Enhanced E-theses Project

Deliverable 10

Considerations with regard to the application of local search engines within an OAI-ORE environment for Enhanced E-theses

Document Description

Project	
Title:	Enhanced E-theses
Start date:	1st December 2007
Funding Agency	Knowledge Exchange
Document	
Deliverable number:	D10
Deliverable title:	Considerations with regard to the application of local search engines within an OAI-ORE environment for Enhanced E-theses
Actual Date of Delivery:	February 2009
Author:	Eric Sieverts (Utrecht University Library)
Editor(s):	Martin Slabbertje (Utrecht University Library)
Workpackage:	WP10
Workpackage title:	Interoperability between data- and service providers in terms of the exchange of complex objects
Version/Revision:	1.1
Draft/Final:	Final
Total number of pages: (including cover)	6
File name:	Enhanced Etheses project_deliverable 10_v1.1.doc

Introduction

In any digital information environment, **search** is considered a vital and mandatory functionality. In order to be of proper value, search must be carefully adjusted to the specific character and structure of the searchable material. For a proper configuration of search functionality in an ORE-based environment, as in the Enhanced E-theses project, we will discuss the following issues.

1. Granularity issue: what are the individual units to be retrieved?
2. Indexing issue: what portions of text and/or metadata and/or documents must be made searchable?
3. External aggregation issue: what material in a link-based environment (as ORE actually is) that is located outside the local collection, must also be indexed?
4. Ranking issue: what additional factors, specific for the ORE-structure, should be taken into account, if the search system supports relevance ranking?
5. Presentation issue: which factors concerning granularity and linking should be taken into account when presenting search-results?
6. Functionality issue: what advanced search functionality - above simple Google-like search - is useful to implement?

1. Granularity

The concept of what are to be considered as individual items, is an important one for retrieval systems. First of all in order to decide what should be considered as units within which terms must occur, for satisfying Boolean or other term co-occurrence relations. In the second place, these units determine what the individual retrieval results are, which should be presented to the user.

In many situations, it is obvious what the individual items are: these can be the individual records in a database-structured retrieval system, the individual text files or e-mail messages in a desktop search or the individual web pages in a web search engine. In the present situation with compound objects, it is less obvious, however.

In the ORE-framework aggregations of resources are specified. In principle each resource can be an aggregation of other resources. In this environment, it seems most appropriate to allow any resource to be a retrievable item, whether it is a comprehensive aggregation of for instance a complete thesis, or just a small component, like an illustration or a dataset, which itself has no further aggregation of smaller components.

For one thing, this would be a pragmatic choice. If we wanted only comprehensive "top level" resources to be separately retrievable items, an automatic system would have to determine autonomously what the appropriate parent aggregations are, which should serve as "top level" resources. Such an automated decision would be almost impossible to make, since there is no formal attribute that characterizes a resource as "top level", whereas even very comprehensive resources can in principle be aggregated themselves within another resource and, formally speaking, would therefore not be "top level". Moreover, from a retrieval point of view, such a choice has the advantage that retrieved results are always most closely connected to the precise context of the subject query: it is better to retrieve the pertinent chapter or section that contains the answer to a query, than to find a whole book or thesis. In principle, there would be a disadvantage as well, since a retrieved result set could easily contain duplicates: a parent resource and some connected sibling resources, all of them satisfying the query, would be presented as separate entities.

In the specific context of the present Enhanced E-theses project, the situation is even somewhat different. In the next section it will be discussed that it is most appropriate to index just the metadata connected to resources, instead of the full-text of their document files. In the Enhanced E-theses framework, however, comprehensive metadata are only connected to the theses as a whole. In this case, the choice to consider any resource that has metadata connected to it, as a possibly retrievable item, is a pragmatic one for another reason as well. Generally, the lower level sibling components will have such sparse metadata connected to them, that their chance to be retrieved on any query, will mostly be negligibly small, compared to the theses as a whole, so that the disadvantage of retrieved duplicates, mentioned in the previous paragraph, will practically not occur.

2. Indexing

Besides the granularity issue, equally important for any retrieval system is what portions of digitally available textual data should be indexed. In many situations the answer to this question is quite obvious. If the items just consist of textual documents, without explicit metadata portions, it is mostly the full-text of the documents themselves that should be indexed. In situations where both full-text and sufficiently comprehensive metadata are present, a proper decision must be made, however. This is even more compelling in situations where metadata and full-text documents are stored in separate instances and sometimes even at separate locations. In the present case, the compound character of the documents complicates the issue even further. The full-text of an item (an e-thesis) can be split up over a number of separate files, for instance the chapters. And even separate chapters of a thesis can have separate metadata associated with them. As such, the considerations about this issue are closely related to those discussed for the granularity issue.

In general, the indexing of full-text documents themselves is the easiest way to make them searchable. This procedure does not necessarily generate the best search results, however. Generally, searching in more condensed representations of the content of documents, like abstracts, is considered to result in better precision (and consequently better user satisfaction).
Therefore, there is often a tendency to search metadata instead of full-text, under the necessary condition that sufficiently comprehensive metadata are available, including good quality abstracts.

Besides the argument above to restrict search simply to metadata representations of full-text documents, as derived from general retrieval practice, there is also an argument that is specific for the ORE-environment. Aggregations can contain various versions of the same text documents, either in different formats (e.g. PDF, MS-Word and HTML versions) or versions representing the successive stages in the course of their evolution. Indexing all of them would provide (near) duplicates in retrieval results, if no special measures are taken to merge them in result lists.

In the OAI-ORE environment there are separate aggregations for metadata, encoded in XML. Although there is no formal characteristic for metadata files, these metadata can be recognized from the fact that they are XML files and (at least for Dublin Core metadata within the Enhanced E-theses project) that they are of the objecttype info:eu-repo/semantics/descriptiveMetadata. Metadata need not always to be just DC; also MODS metadata sets can exist. Since a resource can have two (or even more) different metadata resources connected to it, the respective metadata files will be combined within an aggregation as well. If for the same resource, metadata files according to more standards are present, it must decided whether all of them should be indexed, or that priorities have to be specified.

When limiting the indexing of Enhanced E-theses to just metadata resources, a necessary condition must be satisfied: the metadata must be sufficiently rich to serve as a comprehensive representation of the contents of the documents. This means that either authors themselves must provide such rich metadata, or mechanisms for automatic feature extraction must be incorporated in the workflow, so that abstracts or fingerprints are being generated from full-text documents and included in the metadata. Also specific elements of the full-text documents themselves could already serve this purpose, by copying them to the metadata. These could be a formal abstract, an introduction that contains the main topics covered in the publication, or a table of contents where the text of section headings represents all the topics. Metadata fields which could contain these derived or generated content representations are for instance the Dublin Core <dc:description> field or the MODS <abstract> or <tableOfContents> fields.

3. External aggregations

In order not to have to index the whole universe, indexing will in principle be restricted to (metadata of) resources that are stored in the repository of the own organisation. However, it may occur that

- a local resource also aggregates resources from external repositories,
- a local resource is aggregated within a resource from an external repository.

The first situation will immediately be visible, the latter can be recognised by an "ore:isAggregatedBy" attribute, although we must realize that in practical situations this attribute will often be missing, since it is not yet automatically generated.

In these situations it has to be decided whether (metadata of) these "linked" or "backlinked" external resources must be indexed as well. If this is the case a next question is how many levels of aggregation-links must be followed. For the present Enhanced E-theses application, it is suggested to restrict external indexing to a single level, both when linking "down" to an aggregated resource and when linking "up" to an aggregating resource. Additional requirements for such external resources to be indexed, is that they are freely accessible and that they also consist of metadata.

4. Ranking

Modern retrieval systems use techniques for relevance ranking. Results with the largest probability to be relevant are presented first. Various factors relating to the words in the query and the query terms that are present in the retrieved documents, play important roles in most ranking models. However, Google has taught us that also other factors, like linking patterns can be an important addition to the ranking rules.

In the OAI-ORE framework linking between document components plays an important role, but it is a completely different one than the role of hyperlinks in web pages. Therefore it is not evident that aggregation linking should be taken into account for ranking in the present situation. On the other hand, it can not be ruled out that a model can be conceived in which ranking can improve by taking into account aggregation linking patterns and, for instance, the presence of multiple versions of resources. This requires further study, however.

5. Presentation

When separate components of documents or various versions of them are being retrieved as separate items, there must be a mechanism to collapse them to single units. On the other hand it is useful if it is visible which separate parts of the document satisfy the query. So it must be possible to unfold these components when required, as well. For this purpose the proposed mechanism to automatically generate maps of the aggregation structure of the resources, on the basis of "aggregates" and "aggregatedBy" attributes, can play a role here as well.

6. Search functionality

Besides a standard Googlian "all fields" no-operator syntax, most search systems also provide advanced search screens. They mostly support simple Boolean search and some fielded search, dependent on the presence of structure in the searchable material. In the present Enhanced E-theses environment, the metadata records contain a number of fields, for which a field-limited search may be useful. Which fields could serve this purpose should further be analysed on the basis of the metadata models in use.
In addition it will be useful if search results can be limited to theses that contain for instance sound or video files or data sets.