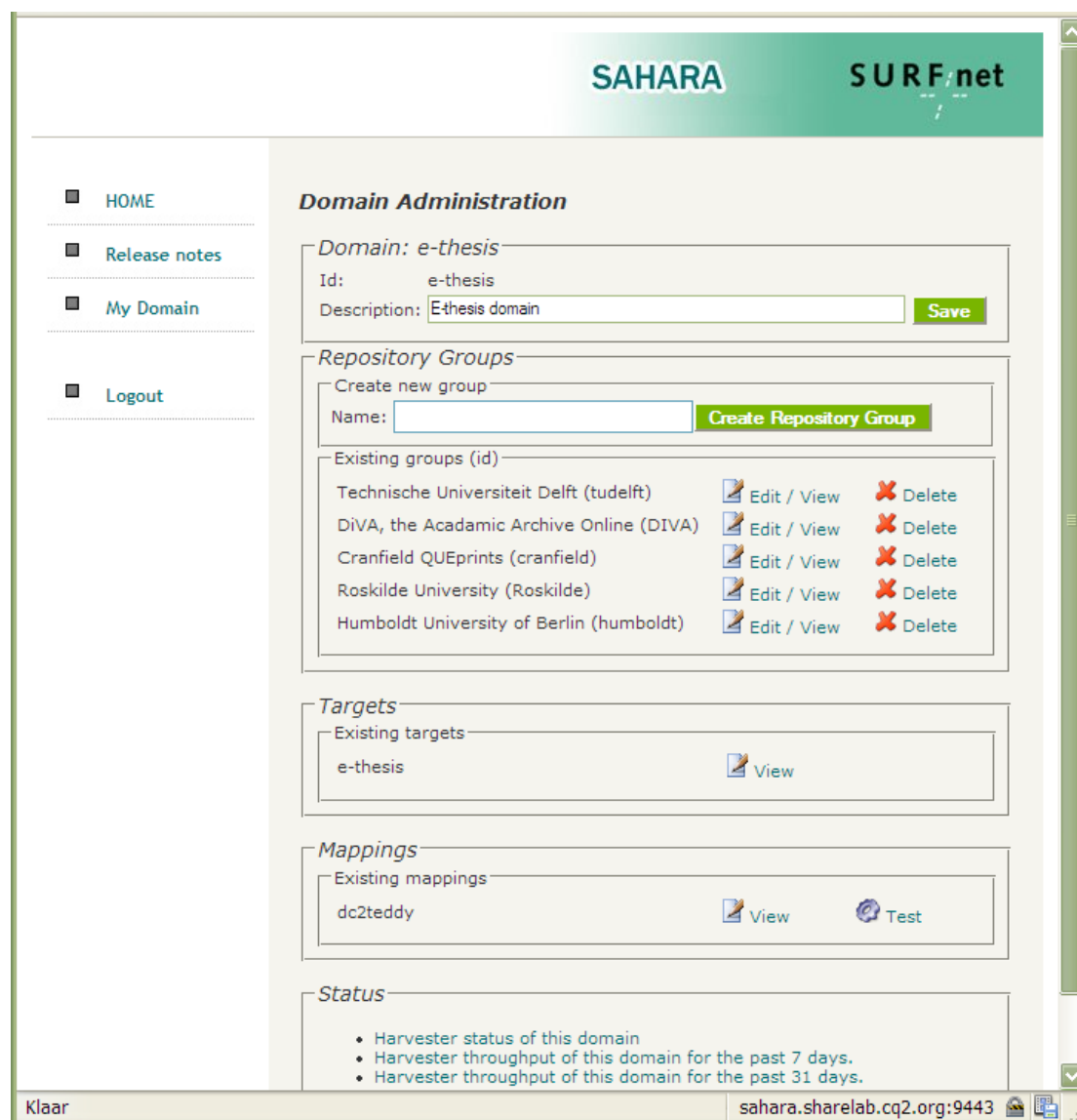


**FIGURE 13: SAHARA ARCHITECTURE**

The collector part of Sahara is the part that interacts with the repositories to retrieve the information stored in those repositories. The format of contents that the harvester retrieves must comply with the OAI-PMH protocol in order for the harvester to harvest the contents successfully.

The harvester is also capable of converting the retrieved data from any given structure into another (mapper). After a possible conversion, the harvester can upload the data to a designated target (uploader). In the demonstrator, this is the Meresco Metadata Management component.

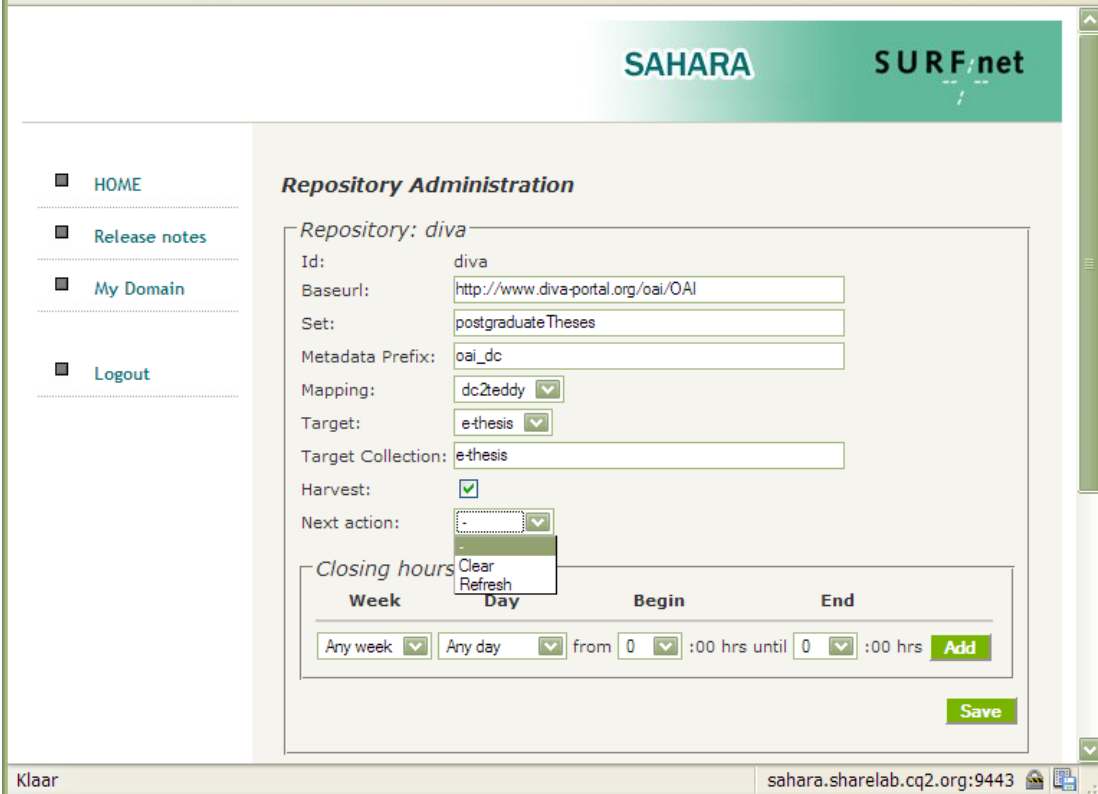
The web-control panel, as the name suggests, controls the behaviour of the harvester by means of a website. This ranges from where on the internet the harvester should contact the repositories to what the harvester should do with the retrieved data. The web-control panel provides the means to add, modify and remove repositories, to group them under a single name and to specify the way the harvester should treat the harvested contents as mentioned above. The web-control panel also provides means to grant access to certain repositories to users by placing them in domains and allowing the application administrator to link users to these domains.



**FIGURE 14: SCREENSHOT OF THE SAHARA WEB-INTERFACE: THE MAIN SCREEN SHOWS THE REPOSITORY GROUPS THAT ARE HARVESTED FOR THE E-THESES DEMONSTRATOR**

Besides providing an interface for humans, the web-control panel also provides an interface for other systems. This interface is the link between the web-control panel and the actual harvester. Using this interface, called SaharaGet, the harvester can retrieve the information it needs to harvest and process retrieved contents.

Sahara has been in production as of 2002 and, since then, additional features have been added to make sure that repositories get harvested, no matter what happens. It can handle repositories being off-line often, variations in metadata formats, expired resumption tokens, etc. Experience has shown that it takes more than OAI-PMH to get repositories harvested without constantly having to look after the process. This experience has been incorporated into Sahara.



**SAHARA** **SURFnet**

**Repository Administration**

*Repository: diva*

Id: diva

Baseurl:

Set:

Metadata Prefix:

Mapping:

Target:

Target Collection:

Harvest: ☒

Next action:

*Closing hours*

Week	Day	Begin	End
<input type="text" value="Any week"/>	<input type="text" value="Any day"/>	from <input type="text" value="0"/> :00 hrs until <input type="text" value="0"/> :00 hrs	<input type="text" value="0"/> :00 hrs

Klaar sahara.sharelab.cq2.org:9443

**FIGURE 15: SCREENSHOT OF THE SAHARA WEB-INTERFACE: HARVESTING INFORMATION IS FILLED INTO THE REPOSITORY ADMINISTRATION PAGE.**

## 6 PRACTICE: INTEROPERABILITY ISSUES AND RECOMMENDATIONS

*Experience teaches only the teachable.*

*- Aldous Huxley (1894 - 1963)*

In the drawings in Chapter 3 we have shown the flow of metadata information that travels from the repository to the user. That chapter shows a theoretical process that only exists in an ideal and perfect universe. Below, one will find the issues that we have encountered during the creation of the demonstrator portal which is based upon multiple repository systems.

To have some analogy with Chapter 3, we use mainly the same groupings of the subsections and follow the numbers in Figure 2.

### 6.1 COLLECTION OF THE REPOSITORY SYSTEMS

A repository, the data provider, has a collection filled with internal records containing an object and the description of the object. The description provides us with information about the object. This is what we call 'metadata'. The metadata are stored in the database in an arbitrary format designated by the repository system.

#### 6.1.1 GENERIC PROBLEM: THE EXTERNAL RECORD HAS A FINER GRANULARITY THEN THE INTERNAL RECORD

*The internal record might not have the fine granularity that the output format of an external record demands. This leads to unclear and ambiguous interpretations of the content of the external record.*

*The example in Table 6 shows that the information granularity of the internal format is poor and yet the external format is very rich. In this situation one does not know where to put the content. For example, the internal database of the repository has a field called 'Author name'. In this field the author name is written in one string: "Doe, John". When the repository wants to put this name in a rich format, it does not know where to put the string "Doe, John"; in the family name field, the maiden name or first name field, or all three? The repository internal format simply does not have the information granularity to be able to support a rich external format.*



Internal fields		External fields
Author name	Convert to or or	<ul style="list-style-type: none"> <li>Family name of the author</li> <li>Maiden name of the author</li> <li>First name of the author</li> </ul>
Format	Convert to or or	<ul style="list-style-type: none"> <li>MIME type of the document</li> <li>File size of the document</li> <li>MD5 Hash code of the document</li> </ul>
Identifier	Convert to or or or	<ul style="list-style-type: none"> <li>ISBN Identifier</li> <li>URN:NBN Identifier</li> <li>DOI Identifier</li> <li>File URL reference</li> </ul>

TABLE 6: EXAMPLE OF POOR INTERNAL FORMAT AGAINST RICH EXTERNAL FORMAT

### Recommendation for data provider:

Use a repository system that has a fine information granularity structure, and therefore is 'future-ready'. Further study has to be made to provide recommendations for internal information formats that are compatible for interoperable uses.

In the mean time, to create the highest level of interoperability, use the DRIVER<sup>31</sup> guidelines on specification for the external format in simple Dublin Core. The DRIVER guidelines are created from experiences and lessons learned on a European scale.

## 6.2 MAPPER OF THE REPOSITORY SYSTEMS

When an internal record in this arbitrary format is given to someone to read, this person might not be able to understand the meaning of each record field. Therefore by default, the mapper of the repository transforms the internal format to the simple Dublin Core format. Dublin Core is a metadata format that has a limited number of basic/generic fields.

### 6.2.1 GENERIC PROBLEM: THE INTERNAL RECORD HAS A FINER GRANULARITY THEN THE EXTERNAL RECORD, AS A RESULT CREATING MORE AMBIGUITY

*This ambiguity problem occurs when there is no room in the external format to express a specific concept. For example, in what field do the supervisor and the juror go when only the field 'contributor' is available? And how can a reader of the external format make the distinction between supervisor and juror when the extra information is dumbed-down<sup>32</sup> to 'contributor'? When there is no exact match for a concept in the internal format with a concept in the external format, this can*

<sup>31</sup> DRIVER is a project on Institutional Repositories (IR) in Europe. The DRIVER project has produces guidelines on interoperability. These guidelines come out of practice and lessons learned from other projects like DARE (NL) and DINI (DE). The guidelines are pragmatic in setup and the only one at this moment. I is a good source for solutions for general IR issues. For more information on DRIVER, see: <http://www.driver-support.eu>

<sup>32</sup> Dumbing-down: a process to fit the content of rich and fine granular metadata format into a less-rich metadata format. As a result, one will miss information which makes it harder to interpret.

lead to combinations and/or duplicates of fields, or fields in the external format that cannot be filled-in.

For example the internal format can define roles for a person like an author, chairman, supervisor, juror etc. Yet the external format only provides two roles, creator and contributor. The concept 'author' is close to the concept 'creator', the other roles have to be put in the concept 'contributor'.

Internal field name		External field name (sDC)
Role.author	Convert to	Creator
Role.editor	Convert to	Contributor
Role.supervisor	Convert to	Contributor
Role.promotor	Convert to	Contributor
Role.commission member	Convert to	Contributor

**TABLE 7: EXAMPLE OF THE DUMBING-DOWN PROCESS IN THE MAPPER. INFORMATION ABOUT DIFFERENT ROLES (EDITOR, SUPERVISOR, PROMOTER, COMMISSION MEMBER) IS DUMBED-DOWN IN SIMPLE DUBLIN CORE TO ONE CONCEPT (CONTRIBUTOR).**

### Recommendation for the data provider:

We recommend not to use Simple Dublin Core as an internal format. Use an internal metadata format with a fine granularity and high detail, to be able to provide in the (near) future richer metadata to service providers who require more detail.

For external mappings from a rich internal format like MARC21 to an external poor format like simple Dublin Core, we recommend to look for and use known mappings that are supported by active communities. These mappings from active communities have been tested, gone through the disambiguity process and brought into practice. An example of community agreements on crosswalks are crosswalks from the Eprints Application Profile to simple Dublin Core<sup>33</sup>.

## 6.2.2 GENERIC PROBLEM: SIMPLE DUBLIN CORE LACKS CONTENT GUIDELINES FOR INTEROPERABLE UTILISATION

*Simple Dublin Core is a metadata format developed in 1995, Dublin, Ohio, USA. The Core is "a set of semantics for Web-based resources [which] would be extremely useful for categorizing the Web for easier search and retrieval."*<sup>34</sup> This Core set consists of 15 fields to represent the basic semantics of a resource. Of course, this simple Core set does not include e-Thesis specific fields, but a greater problem is that the content of these semantic fields can be virtually anything - which makes it a weak format to set up interoperable services.

<sup>33</sup> Mapping from Eprints Application Profile to simple Dublin Core can be found at: [http://www.ukoln.ac.uk/repositories/digirep/index/Mapping\\_the\\_Eprints\\_Application\\_Profile\\_to\\_Simple\\_DC](http://www.ukoln.ac.uk/repositories/digirep/index/Mapping_the_Eprints_Application_Profile_to_Simple_DC)

Or an example for mappings with a Dspace repository system can be found at: [http://www.ukoln.ac.uk/repositories/digirep/index/Mapping\\_the\\_Eprints\\_Application\\_Profile\\_to\\_DSspace\\_metadata](http://www.ukoln.ac.uk/repositories/digirep/index/Mapping_the_Eprints_Application_Profile_to_DSpace_metadata)

<sup>34</sup> <http://dublincore.org/about/history/>

We give a few examples based on the two external records we have used earlier from DIVA and Humboldt (see Table 1 and Table 2). Below is a short list from the Humboldt record.

```
<dc:subject>Langzeitverlauf</dc:subject>
<dc:subject>Neuropsychologie</dc:subject>
<dc:subject>Magnetresonanztomographie (MRT)</dc:subject>

<dc:subject>long-term outcome</dc:subject>
<dc:subject>neuropsychology</dc:subject>
<dc:subject>magnetic resonance imaging (MRI)</dc:subject>

<dc:subject>YE 4500</dc:subject>

<dc:description>Akute Enzephalitiden treten überwiegend .....</dc:description>
<dc:description>Acute encephalitis occurs mainly sporadically .....</dc:description>

<dc:date>2001-11-06</dc:date>
<dc:date>2001-10-13</dc:date>

<dc:type>Text</dc:type>
<dc:type>dissertation</dc:type>

<dc:format>text/html</dc:format>

<dc:identifier>http://edoc.hu-berlin.de/habilitationen/schielke-eva-2001-11-06/HTML/index.html</dc:identifier>
<dc:format>application/pdf</dc:format>

<dc:identifier>http://edoc.hu-berlin.de/habilitationen/schielke-eva-2001-11-06/PDF/Schielke.pdf</dc:identifier>

<dc:language>eng</dc:language>
```

**TABLE 8: EXAMPLE OF AMBIGUITY IN SIMPLE DC FORMAT. (SAMPLE OF A HUMBOLDT RECORD)**

Firstly, have a look at the subject and description fields. The first fields are in German, the next in English. Only humans notice this difference by recognising the text codification. In simple DC there is no room to define the language of the metadata itself.

The last subject field is some sort of code. DC does not provide room to define what this code is and what it represents. For example, try yourself to recognise what the code TR458 and E901 is used for. E901 is a European Union approved food additive code and it represents Beeswax. TR458 could be a flight number, or a car registration number in South-Africa. Without any context a human has trouble recognising that this code is about food additives. Cognitive recognition depends on the context, knowledge about the existence of such code and other complex constructs. Imagine the trouble a machine must go through for recognition. To leave a clue for the machine how to interpret the code, some additional information should be provided. (See Xmetadiss or the E-Prints Application profile for such support)

The date fields have different dates. Not even a human can interpret what each date represents (date of publication, graduation, presentation etc.).

The language field contains a three letter code. We can guess it might be representing the English language encoded in the ISO639-2 standard, but we can never be sure. When we look at the DIVA external record in Table 1 we find, in the language field, the content 'en\_UK'. So, even when our best guess for one repository system is a good one, it might not count for the other.

*Simple Dublin Core does not provide guidelines for the content of the metadata fields in general that answer questions like: "what iso standard shall I use to encode the language field?" or "at what granularity shall I set the date field?"*

*More of these differences can be found in the Metadata analysis (see work package 2 in the annex).*

*To create an interoperable service on a European scale, with hundreds of repository systems, simple Dublin Core appears not to be the best choice without proper agreements on standardisation of the content.*

### **Recommendation for data provider:**

To provide metadata that has a high interoperability factor one has to use a common format and use common content encodings. We therefore recommend look at a project like DRIVER, which has experience and provides guidelines to prevent interoperability problems. We advise the use of simple DC recommendations made by the DRIVER project. As an addition for e-Theses, we advise following the recommendation described in section 6.2.3.1 "Ad 1. Recommendation to adapt simple Dublin Core metadata for e-theses".

This recommendation is based upon experiences explained in the metadata analysis in work package 2 (see the annex) and the work of experts at the Knowledge Exchange meeting in January 2007 who have recognised and taken this advice into account.

### **6.2.3 E-THESIS RELATED PROBLEM: SIMPLE DUBLIN CORE LACKS E-THESIS SPECIFIC EXPRESSIONS.**

*The simple Dublin Core format that is used has, in practice, a high level of ambiguity (like date type and contributor role) and is not able to express basic e-Theses concepts (like grade, and level).*

*In Table 10, one can see the features that are needed to describe e-theses properly for interoperable use. (More explanation will follow below.) Compare this with Table 9, the simple Dublin Core elements, and one will notice that some of the requirements match the simple Dublin Core elements, and some do not. In particular, the e-Thesis specific requirements, like degree name and degree level, cannot be met in simple Dublin Core.*

<b>simple Dublin Core elements</b>	
Label:	<b>Contributor</b>
Definition:	An entity responsible for making contributions to the resource.
Label:	<b>Coverage</b>
Definition:	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Label:	<b>Creator</b>
Definition:	An entity primarily responsible for making the resource.
Label:	<b>Date</b>
Definition:	A point or period of time associated with an event in the lifecycle of the resource.

Label:	<b>Description</b>
Definition:	An account of the resource.
Label:	<b>Format</b>
Definition:	The file format, physical medium, or dimensions of the resource.
Label:	<b>Identifier</b>
Definition:	An unambiguous reference to the resource within a given context.
Label:	<b>Language</b>
Definition:	A language of the resource.
Label:	<b>Publisher</b>
Definition:	An entity responsible for making the resource available.
Label:	<b>Relation</b>
Definition:	A related resource.
Label:	<b>Rights</b>
Definition:	Information about rights held in and over the resource.
Label:	<b>Source</b>
Definition:	The resource from which the described resource is derived.
Label:	<b>Subject</b>
Definition:	The topic of the resource.
Label:	<b>Title</b>
Definition:	A name given to the resource.
Label:	<b>Type</b>
Definition:	The nature or genre of the resource.

**TABLE 9: THE 15 SIMPLE DUBLIN CORE ELEMENTS****Background for data and service provider:**

*To understand what e-theses information we need, we have to work on a 'demand-driven' basis. This demand comes from the user who visits an interoperable e-theses service. This visitor wants to look for certain information. For example, this visitor is a scientist and likes to search for the latest e-theses within his field, to see when an e-theses has been published, look for the graduation date, title and name, see who granted the theses, show the promoter-apprentice network, etc. For this reason we have set up a functional specification for the demonstrator. Look at the Annex for Workpackage 1 for more information.*

Next, we have created a matrix with the Knowledge Exchange workshop participants<sup>35</sup> to define priorities of the features. In this report we define three levels of features: A. e-Thesis specific features; B. service specific features; and C. generic features. All these features have been put in a matrix with an implementation priority set to mandatory, required and nice to have.

		Mandatory	Highly recommended	Nice to have
Level	Feature description			
C. Generic	Title	✓		
	Author	✓		
	Abstract		✓	
	Language of the document	✓		
B. Service	Location of the resource	✓		
A. e-Thesis specific	A field that tells this metadata is about a Doctoral thesis	✓		
	A field that indicates who was supervising the author of the Doctoral thesis		✓	
	The date when the Doctoral thesis was published	✓		
	The name of the degree			✓
	The level of the degree			✓
	The country the degree was given, this in case of the cultural differences on the value of the degree level			✓
	The person or organisation who granted this degree			✓
	The discipline of the Doctoral thesis			✓
	The date to drop the embargo (if there is any)			✓

**TABLE 10: PRIORITY MATRIX OF E-THESIS FEATURES**

### Recommendations for Data providers:

At this point, when data providers want to join or be part of a high quality service, we recommend data providers should change and adapt their metadata for interoperable uses. The recommendations below are based on best practices from the e-Theses demonstrator participants, Knowledge Exchange participants, DRIVER and DARE members.

We have two recommendations:

- Ad 1. Adapt Dublin Core as much as possible for the use of Doctoral e-Theses.
- Ad 2. A recommendation to developers who create a generic format that fits the Academic Information domain, to incorporate specific elements for describing e-Theses.

<sup>35</sup> The priority list was made by e-thesis experts at the Knowledge Exchange workshop on Januari 2007 in Utrecht (the Netherlands).



### 6.2.3.1 Ad 1. RECOMMENDATION TO ADAPT SIMPLE DUBLIN CORE METADATA FOR E-THESES

At the Knowledge Exchange workshop the following recommended use of Dublin Core was made for Doctoral e-Theses. The recommendations in Table 11 below follow the DRIVER guidelines. The rule of thumb recommendations, given below, come from experiences in creating better interoperable metadata especially for the e-theses.

Level	Feature description	DC field with mandatory content = (...) or content encoding = {...}
Generic	Title	dc:title
	Author	dc:creator{bibliographic} <sup>36</sup>
	Abstract	dc:discription
	Language of the document	dc:language{iso639-1} <sup>37</sup>
Service	Location of the resource	dc:identifier{URL}
e-Thesis specific	A field that tells this metadata is about a Doctoral thesis	dc:type(Doctoral thesis)
	A field that indicates who was supervising the author of the Doctoral thesis	dc:contributor{bibliographic}
	The date when the Doctoral thesis was published	dc:date{ISO8601} <sup>38</sup>

TABLE 11: PRIORITY FEATURES WITH SIMPLE DUBLIN CORE

For the full DRIVER guidelines please go to the driver support website. <http://www.driver-support.eu/en/tech/index.html> .

The *Rule of thumb*, when using dc:type with the content 'Doctoral thesis', is that very close attention must be paid to following:

- The dc:date field always must contain the date of publication. (Use only one date field, more date fields will be considered ambiguous. DC has no room to specify other types of dates.)
- And the dc:contributor field always must contain the name of the supervisor. (Using contributor fields with names of other roles will be considered ambiguous. DC has no room to specify other contributor roles.)

<sup>36</sup> Bibliographic name encoding: Normal last name first inversion. (Lastname, Firstname) Example: 'Finnegan, James A.' or 'Pooh, Winnie The'

<sup>37</sup> ISO639-1 : two letter code, see [http://en.wikipedia.org/wiki/ISO\\_639-1](http://en.wikipedia.org/wiki/ISO_639-1)

<sup>38</sup> For ISO8601 see <http://www.iso.org/iso/en/prods-services/popstds/datesandtime.html>

- The rest of the fields should follow the DRIVER guidelines exactly. Please pay attention to the dc:language field that it is only encoded in iso639-1<sup>37</sup>. Also note that the dc:identifier is the only field that contains a URL that points to a full text thesis document or intermediate page with open access to the full text thesis document. The dc:date field must be ISO8601 (YYY-MM-DD). And the dc:creator and dc:contributor fields are formatted in "lastname, firstname" style.



### **6.2.3.2 Ad 2. GENERIC METADATA FORMAT FOR ACADEMIC INFORMATION DOMAIN, WITH E-THESES ELEMENTS INCORPORATED**

This recommendation is for Developers who work on a generic metadata format for the Academic Information Domain. The developers should take in mind that the Thesis, with it's context specific elements, is part of the Academic Information Domain. To research these context specific elements and how to represent these in a format, one should take a look at ETD specific formats like xMetaDiss, UKETD\_DC and ETD-MS. See the appendix section "Existing ETD specific formats" for more information about these formats. A list about ETD specific elements are presented in Table 10.

We recommend on the long run not to implement a specific ETD format. This recommendation has a strategically reason: stop the growth of institution specific metadata implementations and start creating uniformity by incorporating interoperable standards amongst all repositories in Europe. The rationale behind this is the following:

A generic format is more interesting for a broader public, and will generate a bigger mass of repositories. A large number of repositories who externalise a uniform rich and interoperable metadata format is more interesting for service providers to create services, then a small number of repositories supporting different variations of ETD specific formats.

One of the advantages of having a generic format for the Academic Information Domain is, that it will have a larger active community. With a large active community, the maintenance and updating of the format is therefore well supported. This support community is needed for example to create interoperability by eliminating ambiguous elements, and helping each other to use and implementing the format in practical terms.

The developers of repository software will incorporate this Generic Academic format directly into the software for off-the-shelf uses. There is no need for repository administrators to create ETD specific mappings themselves, which reduces errors in syntactic and semantic interpretations.

A candidate model for describing the Academic Information Domain is the "Scholarly Works Application Profile (SWAP)<sup>39</sup>". Humboldt University has a demo<sup>40</sup>

<sup>39</sup> <http://www.ukoln.ac.uk/repositories/digirep/index/SWAP>

<sup>40</sup> [Click this link to view the SWAP format in practice at the Humboldt University. \(OAI-GetRecord\)](#)



running of the SWAP package. And most important that future research needs to be made how this package can support ETD specific content and interoperability.

## 6.3 THE GATE OF REPOSITORY SYSTEMS

After mapping the fields from the arbitrary internal database fields to simple Dublin Core field records, this has to be presented to the outside-world. This will be done by a gate. This gate responds to requests (verbs) from the outside-world with an XML output.

### 6.3.1 GENERIC PROBLEM: THE OAI PROTOCOL V.2.0 IS NOT WELL SUPPORTED BY THE REPOSITORY

*When the repository has a gate that does not fully understand what to do with the OAI-PMH verbs that it receives, it might result in error messages.*

*This occurred, for example, in one of the DARE repositories after an update procedure for PHP<sup>41</sup> where some requests and responses malfunctioned.*

#### **Recommendation for data provider:**

After each server side modification, test also your OAI-PMH gate if it is still functioning as expected.

In the near future it will be possible to test the output of the Repository system automatically by a testing tool. This testing tool is currently under development in the DRIVER project. This testing tool is called a "validator" that tests the level of OAI-PMH 2.0 and DRIVER compliance. The validator will be available at the end of the summer of 2007.

It is also recommended to join a community for the repository software package you are using. In this community you can get tips, best practices, updates, patches, malfunction and security warnings. Such a support community is currently also being developed at the DRIVER support site: [www.driver-support.eu](http://www.driver-support.eu).

## 6.4 OAI-PMH

OAI-PMH is a request-response mechanism described by the Open Archives Initiative. The requests for information are made by harvesters; the responses are given by repository systems. To make the request and response predictable, a communication protocol has been developed which is called Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH). When the Harvester requests a list of records, the Repository responds with an OAI-PMH styled XML file. This OAI-PMH XML file contains records in the simple Dublin Core metadata format by default. XML-encoding is used because it is machine readable, and therefore, can easily be processed by the service provider.

---

<sup>41</sup> **PHP** is a reflective programming language originally designed for producing [dynamic web pages](http://en.wikipedia.org/wiki/PHP). See <http://en.wikipedia.org/wiki/PHP>

*Generic Problem: invalid XML. The XML standard is very strict in its interpretations. A little mistake in the XML can make it impossible for a service provider to interpret the data. In practice these mistakes can be generally divided into three sub problems*

1. Wrong XML encoding scheme (see 6.4.1),
2. not well-formed XML structure (see 6.4.2)
3. and un-recognisable XML structure (see 6.4.3).

#### 6.4.1 GENERIC PROBLEM: WRONG XML ENCODING SCHEME

*At the top of the XML file there is the encoding scheme presented (e.g. UTF-8, see example in Table 12 ). The encoding scheme tells the harvester how to interpret the bits used to create the characters in the rest of the XML file.*

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"> .....
```

TABLE 12: XML ENCODING



#### Recommendation for data provider:

Make sure the XML output is UTF-8 compliant. To get help, use the [UTF8conditioner](#) from Cornell University to make UTF-8 encoded output.

#### 6.4.2 GENERIC PROBLEM: NOT WELL-FORMED XML STRUCTURE – URL ENCODING

*This problem often occurs when parts of HTML are copy-pasted into the database. As a result, the parts of HTML appear in the XML of the external metadata record. This messes-up the XML structure.*

*An example of messed-up XML is shown below in Table 13. It shows a <p> tag in the <dc:description> element. Without a closing tag like </p> the XML is not well-formed. (Even with a closing tag, the XML is not valid. See 6.4.3)*

```
<oai_dc:dc>
  <dc:title>Mixing Oil and Water : Studies of the Namibian Economy</dc:title>
  <dc:creator>Stage, Jesper</dc:creator>
  <dc:description> <b>p</b> This thesis consists of four papers studying economic aspects of natural resource
and environmental management in Namibia.Paper [I] analyses changes in Namibian energy use patterns
between 1980 and 1998. The study finds that, unlike their counterparts in many other developing
countries where energy use has been studied, Namibian energy users appear to have been quite flexible in
changing to energy-saving technologies and to technologies using different energy sources altogether.
</dc:description>
</oai_dc:dc>
```

TABLE 13: EXAMPLE OF NOT WELL-FORMED XML, LOOK AT THE BOLD <P> WITHOUT A CLOSING TAG

*When a 'wrong' character is in the XML file, the automatic machine processing of the XML file will stop. Further harvesting is impossible because the harvester is not able to get past this problem and read the Resumption Token. The resumption token (see 6.4.4) is a ticket from the repository at the end of the XML file to read the next batch of records.*

*The XML basic structure is one of a nested hierarchy. When elements are not well nested, this is called 'not well-formed'. According to the service provider, the data provider probably speaks a known language, but with a strange kind of grammar. The automatic interpretation of such a XML file is not possible.*

### Recommendation for data provider:

These errors occur when data, mostly HTML code from the database, is placed in the XML without transcoding to URL codification.

Make sure the XML output is encoded in URL or ISO-8859-1. The Latin character set can be represented in a hexadecimal and a decimal way. For example a space in hexadecimal representation is written as %20 , in decimal representation &#32;.

The text in the example above apparently has been copied from an HTML page and pasted with HTML tags in the database. The content in the <dc:description> element should not contain tags like in Table 13. The tags should be removed or transformed to a hexadecimal or decimal URL codification. A well-formed XML should look like:

```
<oai_dc:dc>
  <dc:title>Mixing Oil and Water : Studies of the Namibian Economy</dc:title>
  <dc:creator>Stage, Jesper</dc:creator>
  <dc:description> &lt;p&gt; This thesis consists of four papers studying economic aspects of natural
resource and environmental management in Namibia.Paper [I] analyses changes in Namibian energy use
patterns between 1980 and 1998. The study finds that, unlike their counterparts in many other developing
countries where energy use has been studied, Namibian energy users appear to have been quite flexible in
changing to energy-saving technologies and to technologies using different energy sources altogether.
</dc:description>
</oai_dc:dc>
```

**TABLE 14: EXAMPLE OF WELL-FORMED XML, WHERE THE <P> TAG IN THE <DC:DESCRIPTION> ELEMENT IS ENCODED TO &LT;P&GT; . THESE ARE ENTITIES USED IN THE ISO LATIN 1 (ALSO KNOWN AS ISO 8859-1) CHARACTER SET.<sup>42</sup>**

## 6.4.3 GENERIC PROBLEM: XML VALIDATION, NOT RECOGNISED XML STRUCTURES.

*This problem also occurs when copy-pasting HTML. The problem is like speaking English and, all of a sudden, in the middle of the sentence you hear strange words from another language.*

*Using the same example with the HTML <p> tag, according to the oai\_dc schema child elements are not expected in the <dc:description> element.*

```
<oai_dc:dc>
  <dc:title>Mixing Oil and Water : Studies of the Namibian Economy</dc:title>
  <dc:creator>Stage, Jesper</dc:creator>
  <dc:description> <p> This thesis consists of four papers studying economic aspects of natural resource
and environmental management in Namibia.Paper [I] analyses changes in Namibian energy use patterns
between 1980 and 1998. The study finds that, unlike their counterparts in many other developing
countries where energy use has been studied, Namibian energy users appear to have been quite flexible in
changing to energy-saving technologies and to technologies using different energy sources altogether.
```

<sup>42</sup> URL encoding to ISO 8859-1 character sets, see:

[http://www.astro.washington.edu/owen/ROFM\\_CGI/Documentation/SpecialChars.html](http://www.astro.washington.edu/owen/ROFM_CGI/Documentation/SpecialChars.html)

```
</p>
</dc:description>
</oai_dc:dc>
```

**TABLE 15: EXAMPLE OF AN INVALID XML. THE OAI\_DC SCHEMA DOES NOT EXPECT THIS CHILD ELEMENT IN THE DESCRIPTION ELEMENT**

*The XML file is not correctly structured in the way the OAI-PMH scheme prescribes. This means that the communication between the data and service provider does not occur according to the agreed communication protocol. According to the service provider, the data provider probably speaks a strange language. The automatic interpretation of such a XML file is not possible.*

#### **Recommendation for data provider:**

Make regular checks to the XML output. Use the [W3C validator](#) to check the output of the repository.

#### **Recommendation for the service provider:**

Use a harvester system that is not too strict. For the demonstrator, we used a harvester called [SAHARA](#) which can handle this kind of flexibility. The DRIVER project is also planning to use SAHARA.<sup>43</sup>

### **6.4.4 GENERIC PROBLEM: VERY SHORT LIFESPAN OF RESUMPTION TOKEN.**

*In our demonstrator project we found that we harvested only 50 metadata records from the DIVA repository, despite the fact that the DIVA repository system contains over 5000 e-theses! The problem occurs because the harvester didn't return in time to harvest the next batch. This has to do with the lifespan of the resumption token. Hypothesis 1: the harvester is too slow to return to the repository, or Hypothesis 2: the repository system drops the resumption token faster than is reasonable. In practice, most of the time, hypothesis 2 is the case.*

*The resumption token is a ticket the harvester gets from the repository at the end of each XML file. (See Table 3.)*

```
...
<dc:title>
Next-generation extreme ultraviolet lithographic projection systems
</dc:title>
</oai_dc:dc>
</metadata>
</record>
<b>resumptionToken cursor="0" completeListSize="7881">!!!oai_dc!50</resumptionToken>
</ListRecords>
</OAI-PMH>
```

**TABLE 16: EXAMPLE OF A RESUMPTION TOKEN (SEE BOLD TEXT: !!!OAI\_DC!50)**

*The XML file contains a fixed number of records, called a batch. In the example, above, the batch size contains of 50 records. For the harvester to receive the next batch of records it gives the resumption token from the previous batch to the repository system.*

<sup>43</sup> More about the SAHARA harvester can be found in the harvester Quickscan in the Annex.

*To reduce the workload for the DIVA repository system, the harvester continues to harvest other repositories first. When the harvester returns to the DIVA repository system, it finds out the ticket it received is past its expiry date and is not valid anymore.*

*The DIVA repository system does not continue with the next batch, but keeps on sending the first batch forever. The result is that the service provider only has the first load of 50 records.*

#### **Recommendation for data provider:**

1. Keep the resumption token alive/valid for at least 24 hours.
2. The recommended batch size will be between 100 and 200 records. These recommendations are in line with the DRIVER guidelines.

### **6.4.5 GENERIC PROBLEM: BASEURL INTERPRETATION.**

*The repository manager provides the service provider with a URL saying it is the baseURL. When the service provider uses this to harvest, it results in errors. After closer inspection, the given baseURL appears to be a human readable HTML page, instead of the machine readable repository gate.*

*This occurs when the a repository is quickly setup out of the box and the focus is on filling the repository. The results show up nicely on the webpage, not knowing there is also a separate URL for the OAI-PMH gateway.*

#### **Recommendation for data provider:**

Make sure to provide service providers with the URL that can handle OAI-PMH requests and delivers OAI-PMH XML as output.

Test your baseURL by putting ?verb=Identify after the URL. Look at source view of the result. When it returns in OAI-PMH structured XML with information about the repository, then this is the baseURL.

### **6.4.6 GENERIC PROBLEM: FIREWALL BLOCKING THE HARVESTER.**

*The service provider is not able to access the repository gate to harvest. In our experience with the DARE project, this was most confusing for the repository managers. They could simply access the repository gate because they were accessing the repository system on the university campus with a computer also on the university campus.*

#### **Recommendation for data provider:**

When you don't know if the repository is behind a firewall, try to reach it from another ISP. When using a firewall, ask the for the IP-address of the harvester machine from the service providers you want to be harvested from, and add this IP-address to the trusted list.

#### **Recommendation for service provider:**

If the repository is not reachable with the given baseURL, the URL either does not exist anymore, or is probably firewalled. Inform the repository manager you want to harvest the repository and provide the IP-address of the harvester.

#### 6.4.7 GENERIC PROBLEM: CHANGING IDENTIFIERS & UPDATING DATESTAMPS

*The portal service depends on reliable identifiers. The oai identifiers that were once harvested are used to update records and to get the real record from the live repository. For example the oai identifier [oai:diggy.ruc.dk:1800/2081](#) was harvested four months ago. Now, using this identifier results in errors. After further research the identifiers appear to have changed to a different format: [oai:rudar.ruc.dk:1800/2081](#)*

##### **Recommendation for data provider:**

This problem has two solutions. The first one is a short term solution; the second is a robust and permanent solution.

Solution 1: When updating a record by, for example, changing the identifier, the datestamp must also be updated! This is the only way a harvester can recognise and process any changes in the repository in an incremental harvest.

Solution 2: Do not use identifiers that change. We recommend the use of Persistent Identifiers. These identifiers must be independent of system names like 'diggy' or 'rudar' and names of organisations like 'ruc' and 'diva'. In Germany, the use of the URN:NBN namespace<sup>44</sup> is common practice for Persistent Identifiers. In the Netherlands, there are plans to use the URN:NBN namespace as a Persistent Identifier for digital information assets.

##### **Recommendation for service provider:**

In the short term; carry out a full re-harvest on a regular basis because incremental harvesting is not as reliable. Some repositories update the datestamp on some modifications, others not. In the long term, carrying out frequent full harvests is not advised for the following reasons: A full harvest is an intensive process for the repository, the bulk of data to be transferred uses a lot of the network capacity, and the number of service providers that harvest is increasing, demanding more from the repository system.

### 6.5 THE COLLECTOR OF THE HARVESTER FROM THE PORTAL

As the Collector module handles the communication in OAI-PMH language with the Repository gate, it sends the received records to the Mapper.

---

<sup>44</sup> See <http://www.persistent-identifier.de/?lang=en> for more information.

### 6.5.1 E-THESIS PROBLEM: COLLECTED METADATA FEATURES DO NOT MEET SERVICE FEATURES.

*The metadata format that is collected depends on the service a service provider wants to provide at the front-end. For example, the simple Dublin Core metadata format is not rich enough to express concepts like degree name, embargo, etc. that are e-theses specific.*

#### Recommendations for service provider:

When the Dublin Core Metadata format is not rich enough to provide the information needed on the front-end, the service provider has to consider collecting metadata from another format. However, ETD-specific formats are not widely used by default. One can see there is a tension between quantity and quality a service is based on.

In section 6.2.3 about “background for data and service providers”, the functional specification has been elaborated for a Service provider who wants to start a Doctoral e-Theses portal. The outcome is that simple Dublin Core can be sufficient when it is adapted. For the time being, this is also our recommendation.

When a service provider wants to increase the detail, he has to harvest a richer format. But also the quantity is important to create a real service. As said before in section 6.2.3.2, future research has to be made on a generic format with ETD specific elements.

On the short term, to get richer metadata, the service provider could harvest the not so widely used ETD specific metadata formats like ETD-MS, UKETD\_DC and Xmetadiss. In other words, be prepared to harvest repositories that offer one of these metadata formats. In practice this means to have crosswalk templates ready in your Mapper component to transform these formats into the format your search engine is using.

#### Recommendations for Data providers

Our recommendation is to adapt the Dublin Core for joining basic e-Theses services (see section 6.2.3.1). When have adapted, your repository has become interoperable, which makes it easier for service providers to use your metadata. This is the first step in sharing the work of these young scientists.

On the short term, to join an e-Thesis service that needs richer metadata then specified in section 6.2.3, one could use e-Theses specific formats like ETD-MS, UKETD\_DC or XMetaDiss. Further study could be made on the interoperability of these ETD formats. As said before, in section 6.2.3.2, specific ETD formats are not wide spread and hard to promote. Knowledge from these ETD formats should be used to create a generic format for the Academic Information Domain with ETD elements.

## 6.6 MAPPER OF THE HARVESTER FROM THE PORTAL

For each specific repository the Mapper module can transform and normalise the received metadata record to an internal metadata record. This internal metadata format can differ from the received metadata format. However, in the case of our demonstrator, we receive

simple Dublin Core metadata from the repositories and map it to simple Dublin Core without any normalisation.

### **6.6.1 GENERIC PROBLEM: COLLECTING METADATA WITHOUT NORMALISATION RESULTS IN AMBIGUOUS DATA, WHICH MAKES IT HARDER TO CREATE INTEROPERABILITY**

*An example from our demonstrator experience is that Masters theses are found in the results for the Cranfield repository where there should be only Doctoral theses. See section 6.9 for the complete example.*

*Mapping external records without any normalisation will result in metadata fields with unclear and undefined content. Practically every repository delivers the content of their metadata fields their own way (see the Metadata analysis in the Annex). This has two causes:*

- *For repository administrators it is the easiest way to make an external record to put the content of the database field directly into the output format.*
- *The second cause for undefined ambiguous data is that Dublin Core format does not prescribe how to standardise its content.*

#### **Recommendations for data providers**

See section 6.9.1.2 “Ad Sub-problem 2: mapping and up-scaling” for recommendations to improve interoperability.

#### **Recommendations for service providers**

The best solution is to create national proxy services. These proxies are gateways that on one side harvest the metadata from local Institutional Repositories, and on the other side have an OAI-PMH gate that can be used for service providers to harvest the national proxy. The advantage of a national proxy is that it can centrally normalise the metadata, and deliver unambiguous metadata to service providers. It is recommended that the metadata is normalised using the DRIVER guidelines. (See section 6.9.1.2.)

The short term solution is to normalise the simple Dublin Core metadata yourself. You can do this by looking at the content of the fields and try to convert the differences between repositories to one standard (the DRIVER standard is recommended). For example, the content of the language field in one repository contains the text “English”, in the other repository they use for the English language the text “en\_UK”. Convert all variations to the text “en”, this is a two letter description of the language (ISO639-1).

This has to be done for every type of field of every repository, which is a labour intensive enterprise when offering a high quality service. This is a short term solution, the long term solution will be that repositories offer standardised content that can be easily used for interoperable services.



## 6.7 UPLOADER OF THE HARVESTER FROM THE PORTAL

The Uploader module sends the mapped records to a specific target. This target can be storage in the search engine, a file system etc.

*There are no problems in this section regarding e-theses or interoperability.*

Information about the harvester quickscan can be found in the annex.

## 6.8 SEARCH ENGINE OF THE PORTAL

The stored records are being indexed for higher search performance.

*There are no problems in this section regarding e-theses or interoperability.*

## 6.9 E-THESES FILTER OF THE WEB INTERFACE FROM THE PORTAL

In this demonstrator, the visitor of the portal uses the search interface to find the Doctoral theses he/she needs. Because the harvester collects more than just Doctoral theses, we have added a query, which functions as a filter, together with the visitor query to narrow the search result.

### 6.9.1 E-THESIS PROBLEM: FILTER IS NOT ACCURATE ENOUGH

*For example, masters theses were also harvested from the Cranfield repository, because we didn't harvest separate sets, didn't normalise or make crosswalks. These Nine masters theses are found when 'msc' is entered in the search field. (See Figure 16.) Further explanation of this example can be found in 6.9.1.1.*

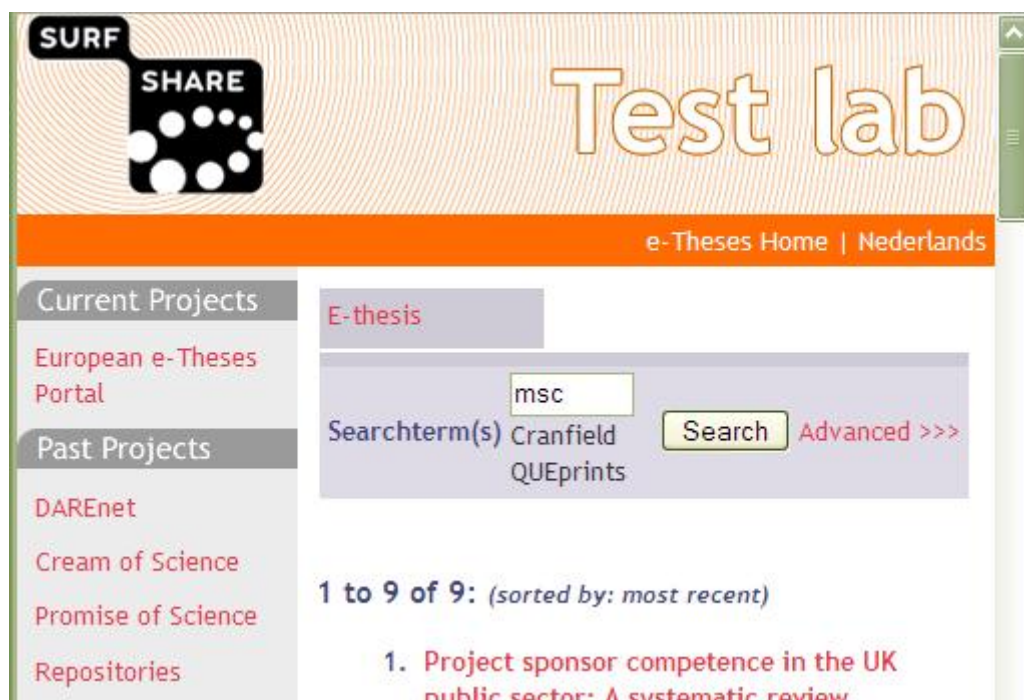


FIGURE 16: EXAMPLE OF MASTERS THESES FOUND IN DOCTORAL THESIS PORTAL.

*This problem is broken down into 3 sub-problems:*

1. *There is something wrong with the filter in the web interface. (see 6.9.1.1) The filter supports too many values for defining a doctoral thesis.*
2. *The normalisation of the Mapper in the harvester of the portal is not turned on. (see 6.9.1.2) This results in different values in the dc:type field and causes the filter to support all values.*
3. *The Collector of the harvester collects a lot more records than just Doctoral theses. (see 6.9.1.3) This causes the service provider to filter on a metadata type (e.g. dc:type). When the value of the type is not standardised, this requires the use of normalisations and filtering implementation which will cause the sub-problems drawn above.*

*All three sub-problems occur in our situation. Sub-problem 1 would not occur when either of the sub-problems, 2 or 3, is solved. Below, we look further into each of the issues.*

### **6.9.1.1 AD SUB-PROBLEM 1: FRONT-END FILTER**

*Sub-problem 1 occurs because the filter recognises too many values in the type field which should identify a Doctoral thesis. See Table 17 for the filter rule in words.*

Show only the records where the field dc:type contains only the content ("text.thesis.doctoral" or "Doctoral" or "Doctoral thesis" or "dissertation")

**TABLE 17: THE FILTER RULE USED TO FILTER OUT DOCTORAL THESES FROM THE REST.**

*In the metadata field of Cranfield the term "dissertation" appears in the dc:type field for both Masters and Doctoral theses. (See the bold text in Table 18 and Table 19.) According to the filter (see Table 17), not only the value "doctoral" (which appears in Cranfield repository) is recognised as a Doctoral thesis but also the value "dissertation" (which appears in Roskilde repository). As one looks closely, the value "dissertation" is also used for the Masters degree in Cranfield, which results in both Masters and Doctoral theses from Cranfield passing through the filter.*

```
<dc:title>Project sponsor competence in the UK public sector: A systematic review</dc:title>
<dc:type>Thesis or dissertation</dc:type>
<dc:type>Masters</dc:type>
<dc:type>MSc</dc:type>
```

**TABLE 18: EXAMPLE OF A CRANFIELD RECORD WITH A MASTERS THESIS, WHICH PASSES THROUGH THE 'DISSERTATION' FILTER**

```
<dc:title>Configuring in high velocity error sensitive circumstances: A grounded study</dc:title>
<dc:type>Thesis or dissertation</dc:type>
<dc:type>Doctoral</dc:type>
<dc:type>PhD</dc:type>
```

**TABLE 19: EXAMPLE OF A CRANFIELD RECORD WITH A DOCTORAL THESIS, WHICH PASSES THROUGH THE 'DISSERTATION' AND 'DOCTORAL' FILTER**

```
<dc:title>Essays in risk</dc:title>
```

```
<dc:type>Text</dc:type>
```

```
<dc:type>dissertation</dc:type>
```

**TABLE 20: EXAMPLE OF HUMBOLDT RECORD WITH A DOCTORAL THESIS, WHICH PASSES THROUGH THE 'DISSERTATION' FILTER**

```
<dc:title>Ældreomsorg i et pædagogisk perspektiv</dc:title>
```

```
<dc:type>Dissertation</dc:type>
```

**TABLE 21: EXAMPLE OF A ROSKILDE RECORD WITH A DOCTORAL THESIS, WHICH PASSES THROUGH THE 'DISSERTATION' FILTER**

The **first solution** might be to exclude the value "dissertation" in the filter, but then no Doctoral theses from Roskilde would pass through the filter. (See Table 20 and Table 21)

A **second solution** could be an improved filter that makes a combination between a repository name together with a dc:type value, like the one in Table 22.

When DIVA Use filter: dc:type="text.thesis.doctoral"

When Cranfield Use filter: dc:type="Doctoral"

When TU Delft Use filter: dc:type="Doctoral thesis"

When Humbolt Use filter: dc:type="dissertation"

When Roskilde Use filter: dc:type="dissertation"

**TABLE 22: BETTER FRONT-END FILTER, BIND DATA PROVIDER WITH APPROPRIATE DC:TYPE FILTER**

A **third solution** would be normalisation of the dc:type value for a Doctoral thesis per repository by the service provider ( see 6.9.1.2).

A **fourth solution** would be standardisation of the type value for a Doctoral thesis by the repositories (see 6.9.1.2).

A **fifth solution** would be to harvest only from sets that contain Doctoral theses (see 6.9.1.3).

A **sixth solution** would be to ensure that the word 'doctoral' always explicitly appears, either alone or in combination with 'thesis' or 'dissertation', in one of the type fields, and also that 'thesis' or 'dissertation' always appears in one of the type fields. However, this not according to the DRIVER guidelines that should make interoperability easier, and therefore not recommended.

### Recommendation for service provider:

A front-end filter can only work when the data is normalised. To get normalised metadata can be done in two ways:

1. Map the raw metadata yourself (see Figure 6). However this is a labour intensive job for a service provider all by himself (see 6.9.1.2).
2. A better approach is to get the metadata only from repositories that use interoperable metadata standards. (Like the DRIVER guidelines).

### **Recommendation for data provider:**

For better interoperability repositories should be using the same guidelines. Our recommendation is to use the DRIVER guidelines. Services can be setup more quickly, thus knowledge sharing will increase. For defining Doctoral theses, at least use the term "Doctoral thesis" in one of the dc:type fields, this is part of the DRIVER metadata vocabulary (See also section 6.9.1.2.)

#### **6.9.1.2 Ad SUB-PROBLEM 2: MAPPING AND UP-SCALING**

*We can narrow this problem down by normalising all written variations of the concept 'Doctoral thesis' to one term. The Mapper module in the harvester can transform for every single repository the variations to one term 'Doctoral thesis' (see Figure 6) by applying a repository specific mapping profile. As a result the front-end filter has only to pass through one term the field dc:type with the content 'Doctoral thesis'.*

*When every repository (r) uses different values for one concept/document type (t), leaves the service provider with  $r \times t$  variations of document types to normalise.*

*For example, in a demonstrator with 5 repositories it is not much work to create 5 repository specific mapping profiles. However making mapping profiles for every repository on a production-ready service with 500+ repositories will be a nightmare.*

*If there isn't any agreement between the service and data-provider about standards and stability, the data provider can change something without notifying the services that rely on the data-provider. As a result any mapping or data processing at the service provider will become unreliable.*

*The real solution is to use the mappers at the repository side to create a uniform, interoperable external format.*

### **Recommendation for data providers:**

We recommend using, at least, the value "Doctoral thesis" in the dc:type field, which is part of the DRIVER metadata vocabulary.

Use a known format like DC, and use known guidelines to make the metadata really interoperable. The DRIVER guidelines are recommended for the interoperable standardisation of metadata. To be more specific on using the DRIVER guidelines for e-Thesis, see 6.2.3.1 'Ad 1. Recommendation to adapt simple Dublin Core metadata for e-theses'. For using e-theses specific metadata format look at 6.2.3.2 'Ad 2. Generic metadata format for Academic Information'. The interoperable usability on these specific formats has to be studied further.

### **Recommendation for service providers:**

To prevent filter problems on doctoral theses, harvest sets that contain only doctoral theses.

### 6.9.1.3 Ad SUB-PROBLEM 3: HETEROGENEOUS COLLECTIONS

*In practice every repository can define a set and its content in its own flavour. A result of this is that not all repositories have placed only Doctoral theses into one set. This means, for example, that what the service provider harvests is a mixture of all kinds of theses - Masters theses, Bachelor theses and Doctoral theses. Other repositories have defined no sets at all, which means that all types of documents (articles, monographs, theses, etc.) are harvested.*

*The problem of these sets containing all sorts of documents is that the service provider relies on the quality of the metadata. When the metadata from a repository is of poor quality or does not conform an international standard, than this data will become less useful.*

*It would be easier for the service provider to harvest from a set where he knows exactly what it contains, for example, a set with only Doctoral theses, where the full text document is freely available without access restrictions and no embargo's. This will increase the quality and reliability for the information the service providers can provide.*

#### Recommendation for data provider

To join a high quality Doctoral e-Thesis portal, we recommend using a set specifically for Doctoral e-Theses and communicating this to the service provider. The criteria for this set is i.) that it only contains records are truly open access, ii.) have no access limitations, iii.) have no embargo, and where the full text is downloadable by anyone. The DRIVER guidelines prescribe to create a specific DRIVER set that contains records that match the DRIVER Open Access criteria. This recommendation adds the ETD specific 'no embargo records' criterion.

An advantage for data providers of placing the Doctoral records in a homogeneous Doctoral e-Thesis set is that it can be managed more precisely by the data provider. For example, to be part of a high quality service, the Doctoral e-Thesis set only contains records that are truly open access, have no access limitations, have no embargo, and where the full text is downloadable by anyone. This can be managed by the repository administrator.

It is very likely a record can be in one or more sets. For example in a Doctoral e-Thesis set, a heterogeneous theses set, and in a set containing all publications about a bio-medical subject. (See Figure 17.)

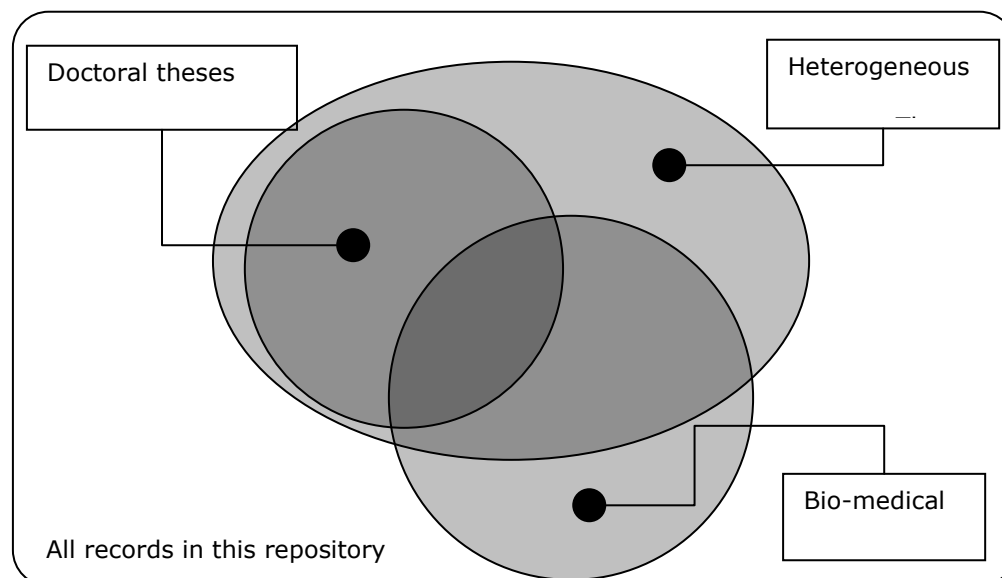


FIGURE 17: A RECORD CAN BE PART OF MORE THAN ONE SET.

### Recommendation for service provider

The service provider relies on the data provider how the metadata is delivered. The delivery can be ambiguous as well. To create a high quality service, make clear to the data provider what you want to receive. In your context we recommend the following criteria:

- 1) Receive theses with a Doctoral degree. (A Doctoral degree can be considered as a quality label for theses.)
- 2) The metadata is DRIVER compliant. (A syntactic and semantic interoperable standard is needed to prevent ambiguity.)
- 3) The harvested records are truly open access. (This means that the records and the full text records have no access limitations, like logins, toll gates, embargo, and campus fire walls.)

## 6.10 RESULTS OF THE WEB INTERFACE FROM THE PORTAL

The search engine generates a set of results matching the requested search query.

### 6.10.1 E-THESIS PROBLEM: CULTURAL DIFFERENCES CREATE AMBIGUOUS METADATA, AMBIGUOUS METADATA CREATES AMBIGUOUS SEARCH RESULTS

*For example: When sorting the search results by date, one should get a list with the newest thesis on top. In theory everything looks fine, but if we look closer to the concept 'newest thesis', we can ask the question: "What kind of date is used to base our list on?" When we look closer in the demonstrator to search for an answer, we find in one record three date fields. (see Table 23). It is unclear what every date might represent. The date of graduation, publication, offence of the thesis, etc. These is still no answer...*

```
<oai_dc:dc >
  <dc:title>Mixing Oil and Water : Studies of the Namibian Economy</dc:title>
  <dc:creator>Stage, Jesper</dc:creator>
  <dc:date>2003-12-02</dc:date>
  <dc:date>2003-12-25</dc:date>
  <dc:date>2004-02-04</dc:date>
  ...
</oai_dc:dc>
```

**TABLE 23: EXAMPLE OF THREE METADATA FIELDS, WITHOUT REPRESENTING WHAT EACH FIELD MEANS**

**Dates & educational system:** Even when we have figured-out in what order to interpret the dates in Table 23 for one repository, this interpretation can be different for another repository. This might be the cause of the under laying organisational or educational processes that differ per country or organisation. For example in Germany people first graduate, and then publish their work month's later. In England this is the other way around. This will result that in Germany the newest date represents the publication, and in England the newest date represents the graduation.

**The language & archival preferences:** When we look at the metadata of the several repositories in the demonstrator project, we can see that that the language in the metadata differs is used differently in each repository. For example the Cranfield repository describes its metadata in English (which happens to be also the native language), Humboldt uses both German (native) and English to describe their metadata, the TUDelft and Roskilde use English or the native language, the style can differ per record.

**The roles in the Theses context:** In the context of a thesis there are different roles involved by the graduation of a student. All these roles have made some contribution to the thesis. And to make it worse these roles differ in name and function per country or University. For example do the words 'supervisor' and 'promotor' refer to the same role (person with a function)? And is the function of a juror the same in the Netherlands as in Sweden? Does this person make the same contribution in the Theses process? This information definitely cannot be put into simple Dublin Core where we only have one field: 'Contributor'.

As we saw before in 6.2.1, when using simple Dublin Core it is not possible to determine the role of the contributor. When the service provider wants to set up a service that, for example, can visualise a network of master-pupil relationships, he most likely uses the supervisor and author concepts. When the additional role information is dumbed-down to only the concept of contributor, this is not possible anymore.

**Weight of the degree:** One can have the title name PhD, and the Doctoral degree, but does it count as much in every country? For example, in England it takes two years to write a thesis and graduate, in Germany it takes four years. Does the German Doctoral degree count more then an English one, or do the English work harder?

It is clear that the Bologna process has not yet made formalisations for the Doctoral degrees as it has done for the Master and Bachelor degrees. It is very hard to create metadata to describe these differences.



## Recommendation for service provider

To create a basic interoperable service we recommend to harvest from repositories who have standardised their metadata using the DRIVER guidelines and the e-theses adoptions recommended in this report. (see top-left quadrant of Table 24)

	Poor formats	Rich formats
Quality standards	Simple DC with DRIVER guidelines & ETD recommendations	A Generic Academic format,  (and some ETD specific formats)
Quantity standards	Simple Dublin Core (without standardisation)	A Generic Academic format

**TABLE 24: A QUADRANT ON POOR VS RICH METADATA AND STANDARDS THAT PROVIDE THE SERVICE PROVIDER WITH QUALITY METADATA, AND FORMATS THAT PROVIDE A LOT OF METADATA RECORDS**

The table above provides an overview of metadata formats that are either poor or rich in their information detail and quality standards to create interoperability or quantity standards to create mass.

If you want to create a rich interoperable service, look at right side of the quadrant. You will find a Generic Academic format that unambiguously incorporates cultural differences and that also can create mass (lots of repositories use this format). However, this format has to be developed. On the other hand there are ETD specific formats you could use. However, further study needs to be made whether these formats comply to standards that make interoperability possible.

### 6.10.2 GENERIC PROBLEM: ACCESS TO REPOSITORY OR FULL TEXT DOCUMENT RESTRICTED BY EMBARGO

*We have also experienced a situation whereby access to the repository intermediate page was possible, but downloading the PDF file was blocked because of embargo restrictions. This raised many questions from DAREnet users, who said: "how come the papers say DAREnet is so great by offering open access to document, but, in reality, I am not able to access this chapter of this Doctoral thesis. For me, you offer a bad and unreliable service". Of course, we corrected this by informing the data providers of the need to exclude records with an embargo.*

*The service providers assume that they provide their users/visitors with a 100% open access service, but in practice it is not. There are limitations out of control of the service provider, and often data providers are unaware of these limitations.*



### **Recommendation for service provider:**

Check not only the metadata you've harvested, but also go to the link provided in the metadata and try to retrieve the full text document. A method of excluding these embargoed documents in your high quality service is to make sure your data providers don't offer them in the set(s) you harvest.

Most of the time, the link in the metadata will refer not to a full text document directly, but will refer to an intermediate page of the repository. One can consider harvesting documents in the DIDL<sup>45</sup> document; this format is recommended by DRIVER to create compound documents. This allows the repository to define the resource locations independently of the metadata. The DIDL is a wrapper and allows to insert more than one metadata format. The advantage is that the service provider can have direct access to the document. Users visiting your website can download the document with one click, which is much user friendlier than making them detour by pointing to an intermediate page.

### **Recommendation for data provider:**

When the data provider offers a set, he can control and fine tune which documents are within this set and which not. To join a high quality e-Theses service it is recommended this set contains only Doctoral theses and the documents are freely available and have no access restrictions. And one important thing, the Doctoral theses with an embargo on one or more documents<sup>46</sup> are excluded from this set. This prevents a lot of user questions at the service provider side. A practical implementation, an embargo controller, has been developed at the University of Leiden (NL) (See Annex VI).

For creating compound documents it is recommended to use the DIDL document. This is a DRIVER recommendation. This format allows the repository to define the resource locations independently of the metadata. The DIDL is a wrapper and allows to insert more than one metadata format.

## **6.10.3 GENERIC PROBLEM: ACCESS TO REPOSITORY SYSTEM ONLY FROM UNIVERSITY CAMPUS**

*For example in the DAREnet project a student called us that he was not able to access the repository at home, yet he was able to access the repository when on the university campus.*

*In this example the harvester had access to the metadata, this had been indexed by the service provider and made available to the world via the darenet.nl portal. The user, finding the result, was redirected to the repository. This repository blocked all users accessing the repository that were not originating from the*

<sup>45</sup> DIDL example: [http://arno.unimaas.nl/oai/dare.cgi?verb=ListRecords&metadataPrefix=dare\\_didl](http://arno.unimaas.nl/oai/dare.cgi?verb=ListRecords&metadataPrefix=dare_didl)

<sup>46</sup> Most of the time e-theses appear in a compound construct. Every chapter is placed in a separate file. In the above example, a doctoral thesis with an embargo means that, for example, chapters 4 and 5 are not accessible until 3 years after the publication of the thesis. Chapters 1, 2, 3, 6 and 7 are available.

*university campus. The repository administrators, who were testing the repository on the University campus, had no problems accessing the repository.*

#### **Recommendation for service provider:**

Check that what you harvest is also accessible. Create an easy way to invite users to contact you when they encounter a problem. They are your 'eyes and ears'. Communicate with your partner repository about possible access restrictions.

#### **Recommendation for data provider:**

Test your repository from outside the University campus! And look critically to the access limitations of your repository. Make clear with your service provider the access limitations you want to provide.

### **6.10.4 GENERIC PROBLEM: CERTIFICATES AND USER EXPERIENCE**

*Some of the repository systems, like DSpace, use a Transport Layer Security (TLS)<sup>47</sup> by default to prevent eaves dropping when transferring data. For this, the user must go through a process of exchanging and manually accepting digital certificates. These extra manual actions reduce the browser experience, and TLS is not really required for exchanging 'open access' documents.*

#### **Recommendation for data provider:**

From the perspective of the data provider, the more secure their service is the better. From the user point of view, the definition of a good service is where he doesn't have to go through much trouble to get what he wants (download the file). Our recommendation is to use the http protocol without digital certificates and increase user-friendly access.

---

<sup>47</sup> Transport Layer Security (TLS) and its predecessor, Secure Sockets Layer (SSL), are cryptographic protocols which provide secure communications on the Internet:  
[http://en.wikipedia.org/wiki/Secure\\_Sockets\\_Layer](http://en.wikipedia.org/wiki/Secure_Sockets_Layer)

## 7 SUMMARY AND CONCLUSIONS

*Men acquire a particular quality by constantly acting a particular way... you become just by performing just actions, temperate by performing temperate actions, brave by performing brave actions.*

*Aristotle (384 BC - 322 BC)*

This section describes a summary of recommendations for data and service providers. In this summary you can find references to parts of the report for more elaborate information.

### 7.1 GENERIC RECOMMENDATIONS RELATED TO REPOSITORIES

the language used in the metadata fields differs per repository; sometimes all fields are in English; or fields are both in the local language and in English; or fields are only in the local language. It even differs per record. Furthermore, the quality of the data presented differs. The adagio 'garbage in is garbage out' is very much applicable for these kind of search services. Better validation of the data at the repository-side is needed. A further issue is the semantic and syntactic differences in metadata between repositories, which means that the format and content of the information exchange requests are not unambiguously defined. Further standardisation is recommended and references are made to the Guideline developed by the European DRIVER project.

An issue of a generic nature is also that the representation of a complex or compound structure of the thesis or enhanced (multimedia) publication. To create this structure we recommend to use the DIDL document meta-structure. This is a MPEG-21 standard that is flexible enough to support multiple purposes, does not rely on metadata formats and is self-descriptive. DRIVER provides the specifications in the guidelines.

#### 7.1.1 QUESTIONS AND ANSWERS IN A WIDER IR PERSPECTIVE

How to integrate richer metadata, is document type a good choice to base services on, how to get open access full text ETD without an embargo, compatibility, subject classification.

**Q: How can a repository take care of syntactic and semantic interoperability in metadata?**

A: Use commonly used formats and encoding schemes to be more interoperable.

Our recommendation is to use DRIVER guidelines to increase the interoperability on syntactic and semantic level. DRIVER defines a common use of simple Dublin Core metadata and suggests in vocabularies for types and languages. (see 6.2.3)

A: For data providers to be interoperable, to join basic Doctoral e-Theses services, we recommend following:

- At the 1<sup>st</sup> level of interoperability (Technical; protocol), use OAI-PMH for networking and transport.
- At the 2<sup>nd</sup> level of interoperability (Syntax; format), use the simple Dublin Core (OAI\_DC) format.
- At the 3<sup>rd</sup> level of interoperability (Semantics; content), use the DRIVER guidelines<sup>48</sup> with an addition of the e-Theses specific recommendations found in this document at 6.2.3.1.

**Q: Where does a repository have to look for to prevent technical errors makes data unreadable/accessible for service providers?**

A: Repositories should check the validity of their output using the DRIVER validator. Soon available at [www.driver-support.eu](http://www.driver-support.eu) (see 6.3 and 6.4)

A: Repositories should check the availability of the IR outside the University campus. (see 6.10.3)

**Q: Is it possible for service providers to harvest repositories that still have minor technical issues?**

A: Yes, but is important to inform the involved repositories. Service providers could use harvesters with some flexibility. Possible candidates are SAHARA and the PKP-harvester. (4.1)

**Q: Is it possible to create compound documents, but still use the interoperable advantage of simple Dublin Core?**

A: DRIVER recommends MPEG21-DIDL. A DIDL document is an XML format that describes the meta-structure that can wrap the academic work. This is "object oriented" and can include more then one document, video, dataset etc., but also more then one metadata format. All these instances can be part of the academic work, like an e-thesis. A DIDL document can be harvested, and the metadata format inside can be DRIVER compliant. (see: 6.10.2)

## 7.2 E-THESES SPECIFIC RECOMMENDATIONS

To be able to harvest *doctoral* theses, the service provider needs to be able to filter on this document type. Up to now there is no commonly agreed format, which makes semantic interoperability possible. It is recommended to distinguish between the various types of

<sup>48</sup> For DRIVER guidelines look at <http://www.driver-support.eu>

For more information about the e-Theses specific metadata formats (ETD-MS, UKETD\_DC and XMetaDiss), in the context of Doctoral e-Theses interoperability, look at Workpackage 2 "Metadata analysis, Chapter 2".

Information about ETD-MS see: <http://www.ndltd.org/standards/metadata/current.html>

Information about UKETD\_DC see: [http://ethostoolkit.rqu.ac.uk/?page\\_id=72](http://ethostoolkit.rqu.ac.uk/?page_id=72)

Information about Xmetadis see: <http://www.d-nb.de/eng/standards/xmetadiss/xmetadiss.htm>

Information about how to create interoperability with simple Dublin Core for e-Theses, see section 6.2.3.1 "Ad 1. Recommendation to adapt simple Dublin Core metadata" Also see workpackage 5 and 6 about issues and recommendations.

theses in the Dublin Core format "dc:type" and use the following qualifications: 'Bachelor thesis', 'Master thesis', 'Doctoral thesis'. Furthermore, there is a need to standardise on the date field, as various dates may be referred to (date of publication; date of graduation; starting date of the research etc). We recommend to use in the Dublin Core metadata field "dc:date" the date of publication of the doctoral e-theses. A last e-theses specific issue is related to the metadata field "contributor". For a doctoral thesis one could distinguish various 'contributors', like juror, committee member, referee, etc. We recommend to use the contributor field in Dublin Core for the person who supervised the thesis.

### 7.2.1 QUESTIONS AND ANSWERS ON HARVESTING E-THESES

**Q: Is it possible to get open access full text ETD's without an embargo?**

A: It is recommended for data providers, when they want to join a high quality ETD service, to offer a set that contains only open access full text ETD's without an embargo. The advantage for the data provider is that one has more control over the documents that are harvested. (see 6.9.1.3, 6.10.2) A practical implementation, an embargo controller, has been developed at the University of Leiden (NL) (See Annex VI).

A: Service providers harvesting such a set with open access full text ETD's without an embargo from a reliable source have improved control on the quality of the data and can guarantee quality criteria to the end-user. Checking for embargo's, accessibility, non-toll gated and full text availability can best be done at the data provider. (see 6.9.1.3, 6.10.2) At the University of Leiden (NL) an embargo controller, has been developed. This tool check's if the document has an embargo.

### 7.2.2 QUESTIONS AND ANSWERS ON METADATA QUALITY FOR E-THESES

When creating a high quality Doctoral e-Thesis service, the quality of the metadata is the most important component. When, for example, the encoding of the content of metadata of every repository is different, we can say that the metadata quality for interoperable utilisation is low. This is a result of underlying syntactic and semantic differences, which means that the format and content of the information exchange requests are not unambiguously defined. When this is the case, one could also use Google to find one's way through an ambiguous information space.

**Q: How can data providers provide metadata in a way that it is ready for interoperable uses?**

A: To be interoperable at the level of Semantics, one has to increase the quality of the metadata. At the basics we recommend to use the DRIVER guidelines<sup>49</sup>. (see 6.2.3.1, 6.6.1, 6.9.1.1 and 6.9.1.2).

---

<sup>49</sup> For DRIVER guidelines look at <http://www.driver-support.eu>

For more information about the e-Theses specific metadata formats (ETD-MS, UKETD\_DC and XMetaDiss), in the context of Doctoral e-Theses interoperability, look at Workpackage 2 "Metadata analysis, Chapter 2".

Information about ETD-MS see: <http://www.ndltd.org/standards/metadata/current.html>

Information about UKETD\_DC see: [http://ethostoolkit.rgu.ac.uk/?page\\_id=72](http://ethostoolkit.rgu.ac.uk/?page_id=72)

Information about Xmetadis see: <http://www.d-nb.de/eng/standards/xmetadiss/xmetadiss.htm>

In addition for e-Theses we recommend the following, which can be found in this document at 6.2.3.1.:

Level	Feature description	DC field with mandatory content = (...) or content encoding = {...}
Generic	Title	dc:title
	Author	dc:creator{bibliographic} <sup>50</sup>
	Abstract	dc:description
	Language of the document	dc:language{iso639-1} <sup>51</sup>
Service	Location of the resource	dc:identifier{URL}
e-Thesis specific	A field that tells this metadata is about a Doctoral thesis	dc:type(Doctoral thesis)
	A field that indicates who was supervising the author of the Doctoral thesis	dc:contributor{bibliographic}
	The date when the Doctoral thesis was published	dc:date{ISO8601} <sup>52</sup>

**TABLE 25: PRIORITY FEATURES WITH SIMPLE DUBLIN CORE**

Use in the dc:type field the classification for theses the terms "Doctoral thesis", "Master thesis" and "Bachelor thesis"

The *Rule of thumb*, when using dc:type with the content 'Doctoral thesis', is that very close attention must be paid to following:

- The dc:date field always must contain the date of publication. (Use only one date field, more date fields will be considered ambiguous. DC has no room to specify other types of dates.)
- The dc:contributor field always must contain the name of the supervisor. (Using contributor fields with names of other roles will be considered ambiguous. DC has no room to specify other contributor roles.)
- The dc:creator and dc:contributor fields are formatted in "lastname, firstname" style.
- Pay attention to the dc:language field that it is only encoded in iso639-1<sup>37</sup>.
- Also note that the dc:identifier is the only field that contains a URL that points to a full text thesis document or intermediate page with open access to the full text thesis document.

Information about how to create interoperability with simple Dublin Core for e-Theses, see section 6.2.3.1 "Ad 1. Recommendation to adapt simple Dublin Core metadata" Also see workpackage 5 and 6 about issues and recommendations.

<sup>50</sup> Bibliographic name encoding: Normal last name first inversion. (Lastname, Firstname)  
Example: 'Finnegan, James A.' or 'Pooh, Winnie The'

<sup>51</sup> ISO639-1 : two letter code, see [http://en.wikipedia.org/wiki/ISO\\_639-1](http://en.wikipedia.org/wiki/ISO_639-1)

<sup>52</sup> For ISO8601 see <http://www.iso.org/iso/en/prods-services/popstds/datesandtime.html>

- The dc:date field must be ISO8601 (YYY-MM-DD).

**Q: Is it possible to filter e-Theses from a heterogeneous collection?**

A: When data providers use a standard vocabulary on the document type, like in the DRIVER guidelines, filtering on the dc:type field becomes fairly easy and therefore sufficient to extract Doctoral theses from a heterogeneous collection. (see 6.9.1.1)

### 7.2.3 QUESTIONS AND ANSWERS ON RICH E-THESES METADATA FORMATS

**Q: What is the best e-Thesis specific format to use for interoperable services?**

A: the answer to this question relies on two assumptions. One assumption is to create a decent service you will need to have a lot of records from all kinds of repositories. The second assumption is to create a e-theses specific service you will need information that is specifically used in the thesis context.

E-thesis specific formats like ETD\_MS, UKETD\_DC and XMetaDiss support a certain richness that is sufficient to describe information about the thesis context. However, these formats are not widely used, and because of their specific nature, repository developers see no need to incorporate these formats by default. We expect that a generic metadata format for the Academic Information Domain will have more success to be incorporated into various repository systems throughout Europe.

Our recommendation for developers on metadata for the Academic Information Domain, is to incorporate e-thesis specific element into this generic metadata format. (see 6.2.3.2, and 6.5.1).

**Q: Is it possible to define ETD specific elements in simple Dublin Core, like the name, level and definition of the degree, and the various dates for graduation and publication?**

A: No. simple Dublin Core is not rich enough to contain this kind of information. To increase the semantic interoperability we have recommended certain fields to contain specific e-Thesis information to reduce the ambiguity. (see 6.2.3.2, and 6.5.1).

## 7.3 RECOMMENDATIONS RELATED TO INTEROPERABILITY BETWEEN DATA AND SERVICE PROVIDERS

In the information age, the repository has become more than a dead-end for documents after publication. Documents are given a second life and are spread all over the internet. The influence of service providers, demanding specific output formats, is increasing. In the DARE project, the repositories were asked to align their output to deliver the 'same'. This alignment was done by DAREnet, the portal service, and increased the overall interoperability.

As we can see, the service provider can play an important role in solving interoperability issues. The service provider could fix these problems after harvesting, as we saw in 3.5, but to increase interoperability the repositories have to fix these problems to come in line with each other.

The demand for interoperable formats lays at the side of the service provider. Currently, the service providers fix, normalise and crosswalk the differences of every repository to

get a standard syntactic and semantic metadata structure. For 5 repositories this work is very manageable, however, when the number of repositories increase, this will become harder. When every service provider has to make normalisation actions for every data provider, a lot of work has to be handled. To reduce the overall work, data providers and repository software developers have to work on implementing standards that create interoperability on syntactic and semantic levels.

Another demand from service providers, who deliver a specific service (like the e-Thesis demonstrator), is for rich theme specific metadata formats, for example E-Theses metadata formats like ETD-MS, UKETD\_DC and XMetaDiss. We expect that only service providers of some magnitude (like Google), or cooperation structures between service and data providers on a project level can demand that data providers offer a theme specific metadata format.

To be able to offer more than basic services for e-Theses, one has to change the metadata format from simple Dublin Core to a richer and e-Theses specific one. To offer the same quality as the basic recommendation on syntactic interoperability, the e-Theses metadata format has to be unambiguously defined. Currently. It is recommended to make a further study to benchmark richer formats like ETD\_MS, UKETD\_DC and XMetaDiss on syntactic and semantic interoperability. To stimulate the broad take up of various services, data providers have to work on implementing standards that create interoperability on syntactic and semantic levels.

### 7.3.1 QUESTIONS AND ANSWERS ON INTEROPERABILITY

**Q: Is there a metadata format standardisation that can be both, interoperable and have rich features?**

A: We expect that a generic metadata format for the Academic Information Domain will have more success to be incorporated into various repository systems throughout Europe.

Such format will provide mass to build services on, and yet keeps a high level of quality and is rich enough to support e-Theses. Further study needs to be made on such formats. For a example SWAP developed by UKOLN could be a candidate. (see: 6.2.3.2)

**Q: Is it possible to have interoperability on subject classification?**

A: This question is out of our scope. However, when full text indexing is possible, service providers could analyse and categorise the documents themselves. To get to the full text in an automated way, the DIDL document could be a candidate for setting up the infrastructure for such services.

**Q: What interoperable metadata standards should repositories focus on for the short term?**

A: For metadata used externally, we recommend to use simple Dublin Core, but with the DRIVER guideline recommendations incorporated. (see 6.2.1) To focus on interoperable standards and specifications guarantees to reach a broader public, and at the same time the quality stay's high.

**Q: What format to use internally?**

A: What kind of format is out of the scope of our study. However we recommend to use a metadata structure with a fine granularity to be prepared for future developments. (see 6.1.1)



**Q: How should I dumb-down my internal metadata to DRIVER compliant simple Dublin Core?**

A: When dumbing down is needed, use known crosswalks from active communities. This guarantees that ambiguity is resolved. Repository platform related communities do exist. Look at the driver-support page for more information. (see: 6.2.1)

## 7.4 CULTURAL AND EDUCATIONAL RECOMMENDATIONS

In every country the educational processes are different. The Bologna declaration<sup>53</sup> has standardised education in Europe up until the Master's degree. After this degree, there is no clear European or international definition on the post-graduate degree<sup>54</sup>. Not only the graduation and publication process differs, but also the duration of the research process. Therefore the quality of the results in a cross-European search of doctoral theses may vary enormously. (see: 6.10.1)

**Dates & educational system:** Every event in a graduation process is marked by a date. In every country the order of the events may differ. Having a plain list of dates in the metadata, like in simple Dublin Core, is not enough.

**The language & archival preferences:** When we look at the metadata of the several repositories in the demonstrator project, we can see that in each repository a mixture of combinations of the native language and English is used in the metadata. When the native language is used only, the knowledge inside the e-thesis stays isolated for the rest of Europe.

**The roles in the Theses context:** In the context of a thesis there are different roles involved by the graduation of a student. For example a 'supervisor', 'promotor', juror, committee member etc. This information cannot be put into simple Dublin Core where we only have one field: 'Contributor'.

**Weight of the degree:** One can have the title name PhD, and the Doctoral degree, but does it count as much in every country? For example, in England it takes two years to write a thesis and graduate, in Germany it takes four years. Does the German Doctoral degree count more than an English one, or do the English work harder?

### 7.4.1 QUESTIONS AND ANSWERS ON ETD'S INVOLVING CULTURAL ASPECTS

**Q: Is there a metadata format that incorporates all this kind of aspects on culture and education ?**

A: E-thesis specific formats like ETD\_MS, UKETD\_DC and XMetaDiss support a certain richness that is sufficient to describe information about the thesis context. However, our recommendation is to use a generic metadata format for the Academic Information Domain, like SWAP (see 6.10.1). We expect this to have more success to be incorporated into various repository systems throughout Europe (see 6.2.3.2, and 6.5.1). Further research needs to be done on such metadata format.

---

<sup>53</sup> [http://en.wikipedia.org/wiki/Bologna\\_process](http://en.wikipedia.org/wiki/Bologna_process)

<sup>54</sup> <http://en.wikipedia.org/wiki/Doctorate>

Meanwhile, the DRIVER guidelines should be applied when using simple Dublin Core and the ETD recommendations in this report to prevent ambiguity.

**Q: What date should repositories put in the simple Dublin Core date field?**

A: to prevent ambiguity we recommend to export only one date field when using simple Dublin Core. The content of this date field should always contain the publication date.

**Q: What language should repository metadata be in?**

A: in order to get important knowledge out of isolation repositories should have written the metadata in English. Based on the English metadata, one could always decide to translate a thesis.

**Q: What role should be in simple Dublin Core metadata?**

A: in the context of a Doctoral thesis the contributor field should be used only for the supervisor. Again to prevent other roles to be inside this field that might create ambiguity.

**Q: Where to put information about the weight of the doctoral degree?**

A: in simple Dublin Core there is no room for this kind of information. One could use the Publisher field that contains information about the university where the grade is given. This information designates indirectly the origin of distributor of the grade. And the grade-distributor (university) defines the weight of the degree.

## 7.5 CONCLUSION

Doctoral theses contain some of the most current and valuable research produced within universities, but are underused as research resources. Where electronic theses and dissertations (ETDs) are publicly available, they are used many times more often than paper theses that are available only via inter-library loan.

In our experience we can conclude that interoperability works. The advantage of having repositories supporting a common metadata format is that it is very easy to setup services like our European e-Theses portal. However, without making appropriate agreements about the metadata semantics, a lot of issues begin to arise when harvesting metadata with a poor quality. Also issues arise when the metadata format is from a poor information density. Therefore service providers are not able to create rich services that comply to the demands of the functional requirements of a portal or the expectations of a user to look for E-Theses in a way that goes beyond a normal Google search.

This project has proofed that within this repository infrastructure, interoperability of doctoral theses on a European scale is possible. Based on the Lessons Learned we have made these practical recommendations. Adopting these recommendations will improve the interoperability between service and data provider. Basically this is done by using the means we already have, OAI-PMH and simple Dublin Core and clearing out the ambiguity. We can start by working on advocacy and implementation of the recommendations today.

However, we only have reached the first phase. Further work needs to be done to create qualitatively and quantitatively richer services, and thereby make the visibility, retrievability and (re)use of this valuable knowledge possible. This can be done, not by separating the e-Theses developments as a separate entity, but integrate ETD's as part of the developments on the Academic Information Domain for a broader and wide-spread utilisation of interoperability between data and service provider.

# ANNEX

The annexes are the output documentation from the Work Package events. These Work Package (WP) events were created to divide the project in measurable parts.

The table below shows the Work Packages that are used in the project. In the content list below the table show sections with Roman numbers. These sections contain the output deliverables of the WP's.

Work Package	What
0	<p><i>Repository information</i></p> <p>This work document contains basic harvest information about all participating repositories</p>
1	<p><i>Functional specifications</i></p> <p>This work document contains the specifications for the output of the demonstrator. Considering these, the specifications what is needed 'under water'.</p>
2	<p><i>Metadata comparisson</i></p> <p>This work document contains the similarities and differences between the uses of common and specific metadata formats from the participants.</p> <p>-Common: This document will review the use of the interoperable oai_dc format, this results in recommendation for harvester mapping crosswalks.</p> <p>-Specific: Also this document makes an inventory about the different formats the participants use specifically for electronic theses. The impact of recommending one format will lead to mapping crosswalks for repositories.</p>
3	<p><i>Harvester comparisson: quickscan</i></p> <p>This work document compares the 3 harvesters from the participants.</p> <p>-In deliverable 1: The participants only deliver a list of features and the problems they have encountered using the harvester. This deliverable will serve as input for the plugfest to discuss the discrepancies.</p> <p>-In deliverable 2: Discrepancies will be added and one will elaborate more on the comparisson.</p>
4	<p><i>Getting started towards a plugfest</i></p> <p>This work document contains the planning, preparations, input and output requirements for creating a smooth plugfest.</p>
4a	P L U G F E S T

5	<p><i>Issue Log</i></p> <p>This work document keeps track of the problems and clashes that occur during the plugfest.</p>
6	<p><i>Agreements and actions: Elaborating towards a demonstrator</i></p> <p>This work document contains agreements and actions about what needs to be done after the plugfest. The planning, preparations, input and output requirements.</p>
6a	<p><i>Working on the Demonstrator</i></p> <p>At this point the parties that have to make changes use work document 6 to give form to the demonstrator. Changes in harvester, repositories, making crosswalks, adding metadata formats etc.</p>
7	<p><i>Interoperability lessons learned and recommendations</i></p> <p>This final document finds it's input from the issue log, metadata and harvester comparisson. The output will be; recommendations, requirements and planning for setting up interoperable activities with OAI-PMH on an European scale.</p> <p>This document contain recommendations for two types of audience. One is the service provider that wants to create an interoperable service. The second is the data provider who wants to get organized to be involved in an european interoperable activities.</p>
8	<p><i>Demonstrator finnished@ETD-conference</i></p> <p>The grand finale: Presenting the demonstrator at the ETD-conference</p>

## **Annex ..... 68**

### **I. Basic repository information & Metadata analysis [WP:0+2] ..... 72**

Questionnaire results .....	72
BaseURLs .....	72
Sets .....	72
Formats.....	74
Differences and Similarities in oai_dc values .....	75
Overview .....	75
Mapping flow.....	77
Analysis for mapping at Element level .....	78

Analysis for mapping at Participant level.....	79
Normalised values of the search engine index .....	84
Conclusion .....	84
Existing ETD specific formats.....	85
ETD-MS .....	85
UKETD_DC.....	88
xMetaDiss.....	92
<b>II. Functional specifications for an e-Theses portal [WP:1]</b>	<b>96</b>
<b>III. Harvester quickscan [WP:3] .....</b>	<b>97</b>
SAHARA .....	97
Introduction .....	97
OAI-PMH system features .....	98
System features .....	98
Features for the backend user .....	99
Legal features .....	100
PKP Harvester2 .....	101
Introduction .....	101
OAI-PMH system features .....	102
System features .....	102
Features for the backend user .....	102
Legal features .....	103
<b>IV. Issues inventory and recommendations for e-Theses [WP:5+6] .....</b>	<b>104</b>
Executive Summary .....	104
Summary of recommendations .....	104
Discussion .....	107
Outcomes .....	108
Additional Outcomes: List of e-theses specific metadata to be discussed at a broader level.....	110
List of Participants .....	110

<b>V. Screenshots of the Demonstrator [WP:7+8].....</b>	<b>112</b>
Logging in for the harvester .....	119
<b>VI. Example of Embargo handing .....</b>	<b>125</b>
<b>VII. List of Terms .....</b>	<b>127</b>

# I. BASIC REPOSITORY INFORMATION & METADATA ANALYSIS [WP:0+2]

## QUESTIONNAIRE RESULTS

In this section the basic results of the questionnaire are presented.

### BASEURLS

repository	BaseURL
DIVA, Sweden	<a href="http://www.diva-portal.org/oai/OAI">http://www.diva-portal.org/oai/OAI</a>
EThOS, UK	<a href="http://dspace.lib.cranfield.ac.uk/dspace-oai/request">http://dspace.lib.cranfield.ac.uk/dspace-oai/request</a>
TU-Delft, SURF/DARE, Netherlands	<a href="http://repository.tudelft.nl/oai">http://repository.tudelft.nl/oai</a>
Humboldt University of Berlin, GERMANY, Germany	<a href="http://edoc.hu-berlin.de/OAI-2.0">http://edoc.hu-berlin.de/OAI-2.0</a>
RUC Archive, Denmark	<a href="http://dspace.ruc.dk:8080/dspace-oai/request">http://dspace.ruc.dk:8080/dspace-oai/request</a>
Copenhagen Business School Working papers, Denmark	<a href="http://ir.lib.cbs.dk/oai.php">http://ir.lib.cbs.dk/oai.php</a>

The repositories from DIVA, SURF and EThOS are OAI-compliant.

Remarks:

At first the baseURL of RUC did not work. The problem was the capital letters 'OAI' in the URL that was delivered in the questionnaire; these had to be typed in lower case.

Meanwhile, we used the Copenhagen Business School Working papers baseURL for metadata analysis.

### SETS

repository	Set name	Set spec
DIVA, Sweden	Doctoral theses	postgraduateTheses
EThOS, UK	PhD and DBA theses (School of Management)	hdl_1826_26
	PhD and EngD theses - DCMT, Shrivenham	hdl_1826_12
	PhD and EngD theses (School of Applied Sciences)	hdl_1826_23
	PhD, EngD, DM and MSc by research theses	hdl_1826_7



	PhD Theses (IERC)	hdl_1826_15
	PhD theses (School of Engineering)	hdl_1826_18
	PhD and DBA theses (School of Management)	hdl_1826_26
TU-Delft, SURF/DARE, Netherlands	DARE Harvesting	A-set
Humboldt University of Berlin, Germany	Dissertations and Professional Dissertations	pub-type:dissertation
<p>RUC Archive, Denmark</p> <p><i>XML-parse error not well formed</i></p> <p><i>Location:</i>  <a href="http://dspace.ruc.dk:8080/dspace-oai/request?verb=ListSets">http://dspace.ruc.dk:8080/dspace-oai/request?verb=ListSets</a>  row 1, column 2783</p>	Ph.D. afhandlinger / Ph.D. dissertations	hdl_1800_106
	Datalogi: Ph.D. afhandlinger / Computer science: Ph.D. Dissertations	hdl_1800_128
	Journalistik: Ph.d.afhandlinger / Journalism: Ph.D Dissertations	hdl_1800_131
	Kommunikation: Ph.d. Afhandlinger / Communication: Ph.D. Dissertations	hdl_1800_135
	FS & Ph.D. afhandlinger	hdl_1800_242
	Internationale Udviklingsstudier: Ph.d. afhandlinger / Int. Development Studies: Ph.D.Dissertations	hdl_1800_252
	Biologi: Ph.d. afhandlinger / Biology: Ph.D. Dissertations	hdl_1800_259
	Kemi: Ph.d. afhandlinger / Chemistry: Ph.D. Dissertations	hdl_1800_260
	Geografi: Ph.d. afhandlinger / Geography: Ph.D. Dissertations	hdl_1800_261
	Historie og Samfundsforhold: Ph.d. afhandlinger / History and Social Theory: Ph.D. Dissertations	hdl_1800_262
	Matematik: Ph.d. afhandlinger / Mathematics: Ph.D. Dissertations	hdl_1800_263
	Fysik: Ph.d. afhandlinger / Physics: Ph.D. Dissertations	hdl_1800_264
	Filosofi: Ph.d. afhandlinger / Philosophy: Ph.D. Dissertations	hdl_1800_266
	Psykologi: Ph.d. afhandlinger / Psychology: Ph.D. Dissertations	hdl_1800_267
	Sprog og Kultur: Ph.d. afhandlinger / Language and Culture: Ph.D. Dissertations	hdl_1800_268
	Uddannelsesforskning: Ph.d. afhandlinger / Educational Studies: Ph.d. Dissertations	hdl_1800_269
	Ph.D. afhandlinger / Ph.D. dissertations	hdl_1800_106

The TU-Delft from the Netherlands is the only participating repository which has no specific Doctoral Theses set.

## FORMATS

	<b>DIVA, Sweden</b>	<b>ETHOS, UK</b>	<b>Humboldt, DE</b>	<b>Roskilde, DK</b>	<b>TU-Delft, NL</b>
Prefix:	oai_dc	oai_dc	oai_dc	oai_dc	oai_dc
NameSpace:	<a href="http://www.openarchives.org/OAI/2.0/oai_dc/">http://www.openarchives.org/OAI/2.0/oai_dc/</a>	<a href="http://www.openarchives.org/OAI/2.0/oai_dc/">http://www.openarchives.org/OAI/2.0/oai_dc/</a>	<a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>	<a href="http://www.openarchives.org/OAI/2.0/oai_dc/">http://www.openarchives.org/OAI/2.0/oai_dc/</a>	<a href="http://www.openarchives.org/OAI/2.0/oai_dc/">http://www.openarchives.org/OAI/2.0/oai_dc/</a>
Prefix:	marc21		oai_ems		dare_didl
NameSpace:	<a href="http://www.loc.gov/MARC21/slim/">http://www.loc.gov/MARC21/slim/</a>		<a href="http://edoc.hu-berlin.de/xml/schemas/ems/">http://edoc.hu-berlin.de/xml/schemas/ems/</a>		urn:mpeg:mpeg21:2002:02-DIDL-NS
Prefix:	mods		oai_pp		qdc
NameSpace:	<a href="http://www.loc.gov/mods/v3">http://www.loc.gov/mods/v3</a>		<a href="http://www.proprint-service.de/xml/schemas/v1/">http://www.proprint-service.de/xml/schemas/v1/</a>		<a href="http://www.library.tudelft.nl/xmlns/qdc/">http://www.library.tudelft.nl/xmlns/qdc/</a>
Prefix:	oai_etdms	uketd_dc			
NameSpace:	<a href="http://www.ndltd.org/standards/metadata/etdms/1.0/">http://www.ndltd.org/standards/metadata/etdms/1.0/</a>	<a href="http://naca.central.cranfield.ac.uk/ethos-oai/2.0/">http://naca.central.cranfield.ac.uk/ethos-oai/2.0/</a>			
Prefix:	libris-print				
NameSpace:	<a href="http://www.loc.gov/MARC21/slim/">http://www.loc.gov/MARC21/slim/</a>				
Prefix:	libris-electronic				
NameSpace:	<a href="http://www.loc.gov/MARC21/slim/">http://www.loc.gov/MARC21/slim/</a>				
Prefix:	marc-liu				

NameSpace:	http://www.loc.gov/MARC21/slim/				
------------	---------------------------------	--	--	--	--

All participants use DC, but with different interpretations. [metadataPrefix=oai\_dc]

DIVA provided no additional information on their e-theses format in the questionnaire.

After scanning the metadata formats, they use ETDMS 1.0 from ND LTD. [metadataPrefix=oai\_etdms]

SURF, Humboldt, Roskilde and Copenhagen have no e-thesis specific format. [metadataPrefix=oai\_dc]

ETHOS uses ethos-oai metadata scheme [metadataPrefix=uketd\_dc]

## DIFFERENCES AND SIMILARITIES IN OAI\_DC VALUES

This section analyses the differences and similarities of the metadata values of the repository output.

### OVERVIEW

In the table below the Similarities are in Green, the differences in Orange. Yellow means redundant information. Blue means uncommon information.

	Random DiVA record	Random ETHOS record	Random Surf record	Random Humboldt record	Random Copenhag en record	Random Roskilde record
<b>dc:title</b>	title string	title string	title string	title string	title string	title string
				subtitle string		
<b>dc:identifier</b>	resolving URL + database number	resolving URL + database number	resolving URL + database number	URL to file	URL ISBN +	resolving URL + database number
	urn:nbn					some number
<b>dc:creator</b>	surname, firstname of the author	surname, first Letter of the author	surname, firstname of the author	surname, firstname of the author	surname, firstname of the author	surname, firstname of the author
<b>dc:date</b>	YYYY-MM-DD	YYYY-MM-DDThh:mm:ssZ	YYYY-MM-DD	YYYY-MM-DD	YYYY	YYYY-MM-DDThh:mm:ssZ
		YYYY-MM-DDThh:mm:ssZ				YYYY-MM-DDThh:mm:ssZ
		YYYY-MM				YYYY-MM

		(Issue date)				
<b>dc:description</b>	abstract	abstract	abstract	abstract in german	abstract (mixed language)	abstract
				abstract in english		
<b>dc:type</b>	the string "text.thesis.doctoral"	the string "Thesis or Dissertation"	the string "Doctoral thesis"	the string "Text"	<b>no PhD's</b>	the string "dissertation"
		the string "Doctoral"		the string "dissertation"		
		the string "PhD"				
<b>dc:language</b>	language in ISO 639-1	language in ISO 639-1 V. 2 ???	language in ISO 639-1	language in ISO 639-2 (three letters)	language in ISO 639-2 (three letters)	no language defined
<b>dc:contributor</b>	not present	supervisor	supervisor (1 or more)	not present in DC	not present in DC	sometimes present in DC
<b>dc:format</b>		byte size (2 elements)	byte size		not present in DC	byte size
	mime-type	mime-type (2 elements)	mime-type	mime-type		mime-type
Less important elements						
<b>dc:relation</b>	not present	not present	not present	not present	YYYY-some number	some department text
<b>dc:subject</b>	not present	not present	several subject fields present	several subject fields present (also with classification codes)	several subject fields present	not present
<b>dc:publisher</b>	one publisher element present	two publisher elements present (first real publisher, second department)	not present	one publisher element present	one publisher element present	one publisher element present
<b>dc:rights</b>	present	not present	present	not present	not present	not present
<b>dc:source</b>	some number	not present	ISBN	not present	not present	not present

Harvester input						
Base URL	<a href="http://repository.tudelft.nl/oai">http://repository.tudelft.nl/oai</a>	<a href="http://www.diva-portal.org/oai/OAI">http://www.diva-portal.org/oai/OAI</a>	<a href="http://dspace.lib.cranfield.ac.uk/dspace-oai/request">http://dspace.lib.cranfield.ac.uk/dspace-oai/request</a>	<a href="http://edoc.hu-berlin.de/OAI-2.0">http://edoc.hu-berlin.de/OAI-2.0</a>	<a href="http://ir.lib.cbs.dk/oai.php">http://ir.lib.cbs.dk/oai.php</a>	<a href="http://dspace.ruc.dk:8080/dspace-oai/request">http://dspace.ruc.dk:8080/dspace-oai/request</a>
setname	DARE Harvesting	Doctoral theses	PhD theses (School of Engineering)	Dissertations and Professional Dissertations	no sets	Ph.D. afhandlingar / Ph.D. dissertations
setspec	A-set	postgraduateTheses	hdl_1826_18	pub-type:dissertation	no sets	hdl_1800_106

The table above shows the Dublin core elements [oai\_dc] used by all participating parties with a working baseURL. For this table, an XML sample of each participant has been used. Click on the links below to view the Samples.

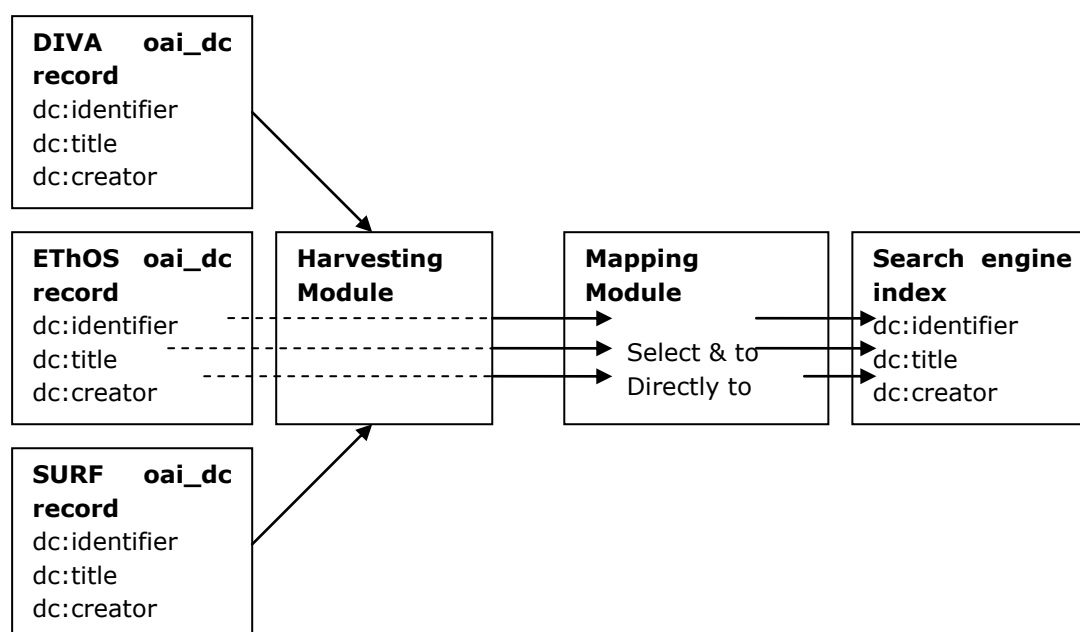
**DIVA:** [http://www.diva-portal.org/oai/OAI?verb=GetRecord&metadataPrefix=oai\\_dc&identifier=oai%3ADiVA.org%3Akh-2711](http://www.diva-portal.org/oai/OAI?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai%3ADiVA.org%3Akh-2711)

**ETHOS:** [http://dspace.lib.cranfield.ac.uk/dspace-oai/request?verb=GetRecord&metadataPrefix=oai\\_dc&identifier=oai%3Adspace.lib.cranfield.ac.uk%3A1826%2F945](http://dspace.lib.cranfield.ac.uk/dspace-oai/request?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai%3Adspace.lib.cranfield.ac.uk%3A1826%2F945)

**SURF:** [http://repository.tudelft.nl/oai?verb=GetRecord&metadataPrefix=oai\\_dc&identifier=oai%3Atudelft.nl%3A371513](http://repository.tudelft.nl/oai?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai%3Atudelft.nl%3A371513)

The colours are used to show what content has a useful common denominator that can be used for mapping to the search engine index.

## MAPPING FLOW



At this point, we assume that all participants' records are harvested in the oai\_dc format. The OAI\_DC format is mandatory in the OAI\_PMH guidelines. However, the interpretation of DC is left very open. This means to be able to create useful output for the e-theses project, the harvester has to normalise the different interpretations. This normalisation is done with a mapping. The mapping is done from a "raw" external DC format and is converted internally to a normalised and polished internal format of DC. For every participant, this mapping is different. To map, we have to know what internal DC content to use. For the time being, we have followed the DARE use of DC guidelines to describe the exact content of each DC element for the normalisation process.

However, to make this a fast implementation of the demonstrator for the plug fest, we did not look too strictly to these guidelines. The common denominator often is used as the normalisation standard. In the table above, the green colours indicate these common denominators.

## ANALYSIS FOR MAPPING AT ELEMENT LEVEL

Below, the DC elements that can directly be mapped are provided, and also those DC elements that need more effort to be put into the search engine index.

### Useful for direct mapping

All participants use the same value types for these elements.

The value types of the following elements can be used for mapping without normalising.

- **dc:title** (title string)
- **dc:creator** (string with author name; last name, first name)
- **dc:description** (abstract string)

### Partly useful for direct mapping

Only two (or more) participants share similar value types for an element. Mapping of content is not possible for all parties, but for some of the participants.

- **dc:identifier**  
*EToS* and *Surf* only use 1 identifier element (URL). The content is a link to the document.
- **dc:date**  
*Surf* and *DIVA* use the same data granularity (YYYY-MM-DD).
- **dc:language**  
*Surf* and *DIVA* use the same ISO format to describe language (ISO 639-1)
- **dc:type**  
*Surf* can map the content directly (string: 'Doctoral thesis')

## Selection crosswalk

Mapping possible after selecting the correct element in metadata.

- **dc:identifier**  
*DIVA*: provides two Identifiers. Map the 1<sup>st</sup> identifier element that contains the link.  
Use this element for mapping to dc:identifier
- **dc:date**  
*EThOS* delivers several dates. Comparison with their uketd\_dc format, shows the 3<sup>rd</sup> dc:date field is the same as the dcterms:issued field.

## Normalisation crosswalk

Mapping possible after normalising the content of a metadata element.

- **dc:type**  
*DIVA*: if dc:type contains('text.thesis.doctoral'), then change content of mapping to 'Doctoral thesis'  
*EThOS*: if dc:type contains('Thesis or Dissertation' or 'Doctoral'), then change content of mapping to 'Doctoral thesis'
- **dc:language**  
*EThOS*: Convert the language format to ISO 639-1 (keep the two characters before the \_underscore\_)

## Useful but data is missing.

- **dc:contributor**  
*EThOS* and *SURF* provide supervisor information in the contributor field.  
*DIVA* and *Humboldt* lack this field. When information is available, it is recommended to add this field.  
*Roskilde* has this field available in XML, but contains sometimes content.

## ANALYSIS FOR MAPPING AT PARTICIPANT LEVEL

Below the mapping processes are shown for each participant.

The colour code:

*Green* = everything OK. The value from the repository metadata element can directly be used to set the value for the metadata element with the same name in the search engine index.

*Yellow* = everything OK, but unnecessary elements are also put in the index (this can be solved, but requires more programming time)

*Orange* = normalisation rules are needed to put the content correctly in the index.

*Blank* = this field is not being mapped.

## DIVA

<b>From repository:</b>			<b>To search engine:</b>
<b>oai_dc</b>	<b>content sample</b>	<b>process</b>	<b>oai_dc</b>
<dc:title>		map value to:	dc:title
<dc:creator>		map value to:	dc:creator
<dc:description>		map value to:	dc:description
<dc:publisher>	KTH, Sweden		
<dc:date>	YYYY-MM-DD	map value to:	dc:date
<dc:format>	MIME-type	map value to:	dc:format
<dc:identifier>	<a href="http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-2711">http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-2711</a>	map value to: (use first item to link to)	dc:identifier
<dc:identifier>	urn:nbn:se:kth:diva-2711	map value to:	dc:identifier
<dc:type>	text.thesis.doctoral	map value to: "Doctoral thesis" if dc:type='text.thesis.doctoral'	dc:type
<dc:source>	91-7170-333-0		
<dc:language>	en	map value to:	dc:language
<dc:rights>	Copyright Jan-Olof Wesström 2000		

Note for quick implementation:

- Both dc:identifiers are indexed, but one must note that only the first dc:identifier will be used as a clickable link.
- When the dc:type element appears with the value 'text.theses.doctoral', there will be one dc:type element in the search engine index with the value 'Doctoral Thesis'

## ETHOS

<b>From repository:</b>			<b>To search engine:</b>
<b>oai_dc</b>	<b>content sample</b>		<b>oai_dc</b>
<dc:contributor>	Hammond, D. W. (supervisor)	map value to:	dc:contributor
<dc:creator>	Luxford, G	map value to:	dc:creator
<dc:date>	2005-11-23T14:32:10Z	map value to:	dc:date
<dc:date>	2005-11-23T14:32:10Z	map value to:	dc:date
<dc:date>	2005-03	map value to:	dc:date
<dc:identifier>	<a href="http://hdl.handle.net/1826/945">http://hdl.handle.net/1826/945</a>	map value to:	dc:identifier



<dc:description>		map value to:	dc:description
<dc:format>	1944 bytes	map value to:	dc:format
<dc:format>	7279452 bytes	map value to:	dc:format
<dc:format>	text/plain	map value to:	dc:format
<dc:format>	application/pdf	map value to:	dc:format
<dc:language>	en_UK	use only the two characters before the _underscore_ to map value to the index	dc:language
<dc:publisher>	Cranfield University		
<dc:title>		map value to:	dc:title
<dc:type>	Thesis or Dissertation		
<dc:type>	Doctoral	map value to: "Doctoral thesis" if record has dc:type='Doctoral' or dc:type'PhD'	dc:type
<dc:type>	PhD		
<dc:publisher>	School of Engineering		

Note for quick implementation:

- All dc:date elements are use in the index
- All dc:format elements are used in the index
- When dc:type elements with the value 'Doctoral' appear, there will appear only one dc:type element in the search engine index with the value 'Doctoral Thesis'
- dc:publisher will not be mapped
- the value of dc:language has to be stripped to ISO 639-1

## SURF

<b>From repository:</b>			<b>To search engine:</b>
<b>oai_dc</b>	<b>content sample</b>		<b>oai_dc</b>
<dc:rights>	(c) 2006 V. Kovalenko		
<dc:creator>	Kovalenko, V.	map value to:	dc:creator
<dc:date>	YYYY-MM-DD	map value to:	dc:date
<dc:identifier>	<a href="http://repository.tudelft.nl/file/415966/371513">http://repository.tudelft.nl/file/415966/371513</a>	map value to:	dc:identifier
<dc:contributor>	Ligthart, L.P., prof.dr.ir. (promotor)	map value to:	dc:contributor

<dc:contributor>	Yarovyi, A., prof.dr.sci. (promotor)	map value to:	dc:contributor
<dc:source>	ISBN:978-9076928-11-1		
<dc:language>	en	map value to:	dc:language
<dc:subject>	ground penetrating radar		
<dc:subject>	landmine detection		
<dc:subject>	clutter suppression		
<dc:subject>	feature fusion		
<dc:format>	2.8 Mbytes	map value to:	dc:format
<dc:format>	application/pdf	map value to:	dc:format
<dc:type>	Doctoral thesis	map value to:	dc:type
<dc:title>	Advanced GPR data processing algorithms for detection of anti-personnel landmines	map value to:	dc:title
<dc:description>	blabla etc..	map value to:	dc:description

**Humboldt**

<b>From repository:</b>			<b>To search engine:</b>
<b>oai_dc</b>	<b>content sample</b>		<b>oai_dc</b>
<dc:title>	Akute Enzephalitiden ...	map value to:	dc:title
<dc:title>	klinisches und ätiologisches ..	map value to:	dc:title
<dc:creator>	Schielke, Eva	map value to:	dc:creator
<dc:subject>	Medizin	map value to:	
<dc:subject>	Enzephalitis	map value to:	
<dc:subject>	Langzeitverlauf	map value to:	
<dc:subject>	YE 4500	map value to:	
<dc:description>	Akute Enzephalitiden treten überwiegend .....	map value to:	dc:description
<dc:description>	Acute encephalitis occurs mainly sporadically .....	map value to:	dc:description
<dc:publisher>	Medizinische Fakultät - .....	map value to:	
<dc:date>	YYYY-MMDD	map value to:	dc:date
<dc:type>	Text		

<dc:type>	dissertation	map content to "Doctoral thesis" if record has dc:type='dissertation'	dc:type
<dc:format>	text/html	map value to:	dc:format
<dc:identifier>	<a href="http://edoc.hu-berlin.de/habilitationen/schielke-eva-2001-11-06/PDF/Schielke.pdf">http://edoc.hu-berlin.de/habilitationen/schielke-eva-2001-11-06/PDF/Schielke.pdf</a>	map value to:	dc:identifier
<dc:language>	ger	map value to:	dc:language

Remarks:

- The language is, I believe, not ISO-639-1. Conversion tables have to be used.
- The Dissertation set apparently also contains other material than 'dissertations'.
- Map content to "Doctoral thesis" if record has dc:type='dissertation' (lower case)
- Two titles appear. Use the first title for presentation.
- The identifier is a direct link to the document; the other participants have a splash-page.

## Roskilde

<b>From repository:</b>			<b>To search engine:</b>
<b>oai_dc</b>	<b>content sample</b>		<b>oai_dc</b>
<dc:creator>	Bredsdorff, Nils	map value to:	dc:creator
<dc:date>	YYYY-MM-DDThh:mm:ssZ	map value to:	dc:date
<dc:date>	YYYY-MM-DDThh:mm:ssZ		
<dc:date>	YYYY-MM		
<dc:identifier>	0909-9174		
<dc:identifier>	<a href="http://hdl.handle.net/1800/243">http://hdl.handle.net/1800/243</a>	map value to:	dc:identifier
<dc:contributor>		map value to:	dc:contributor
<dc:relation>	FS & Ph.D. afhandlinger;13		
			dc:language
<dc:format>	2750158 bytes	map value to:	dc:format
<dc:format>	application/pdf	map value to:	dc:format
<dc:type>	Dissertation	map content to "Doctoral thesis" if record has dc:type='Dissertation' (First Upper Case)	dc:type(Doctoral thesis)

<dc:title>	Forvaltningshistorie forvaltningsvidenskab	og	map value to:	dc:title
<dc:description>	Summary. The aim of the dissertation is twofold		map value to:	dc:description
<dc:publisher>	Institut for Samfundsøkonomi og Planlægning			

Remarks:

- The Dissertation set apparently also contains other material than 'dissertations'.
- Map content to "Doctoral thesis" if record has dc:type='Dissertation' (First Upper Case)
- Contains no language field.
- Use the first date for sorting
- Use the identifier with http://

## NORMALISED VALUES OF THE SEARCH ENGINE INDEX

The output of the mapping is stored in a search engine index. These elements can be used in the front end application. It is nice to know what information one can expect to have in the index.

<b>search engine:</b>		
<b>oai_dc element</b>		<b>value</b>
dc:creator		Author string; Surname, first name
dc:date		YYY-MM-DDThh:mm:ssZ , YYYY-MM-DD and YYYY-MM
dc:identifier		resolving URL + database number and urn:nbn
dc:contributor		Supervisor string; Surname, first name (supervisor)
dc:language		Language in ISO 639-1
dc:format		Byte sizes and Mime-Types
dc:type		Doctoral theses
dc:title		Title string
dc:description		Abstract string

When using this index to the front-end, remember that only EThOS and SURF deliver values for the dc:contributor element.

## CONCLUSION

I used only one sample from each participating repository to make a preliminary mapping. Extensive analysis could be done with a large number of metadata records to find more complications. However, I assume that all other records in the repository are analogous to the sample record. This means, with this assumption, that I can draw the same conclusions when studying one sample record, instead of studying a large number.

To make a deeper analysis, we need to harvest the participants' repositories. More specific issues will then surface.

It is feasible to make a quick and dirty service to harvest e-theses with oai\_dc and present it on a website. However, to describe e-theses a more semantically sufficient format has to be chosen. For example: we know that dc:contributor resembles the name of the supervisor, but when a person does not know of this context and does not know how to interpret this information, the semantics of contributor are very ambiguous.

Therefore, a service provider can disambiguate this information by linking the correct semantics to the values through the use of a better internal format for the search engine index.

Even better, is to let the data providers deliver the semantically correct data by providing an e-theses specific format.

These specific formats already exist. Another analysis has to be made to make recommendations for this e-theses standard. In this e-theses project, we could talk about interoperable metadata formats, but we should talk about semantically-correct metadata schemata for e-theses.

## EXISTING ETD SPECIFIC FORMATS

In this part, a questionnaire has been sent-out to experts of the ETD specific metadata formats. In this regard, I would like to thank Rita Voigt, Paul Needham and Susanne Dobratz for their contributions.

### ETD-MS

(Rita Voigt)

Q	A short description of ETD-MS
A	<p>ETD-MS is an e-theses specific application profile which consists of:</p> <p>standard DC elements (13 out of 15) and recommended qualifiers, i.e. qualified DC</p> <p>and one additional optional e-theses specific element: thesis.degree, which has qualifiers name, level, discipline, and grantor</p> <p>ETD-MS was developed during 2000-2001 by the NDLTD standards committee. No particular improvements since then.</p>
Q	What are the main advantages of ETD-MS?
A	<p>Can be seen as an official recommendation from the NDLTD</p> <p>Widely used in the US</p>

	<p>Provides richer metadata than simple DC</p> <p>Easy to implement, straightforward dump-down to simple DC possible</p>
Q	What are the main disadvantages of ETD-MS?
A	<p>Outdated, obviously no further development planned</p> <p>Too simple, not enough granularity, too few recommendations on proper usage (e.g. order of first name, last name)</p> <p>Too US oriented, reflects mainly features from gaining a doctoral degree in the US</p> <p>Optional language attribute. Should be mandatory in the European context</p> <p>Semantics of the dc.date element unclear</p> <p>No support for complex objects, e.g. compound dissertations consisting of a summary part and several separate already published articles, dc.relation element not present at all in ETD-MS</p>
Q	What was the initial purpose to develop ETD-MS?
A	Identified need for a common ETD specific metadata format
Q	By whom was ETD-MS developed?
A	The NDLTD standards committee
Q	What makes ETD-MS so special?
A	First/early attempt to develop a common metadata format for describing ETDs and include additional elements specific to ETDs. Wide acceptance within the ETD community, especially in the US.

Q	What is the maintenance status of ETD-MS?
A	None.
Q	Give an indication of the number of repositories using ETD-MS.
A	<p>According to The University of Illinois OAI-PMH Data Provider Registry about 100 repositories.</p> <p><a href="http://gita.grainger.uiuc.edu/registry/">http://gita.grainger.uiuc.edu/registry/</a></p> <p><a href="http://gita.grainger.uiuc.edu/registry/ListSchemas.asp">http://gita.grainger.uiuc.edu/registry/ListSchemas.asp</a></p>
Q	Give an indication of the number of service providers actively harvesting the ETD-MS format?
A	<p>I'm not sure, and unfortunately haven't got the time to check this in-depth. At least the ETD union catalog run by OCLC</p> <p><a href="http://www.oclc.org/research/projects/etd/default.htm">http://www.oclc.org/research/projects/etd/default.htm</a></p> <p><a href="http://ndltd.oclc.org">http://ndltd.oclc.org</a> (link broken ☺)</p>
Q	Would you recommend ETD-MS to become an internet standard that is being used for interoperable exchange of ETD metadata? Why / Why not?
A	<p>Yes and no. Yes, because it is easy to implement (e.g. in addition to some more suitable profile), so it should be provided at least until we can agree upon a better one. No, because of the shortcomings listed in question 3.</p>
Q	Provide URL's for further information about ETD-MS.

A	<a href="http://www.ndltd.org/standards/metadata/current.html">http://www.ndltd.org/standards/metadata/current.html</a>
Q	Other relevant information you want to share.
A	See my presentation at the Utrecht workshop Jan. 2007 for any additional information

Example XML output: [http://www.diva-portal.org/oai/OAI?verb=GetRecord&metadataPrefix=oai\\_etdms&identifier=oai%3ADiVA.org%3Akth-2711](http://www.diva-portal.org/oai/OAI?verb=GetRecord&metadataPrefix=oai_etdms&identifier=oai%3ADiVA.org%3Akth-2711)

## UKETD\_DC

(Paul Needham)

Q	1. A short description of UKETD_DC
A	<p>UKETD_DC is an e-theses specific application profile which consists of:</p> <ul style="list-style-type: none"> <li>• standard DC elements and recommended qualifiers, i.e. qualified DC</li> <li>• and uketd_dc 'namespace' (domain specific) extensions – 'uketdterms', which are largely e-theses specific elements: <ul style="list-style-type: none"> <li>○ uketdterms:advisor</li> <li>○ uketdterms:sponsor</li> <li>○ uketdterms:grantnumber</li> <li>○ uketdterms:checksum</li> <li>○ uketdterms:institution</li> <li>○ uketdterms:department</li> <li>○ uketdterms:commercial</li> <li>○ uketdterms:embargodate</li> <li>○ uketdterms:embargoreason</li> <li>○ uketdterms:qualificationname</li> <li>○ uketdterms:qualificationlevel</li> </ul> </li> </ul> <p>UKETD_DC allows harvesting of both metadata and content.</p>
Q	2. What are the main advantages of UKETD_DC?



A	<p>UKETD_DC is compatible with other e-thesis metadata standards, and in particular, it is compatible with NDLTD's ETD-ms standard (<a href="http://www.ndltd.org/standards/metadata/current.html">http://www.ndltd.org/standards/metadata/current.html</a>).</p> <p>UKETD_DC is an extension of qualified DC offering relatively semantically clear metadata, which means it can be widely understood and it 'dumbs down' gracefully and cleanly to simple DC</p> <p>UKETD_DC can easily be implemented in DSpace and GNU Eprints, for which we have developed plug-ins. And it should also prove easy to implement in other software such as Fedora.</p>
Q	3. What are the main disadvantages of UKETD_DC?
A	<p>Being based on qualified Dublin Core, UKETD_DC is a flat metadata scheme. It doesn't support hierarchical metadata and is therefore, in truth limited in its abilities to handle complex digital objects. An example of a problem this causes: Where a thesis consists of multiple files it also has multiple elements related to format and identifier, e.g.:</p> <pre> &lt;dcterms:extent&gt;11375973 bytes&lt;/dcterms:extent&gt;  &lt;dcterms:extent&gt;731 bytes&lt;/dcterms:extent&gt;  &lt;dc:format xsi:type="dcterms:IMT"&gt;application/pdf&lt;/dc:format&gt;  &lt;dc:format xsi:type="dcterms:IMT"&gt;text/plain&lt;/dc:format&gt;  &lt;dc:identifier xsi:type="dcterms:URI"&gt;      http://dspace.lib.cranfield.ac.uk/bitstream/1826/828/2/A.pdf  &lt;/dc:identifier&gt;  &lt;uketdterms:checksum xsi:type="uketdterms:MD5"&gt;      9ac24449b4c9bb61f1d25acbd9752e82  &lt;/uketdterms:checksum&gt;  &lt;dc:identifier xsi:type="dcterms:URI"&gt;      http://dspace.lib.cranfield.ac.uk/bitstream/1826/828/3/E.txt  &lt;/dc:identifier&gt;  &lt;uketdterms:checksum xsi:type="uketdterms:MD5"&gt;      a2b95a14a3320aa1a959501b449cb4f8  &lt;/uketdterms:checksum&gt; </pre> <p>Fortunately, when GNU Eprints and DSpace expose metadata in the OAI interface, the format fields are always in the same order as the dc.identifier fields for the files, making it possible to match the format and identifier elements with each other. However, strictly, this breaks the rules for processing XML – you should not rely on the order that elements appear in. Clearly it would better if the related</p>

format and identifier fields were bundled together within an outer wrapper, e.g.

```
<filebundle>
```

```
<dc:terms:extent>11375973 bytes</dc:terms:extent>
```

```
<dc:format xsi:type="dc:terms:IMT">application/pdf</dc:format>
```

```
<dc:identifier xsi:type="dc:terms:URI">
```

```
http://dspace.lib.cranfield.ac.uk/bitstream/1826/828/2/A.pdf
```

```
</dc:identifier>
```

```
<uketdterms:checksum xsi:type="uketdterms:MD5">
```

```
9ac24449b4c9bb61f1d25acbd9752e82
```

```
</uketdterms:checksum>
```

```
</filebundle>
```

```
<filebundle>
```

```
<dc:terms:extent>731 bytes</dc:terms:extent>
```

```
<dc:format xsi:type="dc:terms:IMT"> text/plain</dc:format>
```

```
<dc:identifier xsi:type="dc:terms:URI">
```

```
http://dspace.lib.cranfield.ac.uk/bitstream/1826/828/3/E.txt
```

```
</dc:identifier>
```

```
<uketdterms:checksum xsi:type="uketdterms:MD5">
```

```
a2b95a14a3320aa1a959501b449cb4f8
```

```
</uketdterms:checksum>
```

```
</filebundle>
```

Q 4. What was the initial purpose to develop UKETD\_DC?

A We developed the UKETD\_DC for practical and pragmatic reasons relevant to the forthcoming EThOS service, offering a one-stop shop for e-theses in the UK:

- We had to match a schema against a UK Core Metadata Set for Electronic Theses and Dissertations (ETDs), which had been defined by the work of a number of JISC-funded e-theses projects.
- It had to be practical and easy to implement with DSpace and GNU Eprints, within a relatively short timescale
- It had to be semantically rich enough to express the core metadata set

	and to fit the needs of the British Library. Simple DC was too simple. METS and MPEG-21 DIDL container schemes, in use with DSpace and/or GNU Eprints were generally only exposing simple DC within them, and the architectures of the softwares didn't really fully support the schemes anyway
Q	5. By whom was UKETD_DC developed?
A	UKETD_DC was developed by the EThOS technical team comprising members from Cranfield University, the British Library, Edinburgh University and Robert Gordon University.
Q	6. What makes UKETD_DC so special?
A	The fact that it will be adopted by virtually all HE institutions across the UK. It will be the first time we have had a national standard for e-theses metadata.
Q	7. What is the maintenance status of UKETD_DC?
A	<p>At the time of writing this, the uketd_dc namespace is still being maintained at a temporary address: <a href="http://naca.central.cranfield.ac.uk/ethos-oai/2.0/">http://naca.central.cranfield.ac.uk/ethos-oai/2.0/</a>. A permanent home stills needs to be identified before the EThOS service goes live and before final 'production' versions of the DSpace and GNU Eprints plug-ins can be delivered.</p> <p>For the permanent namespace, we are recommending a URI along the lines of <a href="http://ethos.bl.uk/namespaces/20060608">http://ethos.bl.uk/namespaces/20060608</a> - it is unambiguously related to EThOS and the use of a 'datestamp' directory allows for future enhancements to the schema without breaking old applications.</p>
Q	8. Give an indication of the number of repositories using UKETD_DC.
A	Currently, only a handful of repositories are using UKETD_DC. However, when the EThOS service goes live, the number of repositories exposing UKETD_DC will increase rapidly - to between 60 and 170
Q	9. Give an indication of the number of service providers actively harvesting the UKETD_DC format?
A	As far as I am aware, one - the British Library
Q	10. Would you recommend UKETD_DC to become an internet standard that is being used for interoperable exchange of ETD metadata? Why / Why not?

A	Yes and no! UKETD_DC – as it stands - will become the <i>de facto</i> UK standard for theses, once the EThOS service goes live. In the short term, this will at least make e-theses metadata consistent within the UK. However, in the longer term, a schema must be developed which will support hierarchical metadata – more like XMetaDiss. METS and MPEG-21 DIDL wrappers, along with a rich metadata schema for ETDs, offer interesting possibilities.
Q	11. Provide URL's for further information about UKETD_DC.
A	<p><i>EThOS XML schemas</i></p> <p>The EThOS XML schema definitions are as follows:</p> <ul style="list-style-type: none"> <li>• <a href="#">uketd_dc.xsd</a></li> <li>• <a href="#">uketddc.xsd</a></li> <li>• <a href="#">uketdterms.xsd</a></li> </ul> <p><i>Local copies of DCMI XML schemas</i></p> <p>For reasons of convenience and performance, we have used local copies of the DCMI XML schemas:</p> <ul style="list-style-type: none"> <li>• <a href="#">dc.xsd</a></li> <li>• <a href="#">dcmitype.xsd</a></li> <li>• <a href="#">dcterms.xsd</a></li> </ul> <p><i>Test instance metadata</i></p> <p>Test instances of the metadata, with and without the OAI-PMH wrapper, are available</p> <ul style="list-style-type: none"> <li>• <a href="#">uketd.xml</a> [<a href="#">validate</a>]</li> <li>• <a href="#">oai-uketd.xml</a> [<a href="#">validate</a>]</li> </ul>
Q	12. Other relevant information you want to share.
A	Although the EThOS project has ended, a follow-up project, EThOSnet, is just starting. It will take EThOS from a prototype to a live service. In the course of the work to be carried out by EThOSnet, UKETD_DC may change and be developed further.

Example XML output: [http://dspace.lib.cranfield.ac.uk/dspace-oai/request?verb=GetRecord&metadataPrefix=uketd\\_dc&identifier=oai%3Adspace.lib.cranfield.ac.uk%3A1826%2F945](http://dspace.lib.cranfield.ac.uk/dspace-oai/request?verb=GetRecord&metadataPrefix=uketd_dc&identifier=oai%3Adspace.lib.cranfield.ac.uk%3A1826%2F945)

## XMETADISS

(Susanne Dobratz)

Q	A short description XMetaDiss
A	Format of the metadata set of the German National Library for online dissertations and post-doctoral theses. The format is co-ordinated with the university libraries in Germany and other countries. The potential of XMetaDiss consists in the use of hierarchical patterns. The basis for the data elements described in the XMetaDiss format is the Dublin Core Metadata Element Set.
Q	What are the main advantages of XMetaDiss?
A	<p>the use for an automatic compiling-process of metadata of online university theses and dissertations by OAI protocol</p> <p>the targeted compatibility with the NDLTD-set ETDMS</p> <p>the use of hierarchical patterns and the avoidance of allocation errors (up to now the allocation occurred only through the sequence of the elements)</p> <p>and the simple transformation potentiality by means of XSLT into other metadata formats, as ETDMS and DC simple.</p>
Q	What are the main disadvantages of XMetaDiss?
A	
Q	What was the initial purpose to develop XMetaDiss?
A	The metadata set is supposed to relieve the set MetaDiss which is up to now embedded into HTML4.
Q	By whom was XMetaDiss developed?
A	Deutsche Nationalbibliothek
Q	What makes XMetaDiss so special?
A	

Q	What is the maintenance status of XMetaDiss?
A	
Q	Give an indication of the number of repositories using XMetaDiss.
A	<p><a href="http://archiv.tu-chemnitz.de/cgi-bin/interfaces/oai/oai2.pl?verb=ListMetadataFormats">http://archiv.tu-chemnitz.de/cgi-bin/interfaces/oai/oai2.pl?verb=ListMetadataFormats</a></p> <p><a href="http://doku.b.tu-harburg.de/oai/oai2.php">http://doku.b.tu-harburg.de/oai/oai2.php</a></p> <p><a href="http://edoc.hu-berlin.de/OAI-2.0">http://edoc.hu-berlin.de/OAI-2.0</a></p> <p><a href="http://edoc.ub.uni-muenchen.de/perl/oai2">http://edoc.ub.uni-muenchen.de/perl/oai2</a></p> <p><a href="http://miami.uni-muenster.de/servlets/OAIDataProvider">http://miami.uni-muenster.de/servlets/OAIDataProvider</a></p> <p><a href="http://psydok.sulb.uni-saarland.de/phpoi/oai2.php">http://psydok.sulb.uni-saarland.de/phpoi/oai2.php</a></p> <p><a href="http://www.freidok.uni-freiburg.de/oai2/oai2.php">http://www.freidok.uni-freiburg.de/oai2/oai2.php</a></p> <p>8 repositories</p>
Q	Give an indication of the number of service providers actively harvesting the XMetaDiss format?
A	Deutsche Nationalbibliothek
Q	Would you recommend XMetaDiss to become an internet standard that is being used for interoperable exchange of ETD metadata? Why / Why not?
A	
Q	Provide URL's for further information about XMetaDiss.
A	<a href="http://www.d-nb.de/eng/standards/xmetadiss/xmetadiss.htm">http://www.d-nb.de/eng/standards/xmetadiss/xmetadiss.htm</a>
Q	Other relevant information you want to share.
A	

Example XML output: [http://edoc.hu-berlin.de/OAI-2.0?verb=GetRecord&metadataPrefix=oai\\_xmetadiss&identifier=oai:HUBerlin.de:10068](http://edoc.hu-berlin.de/OAI-2.0?verb=GetRecord&metadataPrefix=oai_xmetadiss&identifier=oai:HUBerlin.de:10068)



## II. FUNCTIONAL SPECIFICATIONS FOR AN E-THESES PORTAL [WP:1]

In several conference calls a list of functional specifications has been setup for an ETD portal.

These specifications are made to create the demonstrator in the context of the Knowledge Exchange ETD strand. The conference call was between Hubert Krekels (Wageningen University and Research, NL) , Susanne Dobratz (Humboldt University, DE), Gerard van Westrienen and Maurice Vanderfeesten (SURFfoundation, NL)

The functional specifications are divided into two parts. Need to have (\*) and Nice to have.

- Presentation
  - English interface (\*)
  - Multi-language interface
  - Result
    - Metadata in original language(\*)
    - Translated metadata (automatic/change language source)
- Search
  - All fields (\*)
  - Title field (\*)
  - Author field (\*)
  - Restriction/filter
    - Country field
    - Institution field
- Browse
  - Country field
    - Institution field



### III. HARVESTER QUICKSCAN [WP:3]

In this workpackage we are going to provide a feature list of two harvesters that are being used by the partners. SAHARA (NL) and PKP OAI Harvester (UK)

## SAHARA

### INTRODUCTION

The system called *Sahara* consists of two subsystems, namely an OAI-harvester and a webcontrol panel. Throughout this document we will use the term *Sahara* by which we mean the system as a whole. Figure 1 shows how *Sahara* interacts with other systems.

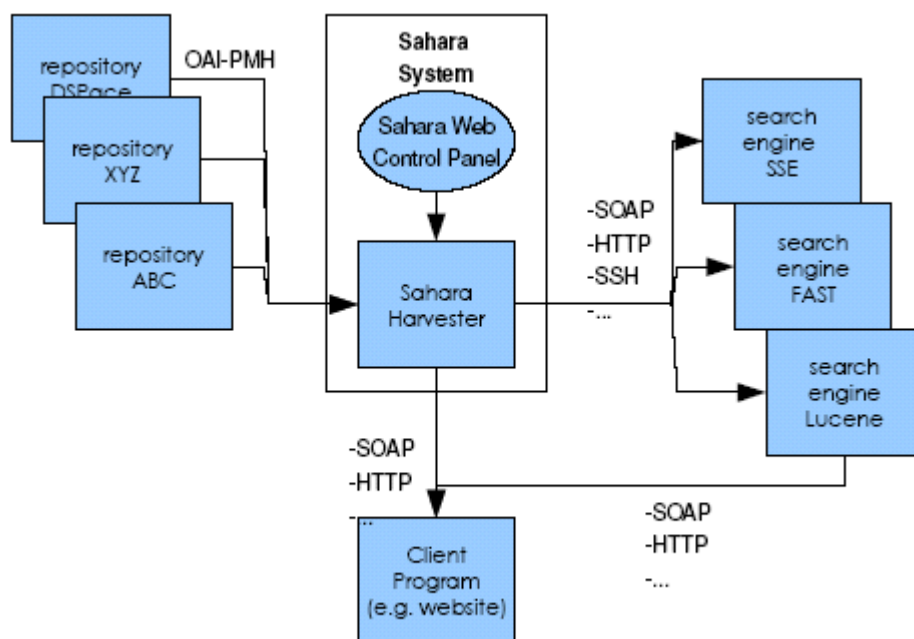


Figure 1: Sahara System

The harvester part of Sahara is the part that interacts with the repositories to retrieve the information stored in that repository. The format of contents that the harvester retrieves must comply to the OAI-PMH protocol in order for the harvester to harvest the contents successfully. Currently no other protocol to harvest repositories is supported by the Sahara.

The harvester is also capable of converting the retrieved data from any given structure into another. After a possible conversion, the harvester can upload the data to a designated target. At the moment the Sahara system supports the SurfNet Search Engine (SSE) or Teddy/Lucene or a local file system as targets.

The web-control panel controls the behaviour of the harvester by means of a website. This web-control panel provides the means to add, modify and remove repositories, to group them under a single name and to specify the way the harvester should treat the harvested contents, as mentioned above. The web-control panel also provides means to grant access to certain repositories to users by placing them in domains and allowing the application administrator to link users to these domains.

Besides providing an interface for humans, the web-control panel also provides an interface for other systems. This interface is the link between the web-control panel and

the actual harvester. Using this interface, called SaharaGet (see Figure 2), the harvester can retrieve the information it needs to harvest and process retrieved contents.

Sahara has been in production as of 2002 and, since then, additional features have been added to make sure that repositories get harvested, no matter what happens. It can handle repositories being off-line often, variations in meta-data formats, expired resumption tokens, etc. Experience has shown that it takes more than OAI-PMH to get repositories harvested without constantly having to look after the process. This experience has been incorporated into Sahara.

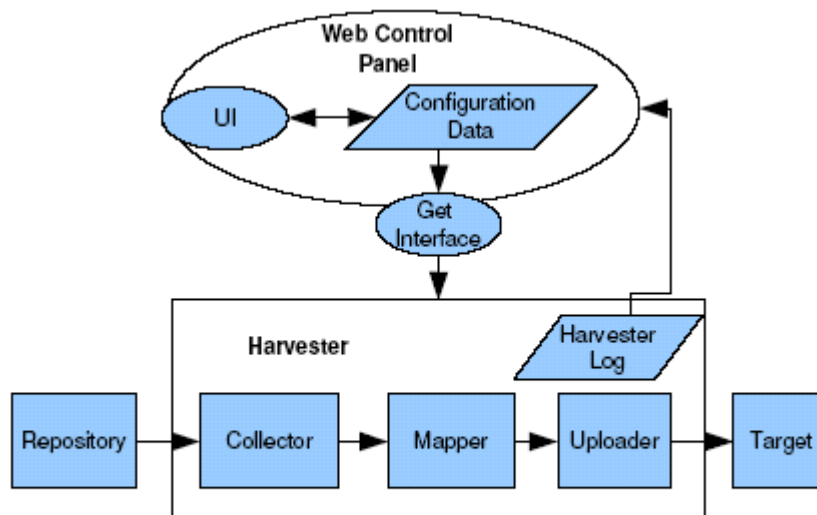


Figure 2: Sahara Architecture

## OAI-PMH SYSTEM FEATURES

- Full OAI v2.0 support; Sahara implements all features of the OAI-PMH protocol v.2.0.
- Selective harvesting; from any date, set specific, metadata format specific.
- Incremental harvesting; continues harvesting from a previous harvest date.
- Deleted records support; support for the three options on the deleted records policy to keep repository content always up-to-date.
- Batch processing; uses resumption tokens to continue with the next batch.

## SYSTEM FEATURES

- Scalable architecture
- Separated domains
- Privileges per domain
- Repository grouping
- Clear harvesting; removes previous data, new harvest
- Automatic refreshing; with the option 'no deleted records' a new harvest takes place, the harvester compares old and new data and updates/replaces old data.

- Looping; non-obstructive harvesting limit's the load on repositories by not starting a new harvest immediately after each batch. Loop means the harvester continues with the next batch after the previous batch of the whole set of repositories
- Metadata independent; harvesting and processing multiple metadata formats.
- Metadata processor (mapping, crosswalking, normalizing)
- Repository closing hours scheduler; a schedule that tells the harvester when not to harvest a repository.
- Save points; keeps old metadata in case of mayor errors.
- Web services; SAHARA-get outputs different information.
- Status reporting; mapping tester, throughput analyzer, global and individual log files.
- Output plug-ins; Output can be sent to multiple backend targets (search engine index, data base, file system).
- Flexible record pre-processing
- Secure Web-Control panel
- Secure upload
- XML Configuration files
- The SAHARA package is built in Python.

## FEATURES FOR THE BACKEND USER

Two different roles: Harvester Administrator and Domain Maintainers

The administrator of Sahara can:

1. Create accounts and delegate administration to these accounts.
2. Configure new domains and assign these to accounts
3. Add mappings
4. Add back-ends

Other user accounts on Sahara are tied to specific domains. These user accounts can:

1. Create repository groups
2. Create repositories and assign these to groups.
3. Assign Mappers to repositories
4. Assign Targets to repositories
5. Start/stop harvesting per repository
6. Schedule automatic re-harvesting
7. Validate responses
8. Validate Mapping in combination with a specific repository
9. Configure metadata prefixes, sets and collections

## LEGAL FEATURES

The SAHARA package is constructed under an Open Source licence.

## PKP HARVESTER2

### INTRODUCTION

PKP Harvester2 has been developed by the Public Knowledge Project (PKP), a research initiative involving the University of British Columbia and Simon Fraser University, funded by the federal authority in Canada.

Harvester2 is the second major version of the PKP harvesting software, which is free, open source software. Following the Model-View-Controller (MVC) pattern (see Figure 3, below), it has been designed to be flexible, robust and easy to maintain.

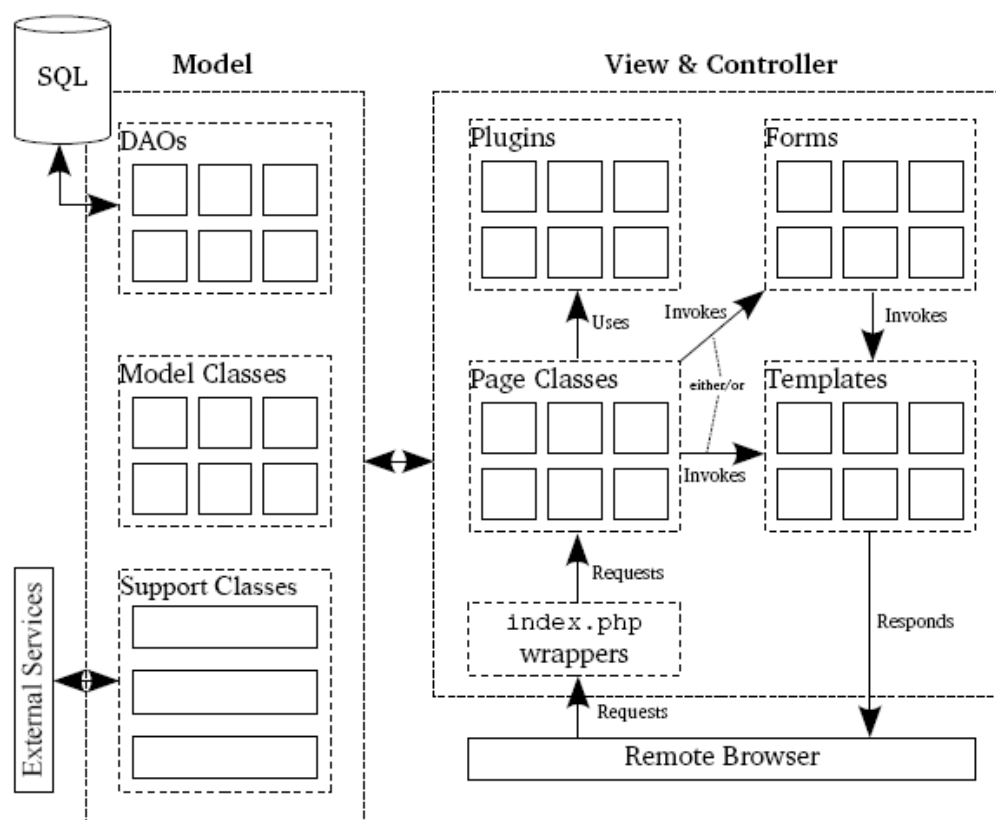


Figure 3. PKP Harvester2 system<sup>55</sup>

The system is comprised of the harvester, a web-based user search and browse interface, and a web-based administrative (backend user) interface.

The harvester requires data providers (repositories) to comply with the OAI-PMH specification in order to harvest their metadata successfully. Out-of-the-box, the harvester supports oai\_dc, MARC and MODS formats, but, through a system of plug-ins, it is relatively easy to add additional metadata schemas as required. By using the plug-in system it is also possible to add additional harvesting protocols, preprocessors, and postprocessors as required.

<sup>55</sup> Figure reproduced from PKP Harvester2 version 2.0 Technical Reference, p8, <http://pkp.sfu.ca/harvester2/TechnicalReference.pdf>, accessed 2007-04-26.

Harvester2 offers its own user search and browse interface, though there is no reason why the underlying database could not be used as a target by other systems and services, either directly or via the postprocessor plug-in system. The simple search facility searches across all harvested records, while the advanced search facility offers users much control over which records are returned from where.

The web-based backend interface allows administrators to control and configure the behaviour of the harvester.

## OAI-PMH SYSTEM FEATURES

- PKP Harvester supports harvesting using versions 1.1 and 2.0 of the OAI-PMH protocol.
- Selective harvesting; from any date, set specific, metadata format specific.
- Incremental harvesting; through cronjobs continues harvesting from a previous harvest date.
- Deleted records processing is not properly supported, so it is necessary periodically to delete and re-harvest a repository completely, but this will be fixed in a future release.
- Batch processing; uses resumption tokens to continue with the next batch.

## SYSTEM FEATURES

- MySQL or PostgreSQL
- PHP
- Apache or Microsoft IIS
- Supports localisation
- Has full support for UTF-8 when using MySQL  $\geq 4.1.1$  (or PostgreSQL  $\geq 7.1$ ) and PHP  $\geq 4.3.0$  with mbstring enabled
- XML configuration files
- Supports preprocessor and postprocessor plug-ins
- Extensibility through the plug-ins system. Future releases are likely to include:
  - SRW/SRU Service
  - SRW/SRU Clients
  - OAI Provider
- Clear harvesting; removes previous data, new harvest
- The harvester compares old and new data and updates/replaces old data.
- Metadata independent; harvesting and processing multiple metadata formats.
- Secure administrative interface
- User interface look and feel easy to customise through templates and css

## FEATURES FOR THE BACKEND USER

The software has a web-based administration interface. Administrators can:

- Setup and amend site settings
- Add, manage and delete repositories
- Select and extend language settings
- Configure schema crosswalks
- Manage plug-ins
- Perform general administration functions

## LEGAL FEATURES

Harvester2 is licensed under the GNU General Public License v2.

## IV. ISSUES INVENTORY AND RECOMMENDATIONS FOR E-THESES [WP:5+6]

This contains the Knowledge Exchange report on the e-Theses strand of the Institutional Repositories Workshop held in Utrecht January 2007.

### Authors:

Susanne Dobratz, Humboldt-University Berlin, [dobratz@cms.hu-berlin.de](mailto:dobratz@cms.hu-berlin.de)

Hubert Krekels, Wageningen University, [Hubert.Krekels@wur.nl](mailto:Hubert.Krekels@wur.nl)

Maurice Vanderfeesten, SURF Utrecht, [Vanderfeesten@surf.nl](mailto:Vanderfeesten@surf.nl)

Gerard van Westrienen, SURF Utrecht, [vanWestrienen@surf.nl](mailto:vanWestrienen@surf.nl)

## EXECUTIVE SUMMARY

The workshop showed that e-theses are an important part of a university's research output and yet not sufficiently integrated into a European repositories infrastructure and searchable for the specific e-theses information. To inform the work of the e-theses strand, a small demonstrator could show how an integrated search using the OAI-PMH could be used to offer a retrieval portal on a European scale and where the difficulties in terms of metadata are.

The participants of the discussion agreed upon the fact that e-theses should be seen as an integral part of broader institutional repositories. As such they should not follow an own metadata set, but be integrated into a metadata profile for the whole repository. It was suggested that a European e-prints Application profile should be developed. Such a profile should work out the specific metadata that only apply to doctoral e-theses.

During the workshop, work was started to identify key issues in handling doctoral e-theses and to prioritise those issues. This table can be used for planning further European and country specific activities to reach the goal: make e-theses in Europe easier to retrieve and therefore more visible via institutional repositories.

The results of the demonstrator and discussions will be further elaborated and presented as the "European e-Theses Demonstrator" at the ETD 2007 conference in Uppsala.

## SUMMARY OF RECOMMENDATIONS

The discussions at the workshop brought the following results:

1. E-theses have to be seen as part of an overall institutional repositories infrastructure and content. They should not be handled differently from other scientific and scholarly e-papers.



2. The demonstrator showed that it is generally possible to harvest European e-theses using the OAI-PMH. But it was shown in the study and discussions that Simple Dublin core is not enough. The participants agreed that a richer metadata set is necessary to offer a retrieval portal with good quality.
3. Country specific best practice examples like the *XmetaDiss* approach in Germany and the *uketd metadata set* in the UK have been presented, and compared to older e-theses specific metadata sets like the *NDLTD etd metadata set*, the latter was seen to be outdated.
4. It was suggested that a European e-theses application profile for metadata at a European level should be developed, meaning within a European working group. It was suggested that the GUIDE group could be a good umbrella for such further activities.
5. Investigations have to be made to handle e-theses in terms of metadata as part of e-prints. Therefore e-theses specific information has to be encoded into metadata. A first list of e-theses specific elements has been produced during the workshop.
6. National authorities should fund national development of richer metadata schemes. This allows the creation of a demonstrator that shows what is possible in the short term. It is the easiest to build 4 or 5 filters at the demonstration service level, but it is not scalable. The main goal is to get to a higher level based on national richer metadata, and is a good way to take the first steps in achieving a European e-prints application schema.
7. The key issues in handling doctoral e-theses have been identified by the workshop strand participants and they have been prioritised as follows:

I. Richer metadata (11 points)

- Use qualified dc
- Compound docs
- Complex objects
- Technical and preservation metadata

II. Wider IR perspective (8 points)

- How to integrate richer metadata
- Is document type a good choice to base services on
- How to get them
- Compatibility

III. ETD specific issues (7 points)

- Degree, level, definition
- Various dates
- Define minimal requirements

IV. Cultural aspects (5 points)

- Language
- Interpretation and definitions
- Local versus national services

V. Audience (2 points)

- Keep it simple
- Focus on added value metadata
- Think about target groups

VI. Subject classification ( 0 points)

## DISCUSSION

The following issues in the context of e-theses have been mentioned during the discussion:

<b>Pool of issues</b>	<b>Richer metadata</b>	<b>Theses specific</b>	<b>cultural</b>	<b>Audience &amp; services</b>	<b>Subject classification</b>	<b>Repository perspective</b>
complex or compound objects	x					
subject classification					x	
minimum metadata requirements	x					x
national differences			x			
avoid simple dc	x					
thesis specific metadata		x				
access rights						x
target groups		x				x
cultural aspects			x			
Crosswalks on an abstract level						
Graduation Degree + level		x				
Keep it simple	x					x
Identify most important fields for added value				x		
Technical information in metadata	x					
Abstract mandatory	x					
Richer metadata	x					
Part of a wider picture						x
Document type is not a good choice for managing reps						x

<b>Pool of issues</b>	<b>Richer metadata</b>	<b>Theses specific</b>	<b>cultural</b>	<b>Audience &amp; services</b>	<b>Subject classification</b>	<b>Repository perspective</b>
Roles of sponsors etc		x				
ETD specific data integrated in repositories						x
Interoperability vs. manageability						x
How to get content						x
Advisor info		x				
Rights author and user	x					x

## OUTCOMES

As a result of the workshop, several questions could be answered by the group:

### A) What is the dimension of the problem?

- Metadata at unqualified level are not rich enough to really add value in a service demonstrator
- Unqualified DC does not accommodate some essential ETD specific metadata
- E-theses have to be seen as e-publications as part of the wider e-publications picture
- The problem is not very big, it is solvable in our opinion

### B) What can be achieved in term of e-theses during the next 12-24 month?

Within the next six months:

- We want to enhance the demonstrator
- Use existing richer national metadata which are e-thesis specific
- Test, if possible, the e-prints application profile as an input format and output format for the demonstrator
- Figure out the thesis specific elements on a semantic level

Within the next two years

- Use e-prints application profile as a kick off towards developing a European Application profile which incorporates e-theses specific information

- Demonstrator for European E-prints (cooperate within KE, maybe EU- 7th framework programme)

### **C) Which approaches are needed to achieve this goal?**

- Identify e-thesis specific elements on a national level or in broader int. coop.
- Identify e-thesis specific encoding on a national level
- Identify cultural issues, language, graduation level
- Enhance the demonstrator
- Encourage countries to adapt or adopt existing profiles (for a longer term adopt European application profile)
- **Cooperation with other KE workshop strands .....**

### **D) What role could the KE partner organisations play?**

KE partner organisation could initiate, coordinate and finally fund national initiatives to reach national agreements on national e-theses specific metadata sets. They could revise current programs and projects from the perspectives explained in the e-thesis group. KE partners should take responsibility to organise a network on European E-thesis (and investigate if the GUIDE working group can be function as umbrella organisation).

### **E) How much can be done within national programmes?**

Within national programs and activities, awareness could be raised and promoted, directed towards the key issues (see above), national players such as universities, libraries and library organisations should be persuaded that it is worthwhile investigating the problem. The national partners have to give advice to local projects and initiatives, discuss the appropriate issues with them, and finally fund the necessary developments. This is a matter for the KE-partners at a local level.

### **F) Where is an international approach mandatory?**

- Developing the European application profile
- Make recommendations (KE) to data-providers targeted at balancing between data- and service providers. (in this project we are data and service provider)

**G) The over-arching findings and challenges of the discussions have been the following ones:**

- Changing the balance between data and service providers
- Moving from toddlers' speech to adult conversation
- Richer metadata
- Wider IR perspective
- ETD specific issues
- Cultural aspects
- Subject classification

- Rights

#### H) What will be the major breakthroughs if the advice leads to partner actions?

- European Eprints Application profile incorporating ETD spec. Metadata Elements
- European Eprints portal prototype

### ADDITIONAL OUTCOMES: LIST OF E-THESES SPECIFIC METADATA TO BE DISCUSSED AT A BROADER LEVEL

Element	Recom-mended	Man-datory	optional	notes
Title		X		
Creator		X		
Description	X	-		
Date (published)		X		To decide on what kind of date
Type		X		
Language		X		
Contributor	X			
Identifier		X		There has to be found a solution for compound objects
Relation			X	
Subject			X	
Publisher			X	
Rights			X	
Source			X	

### LIST OF PARTICIPANTS

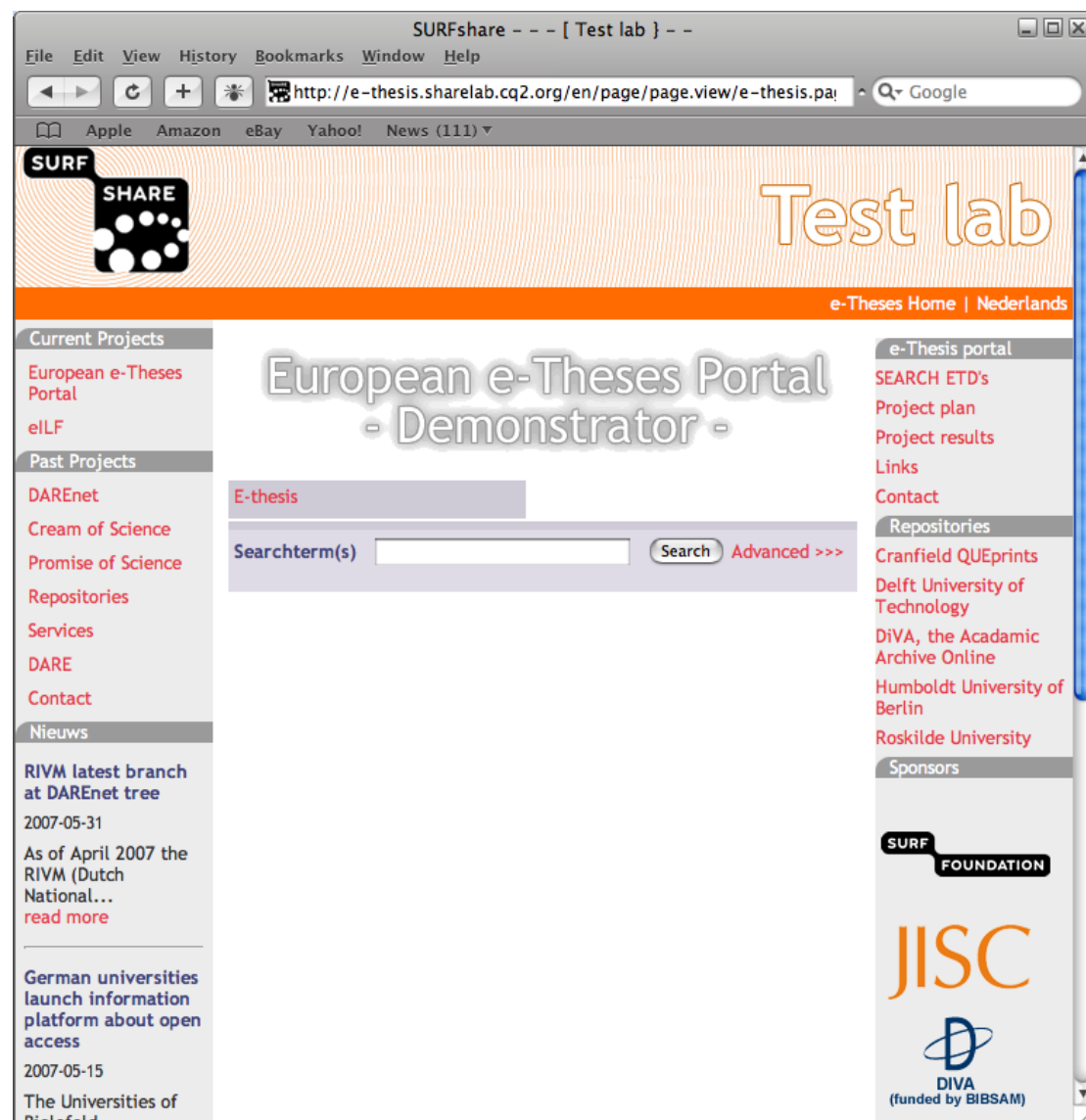
<b><u>MODERATORS:</u></b>		
Gerard van Westrienen	SURF	<a href="mailto:vanwestrienen@surf.nl">vanwestrienen@surf.nl</a>
Hubert Krekels	Wageningen University and Research Centre	<a href="mailto:hubert.krekels@wur.nl">hubert.krekels@wur.nl</a>
Susanne Dobratz	Humboldt University Berlin	<a href="mailto:dobratz@cms.hu-berlin.de">dobratz@cms.hu-berlin.de</a>

<b><u>PARTICIPANTS:</u></b>		
<b>GERMANY:</b>		
Maren Brodersen	DNB (German National Library)	<a href="mailto:m.brodersen@dnb.de">m.brodersen@dnb.de</a>
Kim Braun	University Oldenburg Library	<a href="mailto:kim.braun@uni-oldenburg.de">kim.braun@uni-oldenburg.de</a>
<b>UK:</b>		
Paul Needham	Cranfield University / ETHOS project	<a href="mailto:paul.needham11@btinternet.com">paul.needham11@btinternet.com</a>
<b>The Netherlands:</b>		
Maurice Vanderfeesten	SURF	<a href="mailto:vanderfeesten@surf.nl">vanderfeesten@surf.nl</a>
Arent Bosman	Technical University Delft	<a href="mailto:a.j.bosman@library.tudelft.nl">a.j.bosman@library.tudelft.nl</a>
<b>DENMARK:</b>		
Claus Vesterager Pedersen	Roskilde University Library	<a href="mailto:cvp@ruc.dk">cvp@ruc.dk</a>
Henrik Juul-Nyholm	University of Copenhagen	<a href="mailto:hjn@adm.ku.dk">hjn@adm.ku.dk</a>
<b>SWEDEN:</b>		
Ronnie Kolehmainen	Uppsala University Library, Electronic Publishing Centre	<a href="mailto:ronnie.kolehmainen@ub.uu.se">ronnie.kolehmainen@ub.uu.se</a>

## V. SCREENSHOTS OF THE DEMONSTRATOR [WP:7+8]

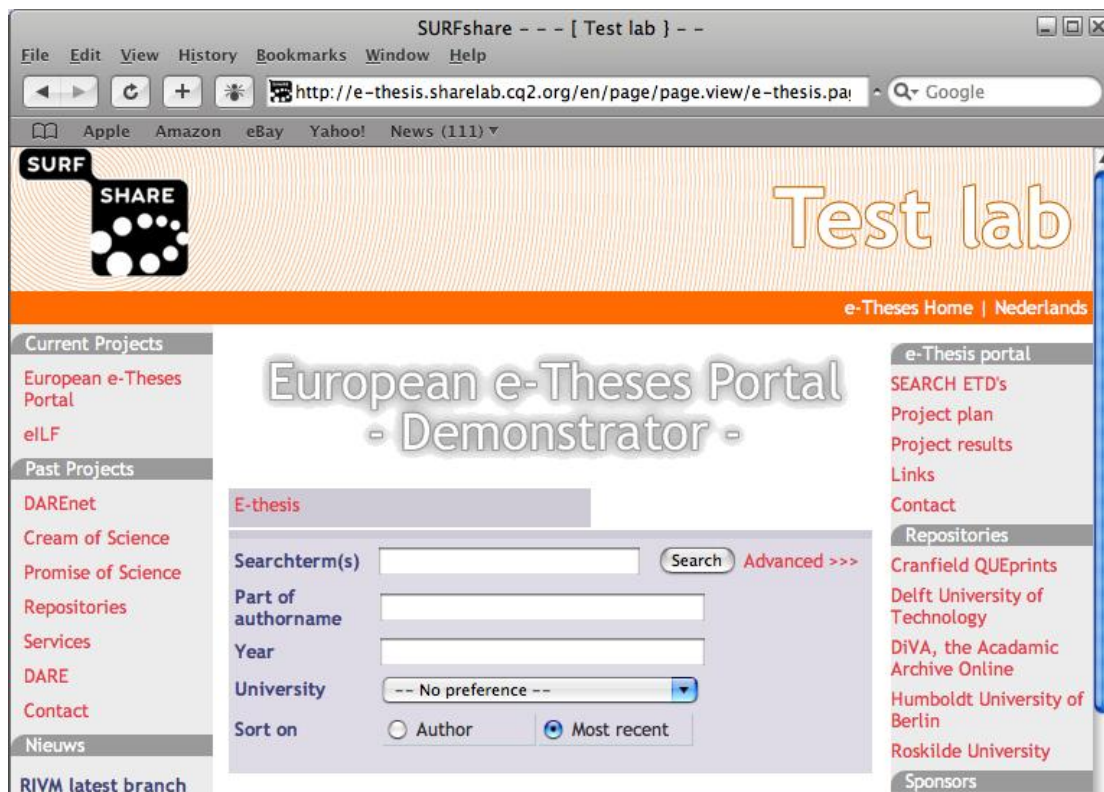
The demonstrator has been 'born' in January 2007 and will cease to exist on June 2007. The URL where the demonstrator can/could be found is at <http://e-thesis.sharelab.cq2.org>

The following pictures will show some screenshots and with some comment in the caption.

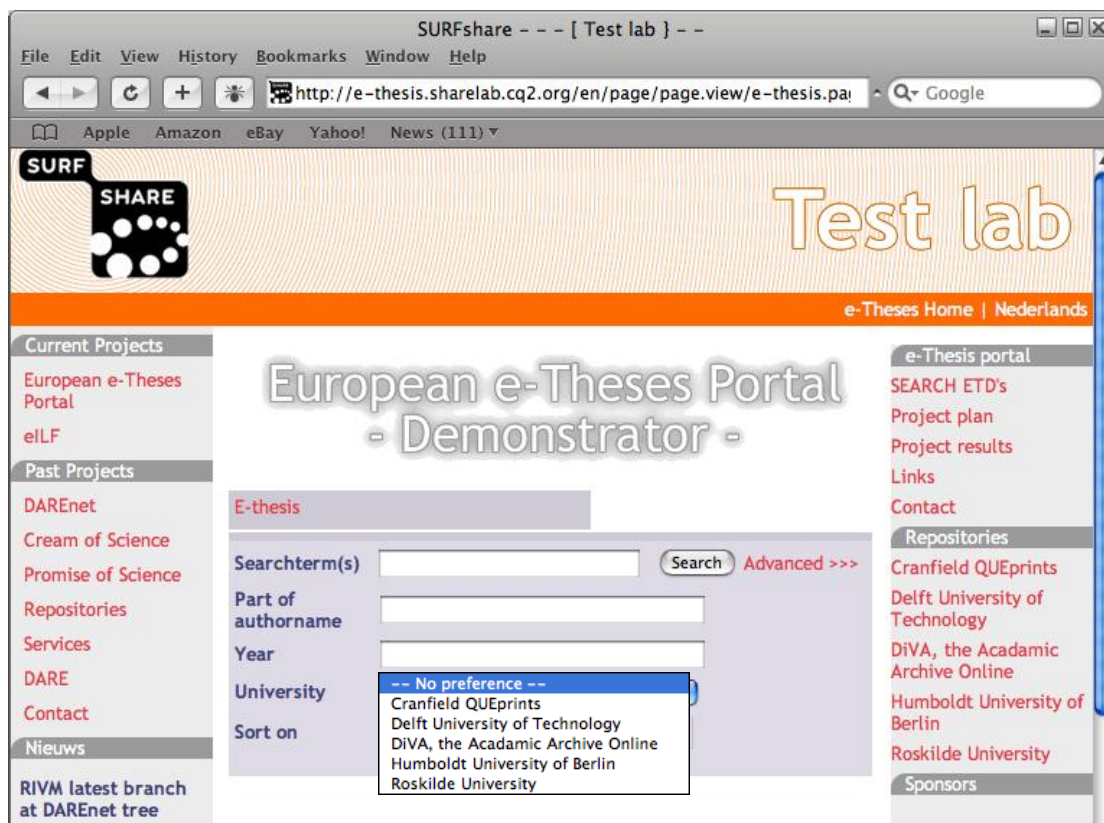


**FIGURE 18: THIS IS THE FIRST SCREEN OF THE DEMONSTRATOR. IT HAS A SIMPLE SEARCH FIELD, A SEARCH BUTTON AND AN 'ADVANCED' OPTION.**





**FIGURE 19: THE ADVANCED OPTION CREATES THE ABILITY TO SEARCH SPECIFIC FOR AN AUTHOR OR YEAR OF (PUBLICATION/GRADUATION?). THE SEARCH CAN BE RESTRICTED TO THE UNIVERSITY AND SORTING THE MOST RECENT DOCUMENT OR THE AUTHOR NAMES IN ALPHABETICAL ORDER.**



**FIGURE 20: THIS FIGURE SHOWS THE REPOSITORIES THAT ARE HARVESTED**



The screenshot shows a web browser window titled "SURFshare - - - [ Test lab ] - -". The address bar displays the URL "http://e-thesis.sharelab.cq2.org/en/page/e-thesissearch.results/sl". The page features a navigation menu on the left with links to "Current Projects", "Past Projects", and "Nieuws". The main content area shows search results for the term "mechanic".

**Test lab**  
e-Theses Home | Nederlands

**Current Projects**  
European e-Theses Portal  
eILF

**Past Projects**  
DAREnet  
Cream of Science  
Promise of Science  
Repositories  
Services  
DARE  
Contact



**Nieuws**  
RIVM latest branch at DAREnet tree  
2007-05-31  
As of April 2007 the RIVM (Dutch National...  
[read more](#)

German universities launch information platform about open access  
2007-05-15  
The Universities of Bielefeld, Goettingen,...  
[read more](#)

[news overview](#)

**E-thesis**  
Searchterm(s)   [Advanced >>>](#)

**1 to 7 of 7: (sorted by: most recent)**

- Non-invasively assessed skeletal bone status and its relationship to the biomechanical properties and condition of cancellous bone**  
2006-04-19 Cook, R. B.;  
Cancellous bone constitutes much of the volume of bone which makes up axial skeletal sites such as the vertebrae of the spine and the femoral neck. However the increased vascularity of cancellous bone compared with cortical bone means that it is more prone to drug, endocrine and metabolic related effects and therefore these skeletal...  
Found in:  [bibliographic data](#)
- Flying GLARE**  
2004-03-23 Beumler, T.;  
At the end of the second millennium did the aircraft industry decide for the first time to apply the fiber metal laminate GLARE in a large quantity on a civil transport aircraft. It was focused on an application of the material on the pressurised fuselage, the decision driven by the demand for weight saving at an affordable cost level....  
Found in:  [bibliographic data](#)
- Quantitative rastertunnelmikroskopische Untersuchungen akustischer Oberflächenwellenfelder auf der Nanometerskala**  
2002-06-19 Voigt, Peter;  
Diese Arbeit befaßt sich mit der SAW-STM-Methode, einer Abwandlung der Rastertunnelmikroskopie (engl. scanning tunneling microscopy) zur hochauflösenden Abbildung akustischer Oberflächenwellen (engl. surface acoustic wave). Das Meßprinzip des SAW-STM beruht auf der Modulation des Tunnelabstandes und der hieraus resultierenden...  
Found in:  [bibliographic data](#)
- Equilibrium and Non-Equilibrium Thermodynamics of Natural Gas Processing**  
2002 Solbraa, Even;  
The objective of this work has been to study equilibrium and non equilibrium situations during high pressure gas processing operations with emphasis on utilization of the high reservoir pressure. The well stream pressures of some of the condensate and gas fields in the North Sea are well above 200 bar. Currently the gas is expanded to a...  
Found in:  [bibliographic data](#)
- A novel Planar Magnetic Bearing and Motor Configuration applied in a Positioning Stage**  
2000-10-03 Molenaar, A.;  
This thesis presents the design and implementation of a fully contactless high precision

**FIGURE 21: SHOWS SEVERAL RESULTS WHEN SEARCHING FOR THE WORD 'MECHANIC'. YOU CAN SEE THIS RESULT COMES FROM 4 DIFFERENT REPOSITORIES. CRANFIELD, TUDELFT, HUMBOLDT AND DIVA.**

Non-invasively assessed skeletal bone status ...es and condition of cancellous bone - DAREnet

File Edit View History Bookmarks Window Help

http://e-thesis.sharelab.cq2.org/en/page/repository.item/show

Apple Amazon eBay Yahoo! News (111)

**SURF SHARE**

**Test lab**

e-Theses Home | Nederlands

**Current Projects**

European e-Theses Portal

eLF

**Past Projects**

DAREnet

Cream of Science

Promise of Science

Repositories

Services

DARE

Contact

**Nieuws**

RIVM latest branch at DAREnet tree

2007-05-31

As of April 2007 the RIVM (Dutch

**Non-invasively assessed skeletal bone status and its relationship to the biomechanical properties and condition of cancellous bone; 2006-04-19T09:13:20Z**

**Cook, R. B.**

<b>Title</b>	Non-invasively assessed skeletal bone status and its relationship to the biomechanical properties and condition of cancellous bone
<b>Author(s)</b>	Cook, R. B.
<b>Contributor(s)</b>	Zioupou, P. (supervisor)
<b>Date</b>	2006-04-19T09:13:20Z
<b>Keyword(s)</b>	Bone Cancellous bone, Fracture mechanics, Osteoporosis,
<b>Summary</b>	Cancellous bone constitutes much of the volume of bone which makes up axial skeletal sites such as the vertebrae of the spine and the femoral neck. However the increased vascularity of cancellous bone compared with cortical bone means that it is more prone to drug, endocrine and metabolic related effects and therefore these skeletal sites are more prone to the bone condition
	mechanic variables points to a clear causal relationship between the bone fracture parameters and bone condition as underlying factors of osteoporotic fractures.
<b>Link</b>	<a href="http://hdl.handle.net/1826/1032">http://hdl.handle.net/1826/1032</a>
<b>Type of Object</b>	Thesis or dissertation Doctoral PhD
<b>Object format</b>	5426642 bytes 13370470 bytes application/pdf application/pdf
<b>Language</b>	en
<b>Publisher</b>	Cranfield University Defence College of Management and Technology; Cranfield Postgraduate Medical School; Department of Materials and Medical Sciences

Also available in XML

[All Repositories](#)

**FIGURE 22: CLICKING ON THE "BIBLIOGRAPHIC DATA" LINK (SEE PREVIOUS IMAGE) THIS REVEALS THE METADATA DELIVERED BY THE REPOSITORY IN SIMPLE DUBLIN CORE**



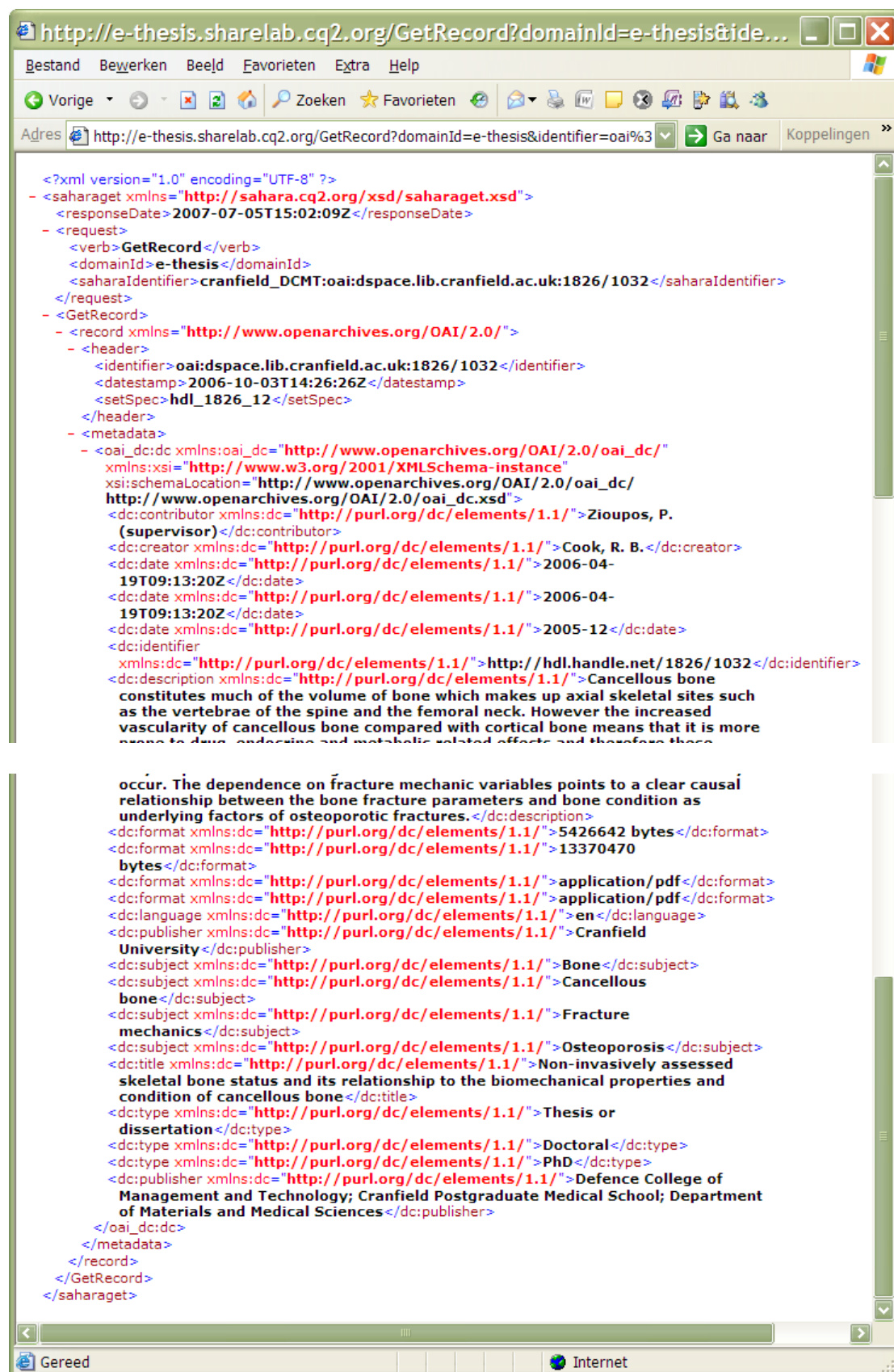


FIGURE 23: THIS SHOWS THE ACTUAL METADATA IN XML FOR THIS RECORD FROM CRANFIELD.

Cranfield QUEprints: Item 1826/1032

File Edit View History Bookmarks Window Help

https://dspace.lib.cranfield.ac.uk/handle/1826/1032 Google

Apple Amazon eBay Yahoo! News (111)

**Cranfield UNIVERSITY**

[Home](#) | [Help](#) | [Feedback](#)

**Cranfield QUEprints**  
The home of research excellence

[Cranfield QUEprints](#) > [Defence College of Management and Technology, Shrivenham](#) > [PhD and EngD theses - DCMT, Shrivenham](#) >

**Search QUEprints**

[Advanced Search](#)

**Browse**

[Communities & Collections](#)

[Titles](#)

[Authors](#)

[By Date](#)

**Sign on to**

[Receive email updates](#)

[My QUEprints](#)

[Edit Profile](#)

**Campus libraries**

[Cranfield](#)

[Shrivenham](#)

**Please use this identifier to cite or link to this item:**  
<http://hdl.handle.net/1826/1032>

**Title:** Non-invasively assessed skeletal bone status and its relationship to the biomechanical properties and condition of cancellous bone

**Authors:** Zioupos, P. (supervisor)  
Cook, R. B.

**Keywords:** Bone  
Cancellous bone  
Fracture mechanics  
Osteoporosis

**Issue Date:** Dec-2005

**Publisher:** Cranfield University

**Abstract:** Cancellous bone constitutes much of the volume of bone which makes up axial skeletal sites such as the vertebrae of the spine and the femoral neck. However the increased vascularity of cancellous bone compared with cortical bone means that it is more prone to drug, endocrine and metabolic related effects and therefore these skeletal sites are more prone to the bone condition osteoporosis. With the bone condition osteoporosis increasing in prevalence it is becoming far more important not only for those at risk of having the condition to be diagnosed earlier, but also for the effects of the condition to be better understood. There is a need for the better clinical management of fractures and for therapies and medical practices that will best avoid the low trauma fractures that

relationship between the bone fracture parameters and bone condition as underlying factors of osteoporotic fractures.


**URI:** <http://hdl.handle.net/1826/1032>

**Appears in Collections:** [PhD and EngD theses - DCMT, Shrivenham](#)

**Files in This Item:**

File	Description	Size	Format	
Thesis.pdf	Thesis text	5299Kb	Adobe PDF	<a href="#">View/Open</a>
Appendices.pdf	Appendixes	13057Kb	Adobe PDF	<a href="#">View/Open</a>

[Show full item record](#)



All items in QUEprints are protected by copyright, with all rights reserved.

Published by Library & Information Service at Cranfield Campus. Copyright © Cranfield University

**FIGURE 24: WHEN CLICKING ON THE TITLE OF A RECORD IN THE RESULT VIEW, ONE IS REDIRECTED TO THE JUMP-OFF PAGE FROM CRANFIELD REPOSITORY. THIS REDIRECTION LINK IS DEPENDENT ON THE CONTENT OF THE DC:IDENTIFIER FIELD IN THE PREVIOUS IMAGE. CLICKING ON THE "VIEW/OPEN" LINK PROVIDES THE PDF.**

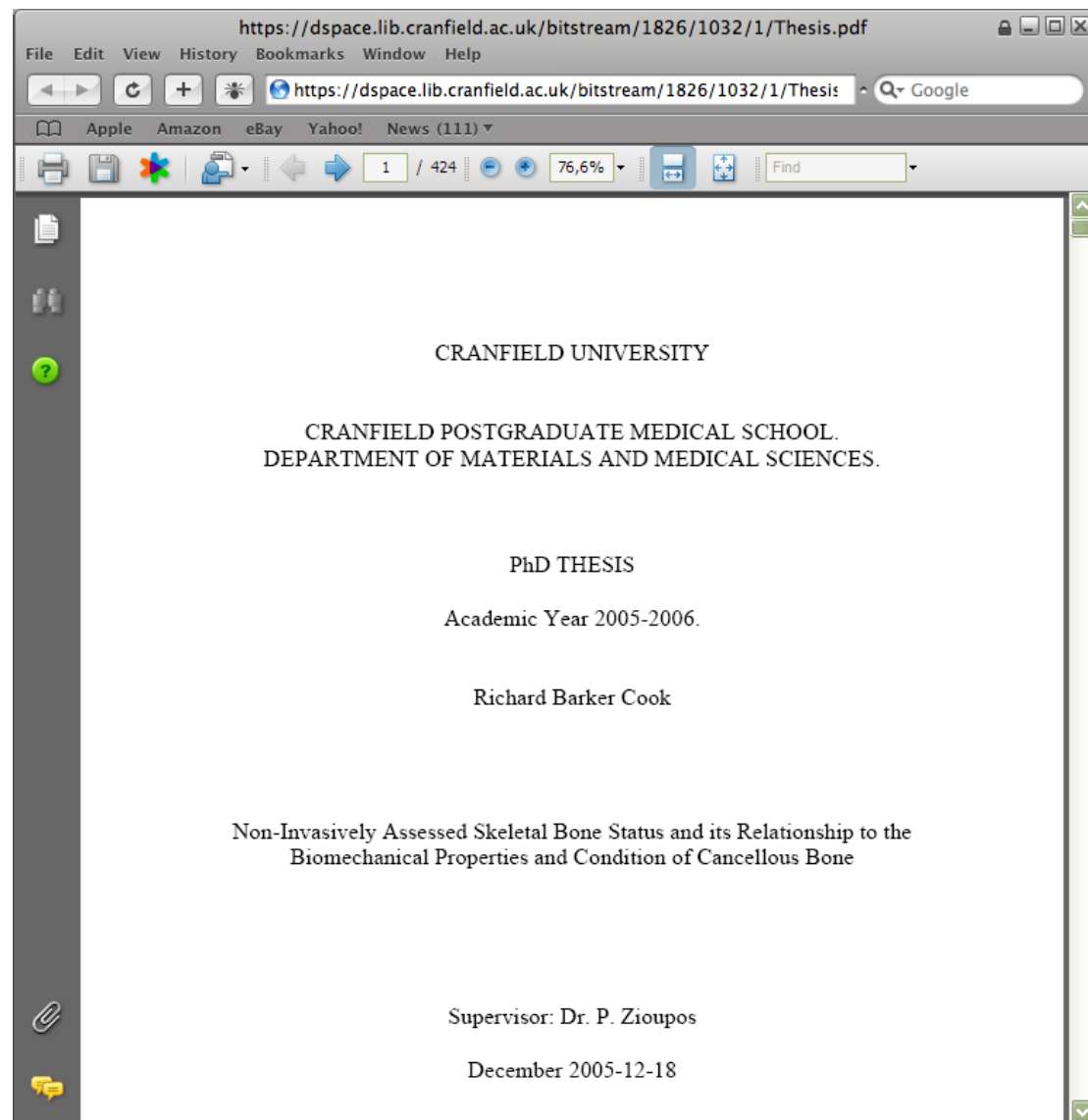


FIGURE 25: ACCESS TO THE PDF IN 3 CLICKS.

## LOGGING IN FOR THE HARVESTER

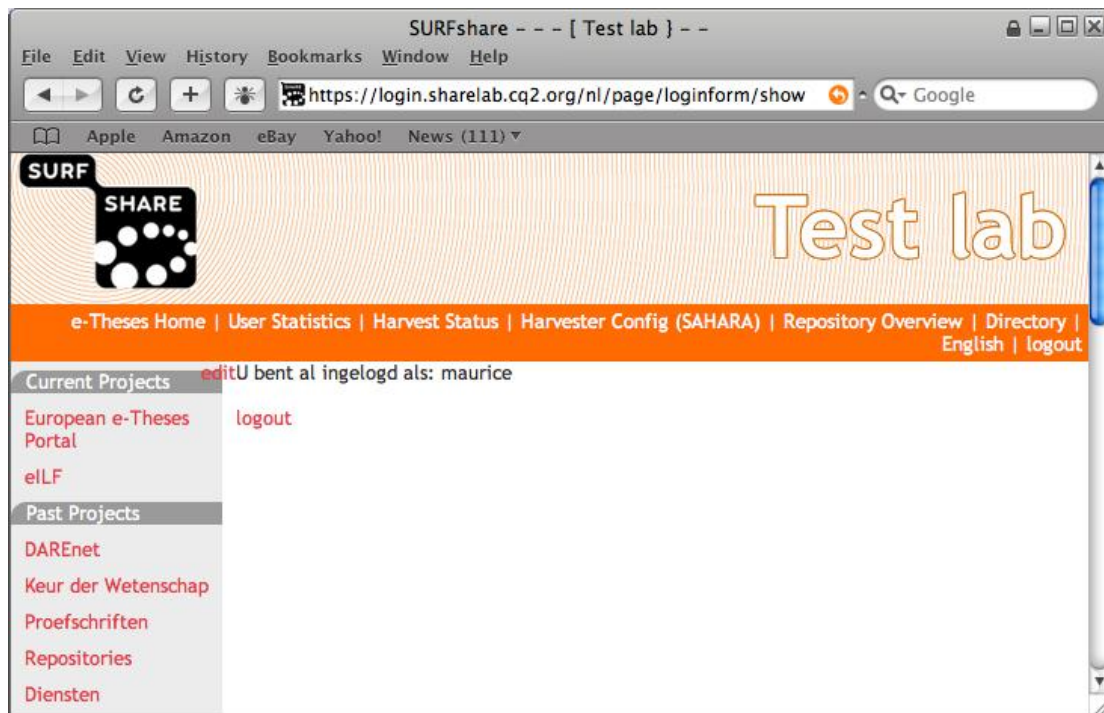


FIGURE 26: LOGIN AT THE FRONT-END

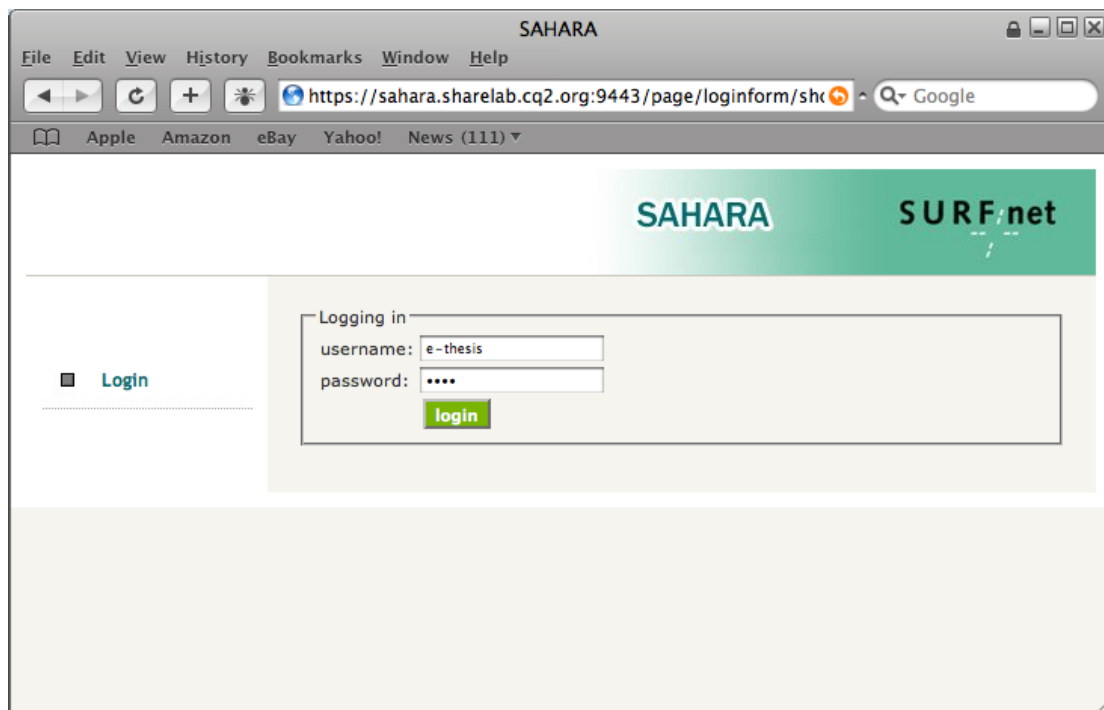
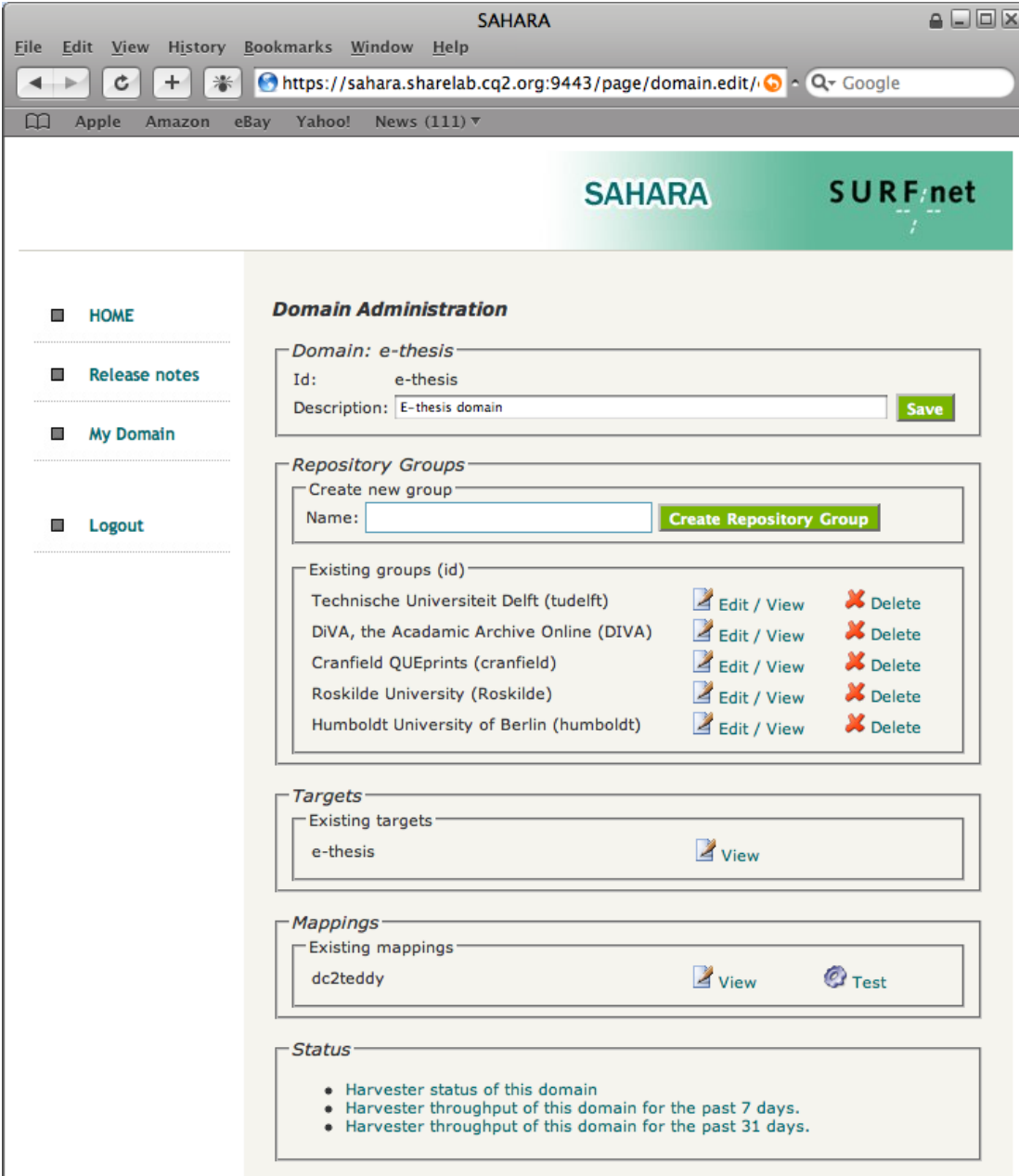












FIGURE 27: LOGIN AT THE HARVESTER BACK-END




The screenshot shows a web browser window with the address bar displaying `https://sahara.sharelab.cq2.org:9443/page/domain.edit/`. The browser's address bar also shows a Google search bar. The page title is "SAHARA". The browser's menu bar includes File, Edit, View, History, Bookmarks, Window, and Help. The browser's toolbar includes back, forward, home, and search buttons. The browser's status bar shows a list of bookmarks: Apple, Amazon, eBay, Yahoo!, and News (111).

The main content area of the browser displays the SAHARA application interface. The interface has a green header bar with the text "SAHARA" and "SURF/net". Below the header bar, there is a left sidebar with a menu containing the following items: HOME, Release notes, My Domain, and Logout. The main content area is titled "Domain Administration" and contains several sections:



- Domain: e-thesis**: This section contains a form with two fields: "Id" (with the value "e-thesis") and "Description" (with the value "E-thesis domain"). There is a "Save" button next to the Description field.
- Repository Groups**: This section contains a "Create new group" form with a "Name:" field and a "Create Repository Group" button. Below this, there is a table of existing groups:

Existing groups (id)	Edit / View	Delete
Technische Universiteit Delft (tudelft)	 Edit / View	 Delete
DiVA, the Academic Archive Online (DIVA)	 Edit / View	 Delete
Cranfield QUEprints (cranfield)	 Edit / View	 Delete
Roskilde University (Roskilde)	 Edit / View	 Delete
Humboldt University of Berlin (humboldt)	 Edit / View	 Delete

- Targets**: This section contains a table of existing targets:

Existing targets	View
e-thesis	 View

- Mappings**: This section contains a table of existing mappings:

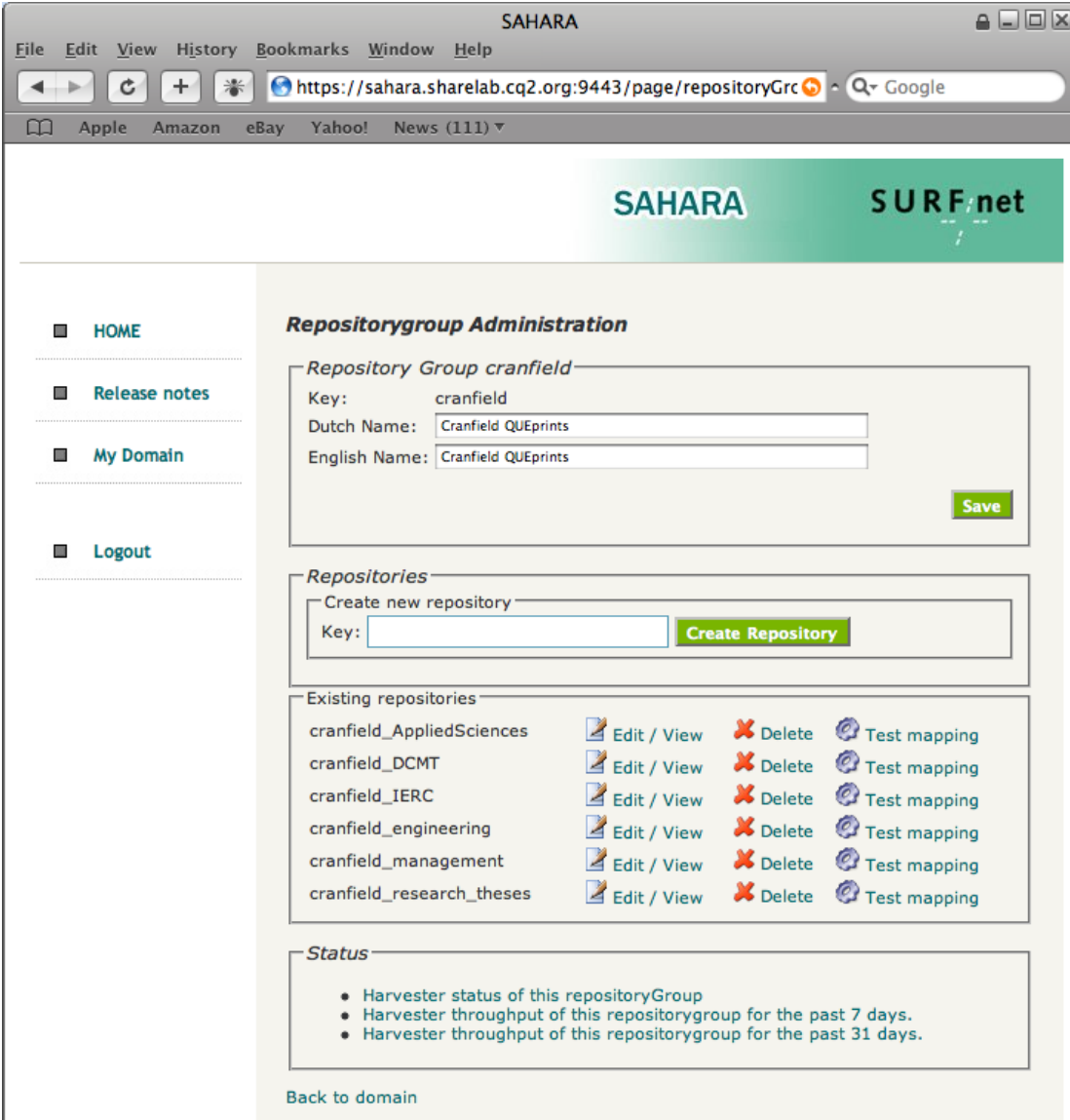
Existing mappings	View	Test
dc2teddy	 View	 Test

- Status**: This section contains a list of status items:

- Harvester status of this domain
- Harvester throughput of this domain for the past 7 days.
- Harvester throughput of this domain for the past 31 days.

**FIGURE 28: THIS SHOWS THE HARVEST DOMAIN. YOU CAN SEE THE REPOSITORY GROUPS OF TUDELFT, DIVA, CRANFIELD, ROSKILDE AND HUMBOLDT**





The screenshot shows a web browser window with the address bar displaying `https://sahara.sharelab.cq2.org:9443/page/repositoryGrc`. The page title is "SAHARA" and the logo "SURF/net" is visible in the top right. The left sidebar contains navigation links: HOME, Release notes, My Domain, and Logout. The main content area is titled "Repositorygroup Administration" and shows the "Repository Group cranfield" configuration. The configuration fields are: Key: cranfield, Dutch Name: Cranfield QUEprints, and English Name: Cranfield QUEprints. A "Save" button is present. Below this is the "Repositories" section, which includes a "Create new repository" form with a "Key:" field and a "Create Repository" button. Underneath is a table of "Existing repositories" with columns for repository names and actions (Edit / View, Delete, Test mapping). The table lists six repositories: cranfield\_AppliedSciences, cranfield\_DCMT, cranfield\_IERC, cranfield\_engineering, cranfield\_management, and cranfield\_research\_theses. At the bottom, the "Status" section lists three items: Harvester status of this repositoryGroup, Harvester throughput of this repositorygroup for the past 7 days, and Harvester throughput of this repositorygroup for the past 31 days. A "Back to domain" link is at the bottom left.

**FIGURE 29: SELECTING THE CRANFIELD GROUP SHOWS US A NUMBER OF SETS WE NEED TO HARVEST THAT CONTAIN DOCTORAL THESES. CRANFIELD HAS MORE SETS, BUT ARE USELESS FOR OUR PURPOSE.**

SAHARA

File Edit View History Bookmarks Window Help

https://sahara.sharelab.cq2.org:9443/page/repository.ed Google

Apple Amazon eBay Yahoo! News (111)

SAHARA

SURFnet

HOME

Release notes

My Domain

Logout

### Repository Administration

Repository: *cranfield\_AppliedSciences*

Id: cranfield\_AppliedSciences

BaseUrl:

Set:

Metadata Prefix:

Mapping:

Target:

Target Collection:

Harvest: ☒

Next action:

#### Closing hours

Week	Day	Begin	End
<input type="text" value="Any week"/>	<input type="text" value="Any day"/>	from <input type="text" value="0"/> :00	until <input type="text" value="0"/> :00 hrs

#### Status

- Harvester status of this repository
- Harvester throughput for this repository for the past 7 days.
- Harvester throughput for this repository for the past 31 days.

#### Useful links

- List all metadata formats.
- Test mapping

(You may need to save first.)

[Back to repositorygroup](#)

**FIGURE 30: CLICKING ON A CRANFIELD REPOSITORY HERE WE CAN INSERT THE HARVEST INFORMATION. BASE URL, THE PARTICULAR SET, THE DESIRED METADATA FORMAT, THE CONFIGURATION HOW THE METADATA IS BEING MAPPED AND THE TARGET, THERE THE METADATA IS BEING PUT.**

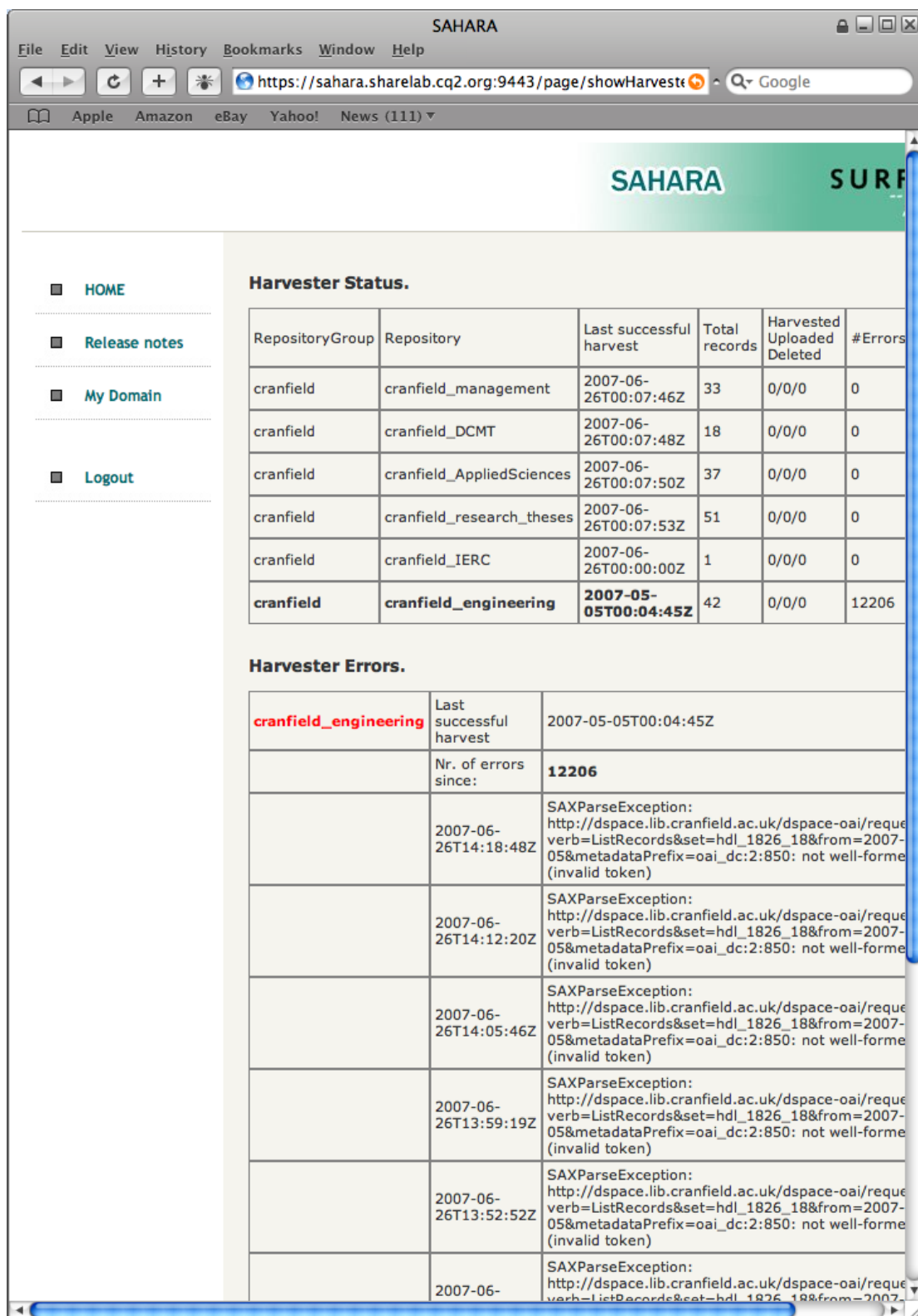
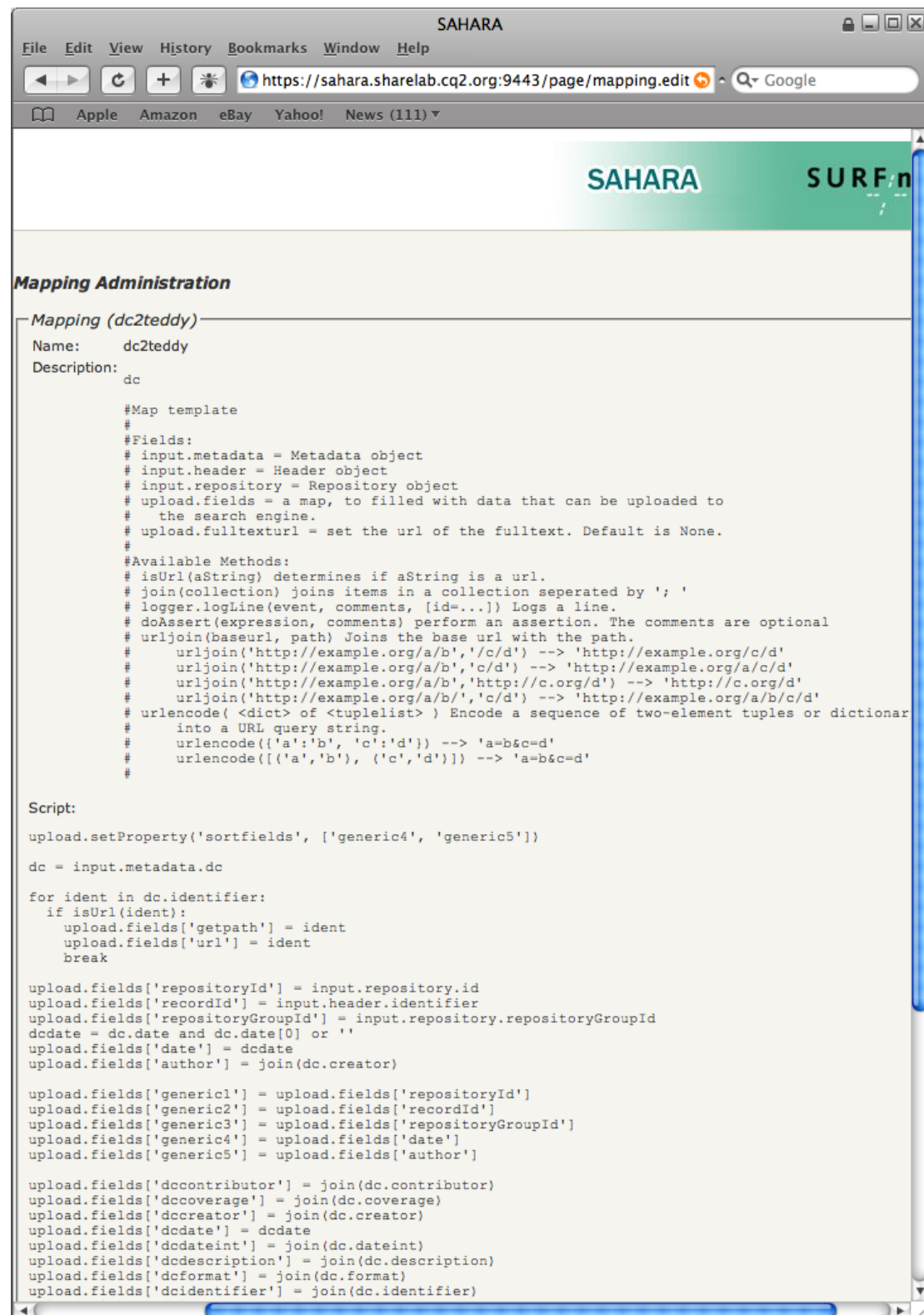


FIGURE 31: THE HARVESTER CAN GIVE INFORMATION ABOUT THE STATUS OF A REPOSITORY. THIS TIME THE ENGINEERING SET OF CRANFIELD GIVE SOME ERRORS.



**FIGURE 32: THE MAPPING IS DONE IN A PYTHON SCRIPT THAT LOOKS FOR A FIELD, DOES SOMETHING WITH IT AND DEFINES AN OUTPUT FORMAT FOR THE CONTENT OF THE FIELD.**

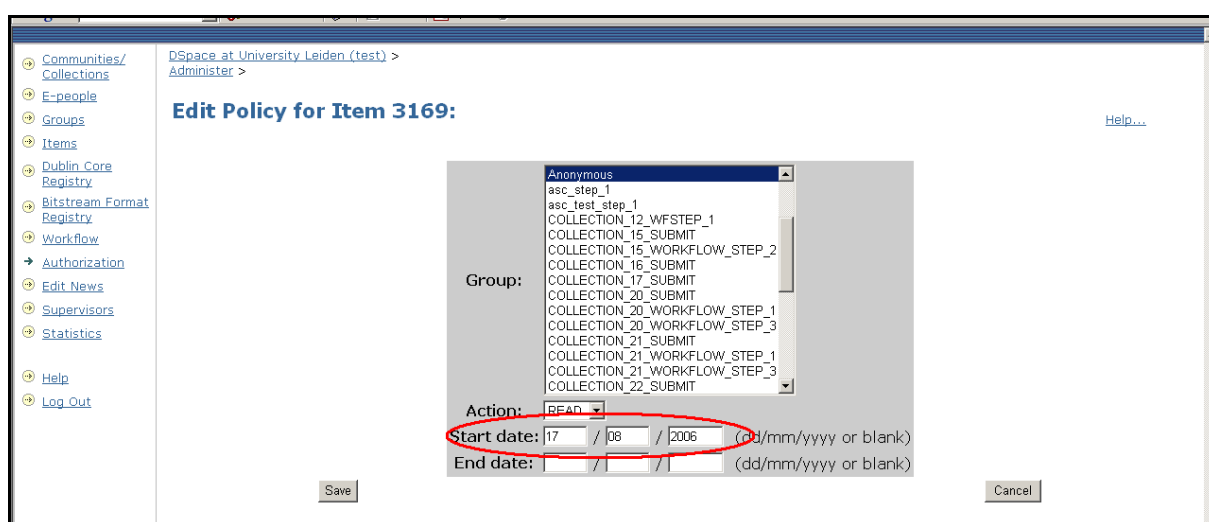
## VI. EXAMPLE OF EMBARGO HANDLING

This section shows how the Dspace Institutional Repository at the University of Leiden handles embargo's to meet DAREnet Open Access criteria. One criteria is that DAREnet only likes to harvest records that point to a full text document free to download.

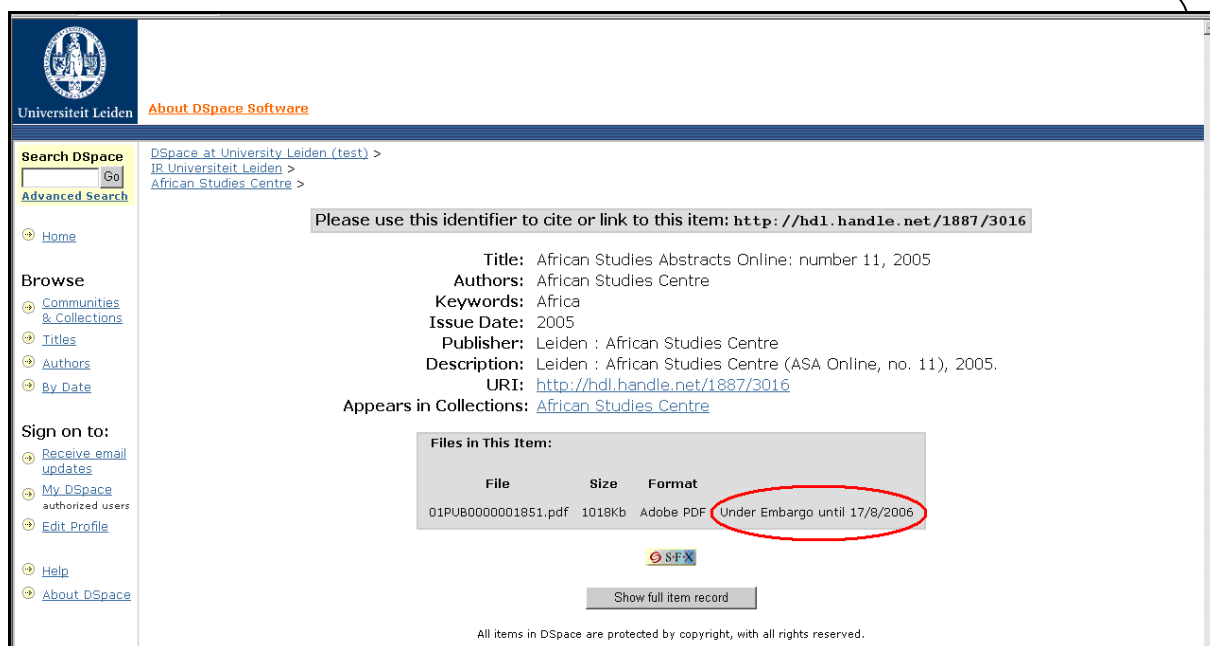
To prevent the fact that DAREnet will harvest publications with an embargo, the University of Leiden chose to make the following actions:

First of all one or more parts of the publication can be placed under embargo, by placing a public release date. When the date is reached, the part is released to the public. When all parts are released from embargo, the record is placed in a separate set. This set can be the DAREnet set with specific Open Access criteria (involving to have the full text available).

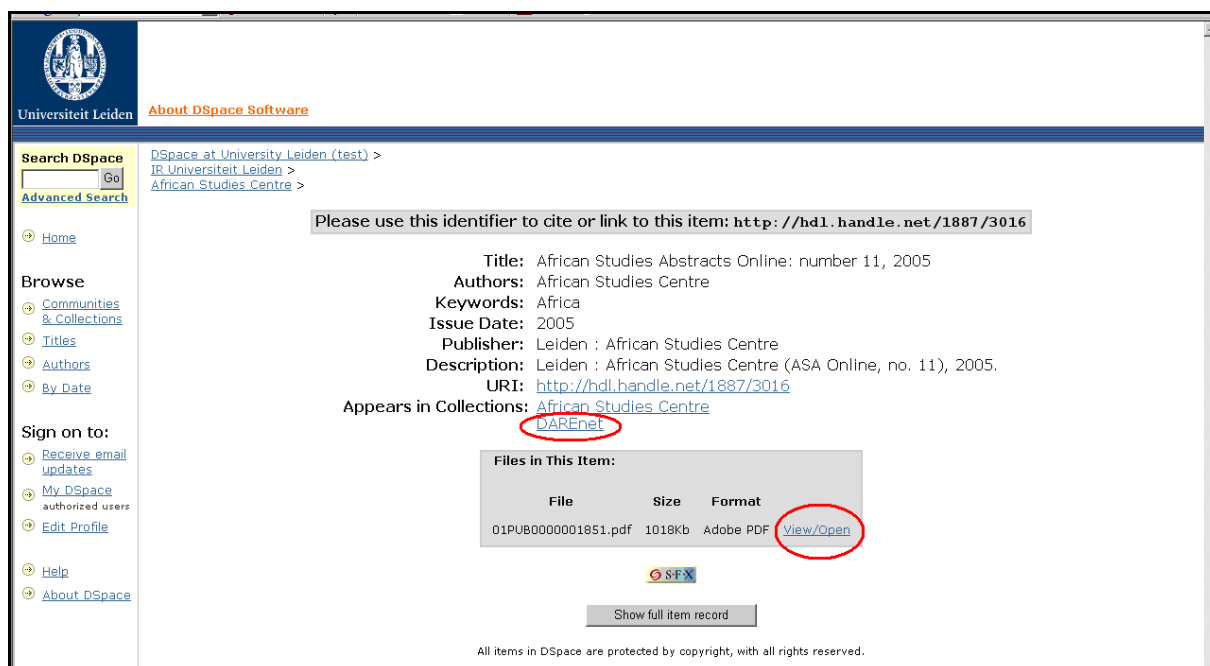
Below, screenshots of the embargo handling can be seen.



**FIGURE 33: A PART OF THE PUBLICATION IS SET TO 17 AUGUST 2006**



**FIGURE 34: UNTIL THE 17TH OF AUGUST, THIS SCREEN IS PRESENTED TO THE USER. (USER CANNOT DOWNLOAD THE FILE, AND THIS RECORD IS NOT VISIBLE UNDER THE DARENET SET)**



**FIGURE 35: AFTER THE 17TH OF AUGUST THIS SCREEN IS PRESENTED TO THE USER. THE PUBLICATION IS VISIBLE, AND THE RECORD IS PLACED IN THE DARENET SET TO BE HARVESTED BY DARENET.**

## VII. LIST OF TERMS

The table below will contain a list of terms and abbreviations that is used in this report.

Term	Explanation
DIDL document	An MPEG-21 wrapper structure to describe compound objects. It is independent of metadata formats and can be used to relate digital objects. These object can be resource locations of bitstreams and metadata formats.
DRIVER project	<p>DRIVER, the “Digital Repository Infrastructure Vision for European Research” is pursued by an EC funded consortium that is building an organisational and technological framework for a pan-European data-layer enabling the advanced use of content-resources in research and higher education. DRIVER develops a service infrastructure and a data-infrastructure.</p> <p>Both are designed to orchestrate existing resources and services of the repository landscape.</p> <p>More information on <a href="http://www.driver-repository.eu">www.driver-repository.eu</a></p>
DRIVER guidelines	<p><i>The DRIVER Guidelines for Content Providers: Exposing textual resources with OAI-PMH</i> shall provide orientation for managers of new repositories to define their local data-management policies, for managers of existing repositories to take steps towards improved services and for developers of repository platforms to add supportive functionalities in future versions.</p> <p>More information on: <b>Fout! De hyperlinkverwijzing is ongeldig.</b></p>
ETD	Abbreviation for Electronic Theses and Dissertation. In our context we use the term for electronic Doctoral theses.
SAHARA	<p>An Open Source robust Harvester used in the DARE program</p> <p>More information: <a href="http://www.uitwisselplatform.nl/projects/sahara/">www.uitwisselplatform.nl/projects/sahara/</a></p>
OAI-PMH Harvester	A piece of software that is used by Service Providers to make OAI-PMH service requests to harvest the metadata from <i>Data Providers</i> .

Repository	Repositories are <i>Data Providers</i> that expose structured metadata via OAI-PMH
SURFfoundation	<p>SURF is the collaborative organisation for higher education institutions and research institutes aimed at breakthrough innovations in ICT. SURF provides the foundation for the excellence of higher education and research in the Netherlands.</p> <p>More information: <a href="http://www.surffoundation.nl">www.surffoundation.nl</a></p>
JISC	<p>JISC's (Joint Information Systems Committee) mission is to provide world-class leadership in the innovative use of Information and Communications Technology to support education and research.</p> <p>More information: <a href="http://www.jisc.ac.uk">www.jisc.ac.uk</a></p>
DIVA	<p>DiVA, the Academic Archive Online (Digitala Vetenskapliga Arkivet in Swedish) is a collaborative effort of a number of universities in Scandinavia which offers both publishing services and technical solutions for local repositories.</p> <p>More information: <a href="http://www.diva-portal.org/about.xsql">www.diva-portal.org/about.xsql</a></p>
IR	Abbreviation for Institutional Repository
Crosswalk	A table that maps the relationships and equivalencies between two or more metadata formats. Crosswalks or metadata mapping support the ability of search engines to search effectively across heterogeneous databases, ie crosswalks help promote interoperability. Source: <a href="#">Dublin Core Metadata Initiative (DCMI) – Glossary</a>
Mapping	See “Crosswalk”. The terms are used mixed in the report, but are analogue to each other.
DARE	<p>The Digital Academic Repositories (DARE) programme is an initiative by the joint Dutch universities that was started in 2003 to make all their research results digitally accessible.</p> <p>More information: <a href="http://www.surffoundation.nl/smartsite.dws?ch=ENG&amp;id=5377">www.surffoundation.nl/smartsite.dws?ch=ENG&amp;id=5377</a></p>
OAI	The Open Archives Initiative.



	<p>The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open access and institutional repository movements. Continued support of this work remains a cornerstone of the Open Archives program. Over time, however, the work of OAI has expanded to promote broad access to digital resources for eScholarship, eLearning, and eScience.</p> <p>More information: <a href="http://www.openarchives.org">www.openarchives.org</a></p>
OAI-PMH	<p>The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. <i>Data Providers</i> are repositories that expose structured metadata via OAI-PMH. <i>Service Providers</i> then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP.</p> <p>More information: <a href="http://www.openarchives.org/pmh/">www.openarchives.org/pmh/</a></p>

More terms can be found at the website from the [Dublin Core Metadata Initiative \(DCMI\)](http://dublincore.org/documents/usageguide/glossary.shtml)

<http://dublincore.org/documents/usageguide/glossary.shtml>