ECHO
European Cultural Heritage Online
WP2 ECHO Search Engine
WP2-TR18-2004 - Version 2

WP2 Note on the DORA Search Engine

Peter Wittenburg
May 18, 2004

In two reports we have described the DORA[1] concept and the underlying mapping scheme (WP2-TR16-2004) and its ontology components (WP2-TR17-2004). In this document we want to describe the search engine and summarize its evaluation[2]. While the DORA document describes the intentions and possibilities, this document describes what was implemented. It is not a technical documentation, but describes to a certain detail which implementation decisions were taken and which problems were encountered. The search engine is based on the mappings as described in the DORA note and in the Ontology note, i.e., it implements the mappings and semantic relations in specific ways to achieve high performance.

The evaluation part has to consider two aspects: (1) The formal correctness of the algorithms have to be checked and (2) the usefulness and appropriateness of the semantics included in DORA has to be evaluated. Finally, answers to the following two questions have to be given:

- Are the chosen semantic relation useful?
- Does metadata interdisciplinary help to answer questions?
- What kind of infrastructure is necessary to overcome current limitations?

It should be noted here that the included number of records is about 95.000 records and that the distribution is uneven. It is obvious that searching only makes sense in large collections such as delivered from Fotothek (75715 records) and languages (17403 records). The relatively small number of records provided by the other repositories at this moment (20 to 1100) limits the strength of the evaluation. Any data that was offered by the data providers was integrated[3].

1. Search Engine

In this chapter we want to describe the actual DORA interface, the harvesting principles, the data correction steps to be taken, the nature of the index creation process and the searching process. It should be mentioned that the DORA engine is implemented largely with Java[4].

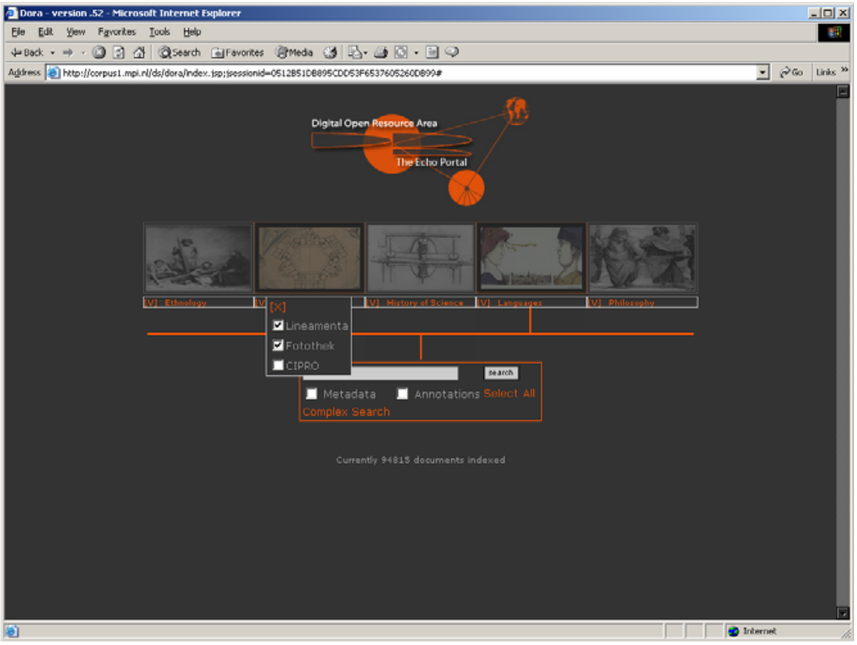
1.1 DORA Interface

The DORA interface was implemented as described in the original DORA document. However, during the ECHO project it became apparent that some of the goals were too challenging to be met within the short period of time. Everyone interested can make use of the DORA engine, it is available under the following URL:

http://corpus1.mpi.nl/ds/dora/

[1] Digital Open Resource Area: see WP2-TR16-2003; web-site to come
[2] The evaluation will be updated in May 2004
[3] In the case of the RMV repository it is being checked why not more than the current 20 records can be harvested.
[4] A technical documentation will go into more detail

The user can select the disciplines and within the disciplines the data providers to be included in the search. The disciplines are indicated by images and the data providers by menu lists. The interface offers two search options: (1) In simple search the user can specify words that are searched for in all metadata fields provided including full-text fields that contain prose-text. (2) In complex search the user can select a view that is derived from the vocabulary used by the different data providers. All details of these views are explained in the DORA note.

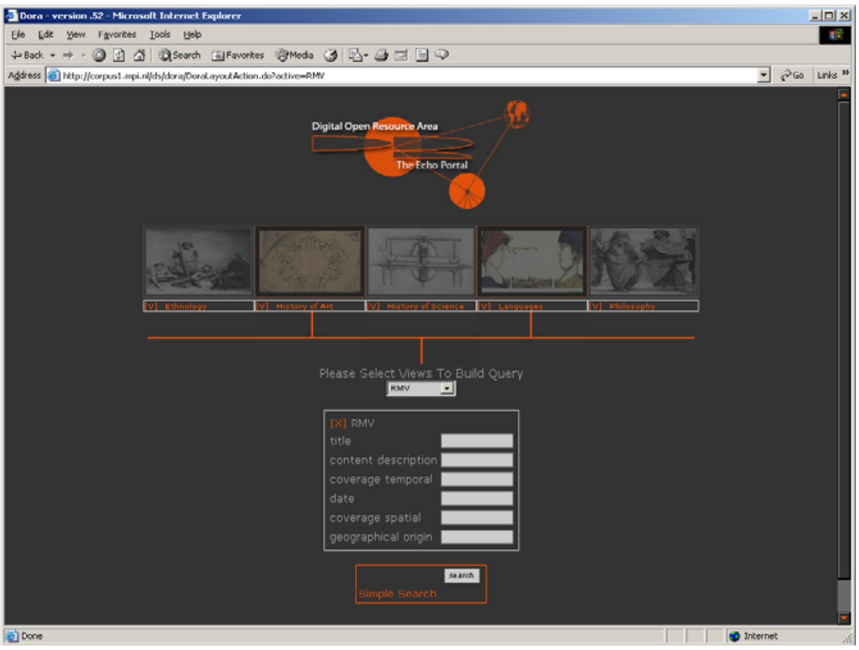

Dora - version .52 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://corpus1.mpi.nl/ds/dora/index.jsp;jsessionid=0512B51DB895CDD53F6537605260D899#
Go Links
Digital Open Resource Area
The Echo Portal
[V] Ethnology
[V
[V] History of Science
[V] Languages
[V] Philosophy
search
Internet

[X]
☑ Lineamenta
☑ Fotothek
☐ CIPRO

☐ Metadata ☐ Annotations Select All

Complex Search

Currently 94815 documents indexed


Dora - version .52 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://corpus1.mpi.nl/ds/dora/DoraLayoutAction.do?active=RMV
Go Links
Digital Open Resource Area
The Echo Portal
[V] Ethnology
[V] History of Art
[V] History of Science
[V] Languages
[V] Philosophy
search
Done
Internet

Please Select Views To Build Query

RMV

[X] RMV
title
content description
coverage temporal
date
coverage spatial
geographical origin

Simple Search

Originally, it was intended to include browsing, geographical browsing and annotations in the search. These features were not implemented. Languages is the only domain where browsing is made available so here it is makes sense to go to the language portal immediately. The geographical browsing turned out to be too difficult to be implemented in the ECHO period. Due to the large scale difference (continents to maps of ancient Rome) we would have needed scalable maps that allow to step down to details of Rome and it was seen as too much work to provide the exact coordinates of all locations involved in the DORA domain. Metadata descriptions do not yet include formal geographical coordinates such that points could be created automatically.

The option to search on annotations is provided and it would not be too difficult to add annotations to the index, however, it is not as effective. Also here some plans were too ambitious to be realized in the short ECHO period. The idea in history of science was to relate web-sites with each other by entering typed relations. These annotations would be very excellent resources to be integrated in searches. Yet no data could be created.

It should be mentioned that the interface is configuration file driven, i.e., it can be easily adapted to other configurations that would imply other

- disciplines
- data providers within them
- views

Every data source in DORA gets an ID which is used as the key to combine different knowledge.

1.2 Harvesting

The way data providers deliver data within ECHO is different as the table indicates.

NECEP	RMV	Languages	Lineamenta	CIPRO	Fotothek	IMSS	Berlin	Philosophy
online XML	online OAI	online XML/OAI	off-line email	off-line email	off-line email	online OAI	not yet up	online XML

Five collections were online and could be harvested according to a various schemes. Three of the interfaces are offering an OAI MHP compliant interface. In the case of languages the XML variant was preferred since it includes all metadata fields. The three data sources extracted files at certain moments and provided them by sending emails. In the latter case a harvesting concept was not applicable.

For those data sources that could be harvested a process file was created. It can be modified in a simple way with the help of a web-interface. The following parameters can be defined via this interface to tune the harvesting engine:

- data provider ID
- frequency of harvesting
- day time to execute the harvesting (hour/minute)
- day to execute the harvesting
- import prefix
- classpath to the data processing programs
- the label of the data provider
- root URL as harvesting address

In addition the file contains parameters such as location of logging information, date and time of last harvesting etc.

The classpath reference is of great importance since it refers to executable code that contains the knowledge about how to grab the data from the specified URL (OAI/XML) and how to preprocess the data delivered from the source.

A log file is created that contains protocol information describing the harvesting process. In addition to the information mentioned above it says how many records were received per source, which type of errors were encountered. This file is also used to document other steps and to protocol the query handling.

1.3 Data Pre-Processing

The data delivered had to be corrected and modified in different ways. Here we can only give a few examples. The purpose of this chapter is not to complain, but to show the problems one is faced with when building an interoperable metadata domain at the various levels. Initiatives such as OAI have a great value, although the metadata harvesting protocol is very simple. Its wide acceptance makes clear to every data provider that it is the task of the data provider to provide correct data and not that one of the service provider. The experience not only in ECHO shows that we are still far away from that goal.

Much effort was due to changes in the data delivered over time. The language domain changed the IMDI version such that new X-paths were necessary and new mappings had to be established. However, this step was an explicit one supported by proper schemas. In many cases changes were done without notice or without providing a schema. Path corrections could only be carried out after visual inspection.

OAI MHP Type of Harvesting (RMV, IMSS)

In the case of OAI harvesting the type of preprocessing was comparatively simple. This has to do with the fact that a validation check is carried out when registering as OAI data provider. A schema has to be provided and the data delivered is validated against this schema, i.e., at the encoding and syntax level correct data can be assumed. Still at the content encoding level some pre-processing had to be carried out, since this is beyond schemas. Due to the limited number of fields in Dublin Core different types RMV chose to package different types of information into one Dublin Core field. During preprocessing this had to be separated again. Also some of the encodings had to be interpreted and modified to separate formal encodings and explanatory (and therefore searchable) strings. In principle, however, the choice of OAI to put all validation errors at the shoulders of the data provider seems to be the best one can do. It requires that the data providers who know their data very well and have the responsibility to clean up all encoding and syntax problems. In general the broad semantic definitions of fields in Dublin Core such as DC:Coverage or DC:Subject make it difficult at the semantic level to create suitable mappings. In some cases it is too early to make statements about the usage of such fields.

XML Type of Harvesting (NECEP, Languages, Philosophy)

In the case of harvesting online available XML data in two cases a schema was available (NECEP, Languages) and validation was carried out by the data provider, so proper metadata was delivered. In the case of philosophy IMDI type of metadata descriptions were created manually from the given texts, therefore also proper schema-based metadata was available. In fact the philosophy data exists from textual descriptions that were interpreted as prose descriptions, i.e., they are not part of the complex search but integrated into the index for simple search.

In the language case a major schema change was done during the DORA work, therefore several utility files containing Xpaths etc had to be adapted. Some repositories such as those created by Lund University within ECHO are still using the old IMDI version, i.e., it had to be noticed which version is used for different parts in the language domain. Therefore, a proper harvesting scheme would have to check regularly the version of the underlying schema to make sure that the settings are still ok. The IMDI import module has the appropriate knowledge and can adapt the import schema, however, the Xpath specifications have to be updated.

Static Harvesting (other providers)

In the case of the other data providers in ECHO static files were exchanged – in general by email. As far as we know XML data was generated by extracting data from relational database repositories of different types. Here many problems were encountered. Again it should be mentioned that our colleagues did their best to provide useful metadata – it's just a picture of the state of technology.

- lack of proper XML headers;
- no UTF-8 character encoding although the XML header claims it[5];
- lack of an XML schema prohibiting any validation;
- invalid XML constructions;
- existence of several XML document headers in one file;

[5] These kind of problems are very serious ones, since during parsing no errors are created. In general errors can only be indicated if searches don't lead to appropriate results. The string "Milano" was not extended due to the geographic thesaurus as subpart of "Italy" and "Europe" since it contained non-UTF-8 character encodings. We assume that some of these errors are still hidden in the index.

- changes of the underlying schema

In the case of the Fotothek it was known that the records are highly nested, so a normalized structure had to be created. It was not always clear to the DORA developers which of the fields had to be replicated.

It became also apparent that the encodings found in the metadata records did not fit with the encodings found in the thesauri for example. Some pre-processing had to be done here as well.

Normalized validated DORA Repositories
Before actually doing any further processing normalized and validated (as far as possible) XML files were created for all repositories. These are part of the DORA ontology, have a documented structure such that the Xpath definitions contained in the various other resources are correct. In general, this pre-processing step was necessary to come to useful repositories, but it took too much time.

When creating these normalized XML files also the punctuation characters were removed from the data to allow proper and easy matching. For presentation purposes the original string is preserved as well.

1.4 Index Creation

Since DORA contains now about 95.000 records and since it can be expected that these numbers will increase rapidly, it was decided to focus on fast indexing mechanisms and to do as much as semantic processing off-line, i.e., not during search. Exploiting the different knowledge components in real time would lead to unacceptable delays. It was decided to use a binary tree where every word found somewhere in the metadata descriptions (including the prose texts) is included as a sequence of nodes. With proper encoding techniques such a tree would guarantee almost equal access times for all queries. It was checked whether an API provided by some of the already existing search engines could be used. Since the search algorithm itself was not seen as the component that would take much time this option was not chosen, i.e., based on existing experience and knowledge a tree-traversing algorithm was programmed.

Before creating the index tree the semantic extension had to take place. To accomplish this first the codes found in the Fotothek and RMV metadata descriptions were replaced by the strings and separated respectively. At the same moment the mapping between the three content thesauri was used to add the appropriate strings (***iconclass2ovm-mapping-v3.xml, ovm2iconclass-mapping-v3.xml, IMDI2iconclass-and-ovm-v1.xml).*** Due to the semantic vagueness of the entries found and of the relations between the thesauri it was decided to not extend to all super-classes in the thesauri. Tests have shown that this would result in an semantic explosion and a decrease in precision[6]. The following example may illustrate the operation.

The following relation is taken from the ***iconclass2ovm-mapping*** file. A specific Iconclass code has relations to two OVM codes.

31D	Iconclass code that maps to OVM classes
human life and its ages	corresponding Iconclass string
OVM.AAC.AAM	OVM code
life cycle	appropriate OVM string
OVM.AAC.AAM.AAA	OVM code
pregnancy, birth and first year	appropriate OVM string

When in a record of the Fotothek repository the entry “31D” is found, it will first be replaced by the corresponding string. Then the two semantically overlapping strings of the OVM thesaurus are added. The resulting entry would be transformed from “31D” to

[6] Here the term “precision” is used known from the field of information extraction. It indicates how many hits were obtained that are inappropriate. A decrease in precision means that too many “wrong” hits were found.

“human life and its ages; life cycle; pregnancy, birth and first year”

In doing so the user would find this entry also if the search string “life cycle” was entered.

For all geographic information a full extension was made. Two thesauri were used: ***ovm-geo-thesaurus-v3.xml; mpi-geo-thesaurus-v4.xml.*** The first is being used for the OVM collection, the second was assembled by looking through all geographically relevant fields including the names of museums, names of languages spoken in that area, etc in the other repositories (for more details we refer to the ontology document). Where possible also other names than the English were added[7]. So if Milano was found, also Milan and Mailand were added.

The ***mpi-geo-thesaurus-v4*** thesaurus also contains mappings to the appropriate categories in the OVM thesaurus. The following example is taken from the ***mpi-geo-thesaurus-v4*** thesaurus.

West Africa	OVM.AAA.AAA.AAE
Benin	OVM.AAA.AAA.AAE.AAA.AAA
Burkina Faso	OVM.AAA.AAA.AAE.AAB.AAA
<lang>Dogon	

It says that Benin and Burkina Faso can be found in West Africa and that the language Dogon is spoken in the area of Burkina Faso. During index creation therefore two types of information were added to an entry such as “Milano”. It would result in the entry

“Milano, Milan, Mailand, Italy, Italien, Italia, Europe, Europa”

This would give the corresponding record as a hit, if for example the string “Italien” would be used to specify the location in a query. In this case hierarchy extension makes sense, since the geographic concepts are exactly defined.

Since only one index is used both for simple and complex search, special care had to be taken how the extension can be done for prose text. For keyword type of metadata elements it was assumed that the vocabulary is used properly, i.e. we expect to find the complete string for an institution such as “Sterling and Francine Clark Art Institute” (an institution in Williamstown/ Massachusetts/USA). This allows us to match the complete string and therefore reduce the chance of fault hits. However, in prose text we may find various variants of such a string such as the “the Art Institute from Sterling and Francine Clark”, nevertheless the search engine should find the entry. We could only implement policies that do not rely on advanced Natural Language Processing. Therefore, during the extension it was allowed to break the found string down and to match for example “Sterling”. Such a policy would increase the risk of false hits, but in case of more information in the query such as “Francine Clark” those records that come from the mentioned institution would get a high rating and appear at the top.

The result of these processes is a large index file that includes all necessary types of information for each node in the tree such as Document ID, Repository ID and Xpath Information. So when a hit was found it can for example immediately be extracted where it comes from.

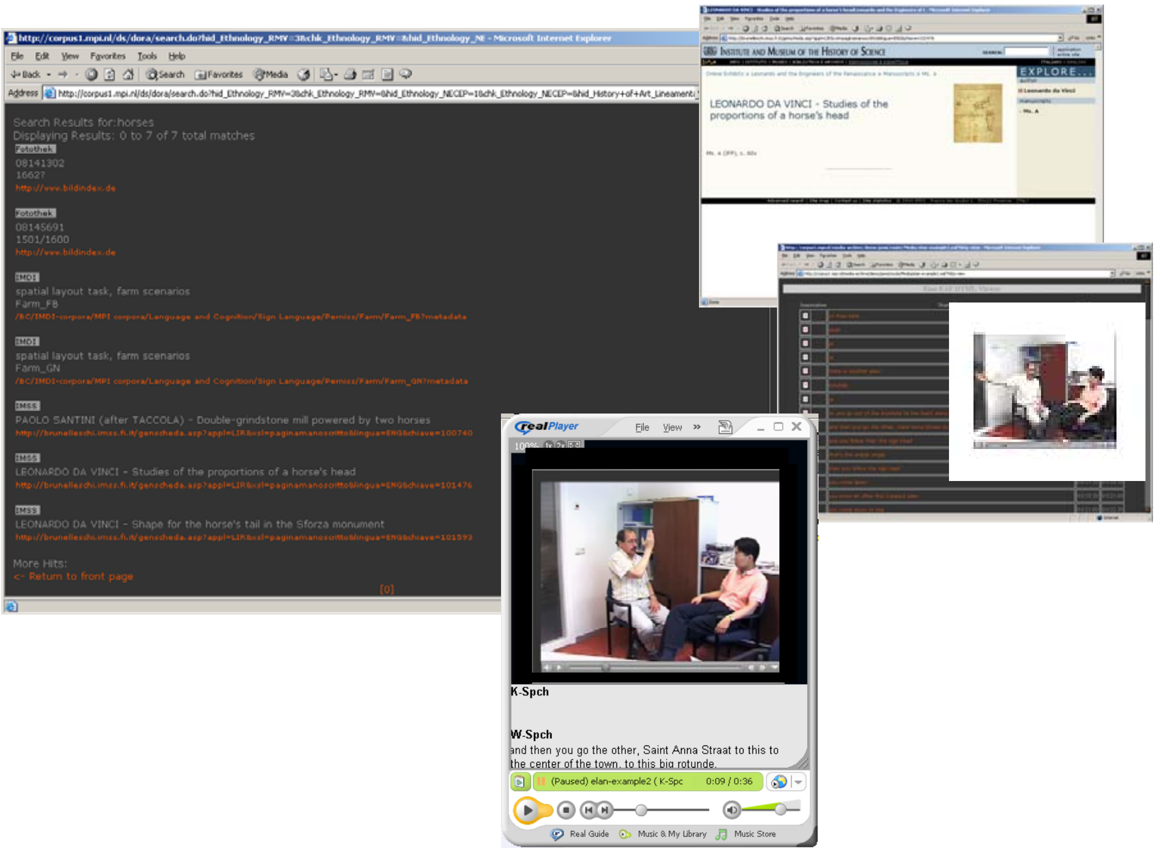
1.5 Searching

Searching is simply done by traversing the binary tree for every entry found in the query. This results in a number of hits which are filtered according to the selections made in the interface. When looking for the string “horse” also the “hits” for “horses” are used which is a morphological variant. Yet no lexical processing is used in the search algorithm.

The filtering includes that for domains, for sub-domains and for the field names for complex search. The latter includes all semantic mapping relations between the metadata categories as explained in the DORA note. In doing so the task of semantic mapping is reduced to a filtering step making mapping very fast.

[7] This could only be done in a limited and unsystematic way to help using the DORA engine.

A simple ranking mechanism is applied in the search algorithm. When two or more separate items as for example in “Sterling and Francine Clark Art Institute” (5 different items) all result in hits, then the hit receives a very high ranking. Further, the number of occurrences of a certain string in a metadata record is used to increase the ranking. Therefore we can speak about three ranking levels: (1) Highest ranking for the co-occurrence of multiple words appearing in the query. (2) Moderate ranking when a word occurs several times in a record. (3) Singular occurrence of one word of the query string.



http://corpus1.mpi.nl/ds/dora/search.do?hid_Ethnology_RMV=3&chk_Ethnology_RMV=&hid_Ethnology_NE - Microsoft Internet Explorer

Address http://corpus1.mpi.nl/ds/dora/search.do?hid_Ethnology_RMV=3&chk_Ethnology_RMV=&hid_Ethnology_NECEP=1&chk_Ethnology_NECEP=&hid_History+of+Art_Lineamenti

Search Results for:horses
Displaying Results: 0 to 7 of 7 total matches

Fotothek
08141302
1662?
http://www.bildindex.de

Fotothek
08145691
1501/1600
http://www.bildindex.de

IMDI
spatial layout task, farm scenarios
Farm_FB
/BC/IMDI-corpora/MPI corpora/Language and Cognition/Sign Language/Pemiss/Farm/Farm_FB?metadata

IMDI
spatial layout task, farm scenarios
Farm_GN
/BC/IMDI-corpora/MPI corpora/Language and Cognition/Sign Language/Pemiss/Farm/Farm_GN?metadata

IMSS
PAOLO SANTINI (after TACCOLA) - Double-grindstone mill powered by two horses
http://brunelleschi.imss.fi.it/genscheda.asp?appl=LIR&xsl=paginamanoscritto&lingua=ENG&chiave=100740

IMSS
LEONARDO DA VINCI - Studies of the proportions of a horse's head
http://brunelleschi.imss.fi.it/genscheda.asp?appl=LIR&xsl=paginamanoscritto&lingua=ENG&chiave=101476

IMSS
LEONARDO DA VINCI - Shape for the horse's tail in the Sforza monument
http://brunelleschi.imss.fi.it/genscheda.asp?appl=LIR&xsl=paginamanoscritto&lingua=ENG&chiave=101593

More Hits:
<- Return to front page
[0]

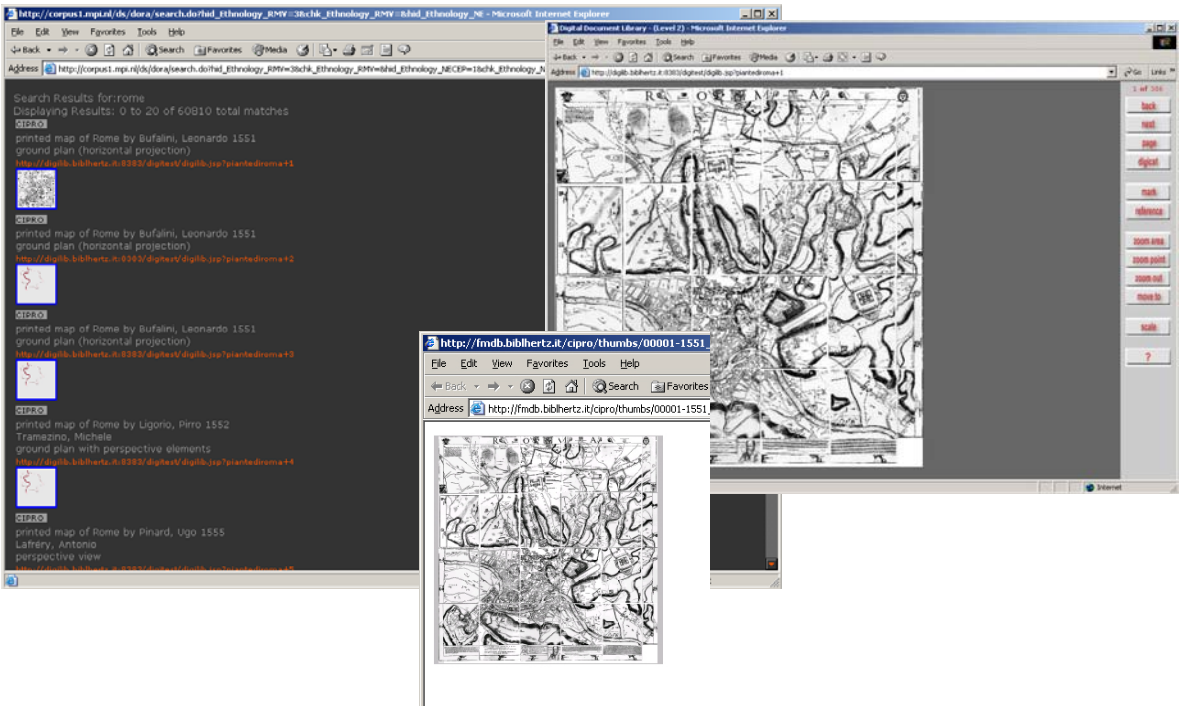
INSTITUTE AND MUSEUM OF THE HISTORY OF SCIENCE

LEONARDO DA VINCI - Studies of the proportions of a horse's head

K-Spch

W-Spch
and then you go the other, Saint Anna Straat to this to the center of the town, to this big rotunde.

(Paused) elan-example2 (K-Spc 0:09 / 0:36



http://corpus1.mpi.nl/ds/dora/search.do?hid_Ethnology_RMV=3&chk_Ethnology_RMV=&hid_Ethnology_NE - Microsoft Internet Explorer

Address http://corpus1.mpi.nl/ds/dora/search.do?hid_Ethnology_RMV=3&chk_Ethnology_RMV=&hid_Ethnology_NECEP=1&chk_Ethnology_N

Search Results for:rome
Displaying Results: 0 to 20 of 60810 total matches

CIPRO
printed map of Rome by Bufalini, Leonardo 1551
ground plan (horizontal projection)
http://digilib.biblhertz.it:8383/digitest/digilib.jsp?piantediroma+1

CIPRO
printed map of Rome by Bufalini, Leonardo 1551
ground plan (horizontal projection)
http://digilib.biblhertz.it:8383/digitest/digilib.jsp?piantediroma+2

CIPRO
printed map of Rome by Bufalini, Leonardo 1551
ground plan (horizontal projection)
http://digilib.biblhertz.it:8383/digitest/digilib.jsp?piantediroma+3

CIPRO
printed map of Rome by Ligorio, Pirro 1552
Tramezino, Michele
ground plan with perspective elements
http://digilib.biblhertz.it:8383/digitest/digilib.jsp?piantediroma+4

CIPRO
printed map of Rome by Pinard, Ugo 1555
Lafréry, Antonio
perspective view

Digital Document Library - (Level 2) - Microsoft Internet Explorer

http://fmdb.biblhertz.it/cipro/thumbs/00001-1551

Address http://fmdb.biblhertz.it/cipro/thumbs/00001-1551

With respect to the hits all information that is provided by the data providers is used to give as quick feedback as possible. In the above figures a few examples are given. The first example is the result of entering “horses” in simple search. It results in 8 hits from three different domains. In the case of the IMSS hits a back link is provided to the web-page with the following object: “PAOLO SANTINI (after TACCOLA) - Double-grindstone mill powered by two horses”., i.e., when clicking on the back link the shown page appears.

In the case of languages when querying for example “wittenburg”, a resource is shown with gesture data. When clicking on the back link one first gets the metadata entry, but can then request the annotations with the appropriate video fragment. Two options are available: (1) The annotations created with ELAN can be viewed with the help of HTML where clicking on an annotation will active the appropriate video fragment. (2) ELAN allows to generate a SMIL[8] object which is addressable via the metadata. When clicking streaming video is shown with subtitles. ELAN allows to select the tiers to be seen and the time fragment that is of interest.

In the third example the word “rome” is entered as query, delivering many hits for example from the CIPRO repository. Here two options are given. When clicking on the thumbnail a larger image of the map is shown. When clicking on the back link a page is offered with showing the appropriate map within the DIGILIB image processing framework. The presentation of the hits and the back link possibilities can certainly be improved, but they were not in the center of the ECHO work. Also some repositories include many resources that are not open.

2. Evaluation

This evaluation is split in three parts. In the first we will make some comments about the formal correctness which we distinguish from the usefulness of the chosen semantic mappings and operations which we will discuss with the help of examples. While in the case of the formal correctness one can speak about “errors”, the semantic mappings are a matter of subjective evaluation. The third part will make statements about the ranking.

2.1 Formal

The formal correctness include all aspects such as

- Are all specifications made in the ontology correctly implemented?
- Are the final metadata files (created by conversion) correct?
- Are the extension mechanisms that create the final index file correct?
- Are the extensions such that we don’t get a semantic explosion?

The latter has also to do with semantic evaluation, so it could also appear under 2.2.

During the last weeks much testing was done to see whether the engine and the underlying mapping files are correct. We distinguish two types of mappings: (1) Those mappings that are specified between the different metadata elements. (2) Those mappings that are established between the thesauri.

The mapping scheme between the metadata elements was provided and discussed very early with the data providing teams. The first version of the DORA document was distributed in late 2003, so that all teams could respond. The corrections we received were integrated. It was checked in detail during the tests whether the mappings are effective while searching. Here the method was to investigate specific examples that were obvious from studying the metadata sets. As far as can be seen from these investigations the specified mappings are used correctly.

The check of the correctness of the implementation of the thesaurus mappings and extensions was especially tested for the geographical elements. Here we discovered a number of errors which mainly had to do with incorrect character encodings in the metadata files. Although UTF-8 was mentioned in the header we found out that this specification was not correct in some cases.

[8] SMIL is a W3C supported standard for media presentations and will be supported by an increasing number of browsers.

Also in some cases additional characters were introduced in the strings. Only by these operational checks we could find out these errors. For the obvious cases corrections were carried out, although we cannot claim that these kinds of problems are completely removed.

Another problem we encountered was that the thesaurus extension leads to an explosion of hits in the case of the content description. In the case of geographical terms we have a well-defined domain that is organized hierarchically. In the case of content descriptions we don't have such a well-structured domain. Both – the application of semantic mappings between nodes of the content thesauri and the hierarchical extension – leads to cycles and an explosion amounting in too many non useful hits. Therefore, we concluded that for the content description within ECHO we will only exploit the mapping specifications and not use the hierarchy information. A more detailed semantic analysis would have to be carried out to come to refinements. This was beyond the scope of the ECHO project.

2.2 Examples and Semantics

First, we will give a number of examples and then give a first evaluation.

Example 1

Simple Search "weapons"
    87 matches are found: Fotothek: 84, RMV: 1, IMSS: 2
Complex Search "weapons"
    Fotothek - Iconography: 84, RMV - Content Description: 1 , IMSS - title: 2

Both search types lead to the same result. In the case of complex search the mapping between the fields becomes effective leading to acceptable results.

Example 2

Simple Search "dogon"
    1 match was found: NECEP: 1
Complex Search "dogon"
    View NECEP - society name: 1 in NECEP
    View IMSS - language: 1 in NECEP
    View DC - language: 1 in NECEP
    View Language - language: 1 in NECEP
Complex Search "mali"
    View Language - country: 1 in NECEP

This example demonstrates the effect of mapping and geographical thesaurus. The language element is mapped to the society name element in NECEP although this is semantically not fully correct. Entering "mali" in the country specification yields a hit since "mali" is seen as a superclass to "dogon". Here a relation type such as "has_language" would be semantically more appropriate.

Example 3

Simple Search "inuit"
    2 matches are found: Language: 1, NECEP: 1
Complex Search "inuit"
    View Language - *: 0 in Language (could not be found in the Language domain)
    View Language – language: 1 in NECEP
Complex Search "greenland"
    View Language – language: 1 in NECEP

The results are similar compared to example 2. It indicates that the element including "inuit" in the language domain is not an element that is used for mapping. It was used as an optional field by one specific researcher.

Example 4

Simple Search “agriculture”

75 matches are found: Language: 73, Fotothek: 2

Complex Search “agriculture”

View Fotothek - iconography: 2 in Fotothek
View RMV – content: 2 in Fotothek
View IMDI – content: 2 in Fotothek

The results can be misleading. The 73 hits for language result from matching with recording place (“southern agriculture kindergarten”) or affiliation of an actor (“ministry of agriculture”). In the case of Fotothek the hits make sense since it is about “harvesting”. The mapping in complex search works properly as indicated. Of course, in complex search the misleading hits from the language domain are not found.

Example 5

Simple Search “clothing”

22 matches: Language: 8, RMV: 8, Fotothek: 6

Complex Search “clothing”

View RMV – content: 8 in RMV, 6 in Fotothek
View Fotothek – iconography: 8 in RMV, 6 in Fotothek
View Language – content: 8 in RMV, 6 in Fotothek

Again the rich annotations that are inserted in various free-text fields in the language domain lead to not useful hits. They are about chats at the bakery shop and the clothes people are wearing – so it’s not about clothing as an object which may be intended by the person specifying the search. The results for complex search from different domains shows the correctness of the mappings. The language hits are excluded, but the others are found.

Example 6

Simple Search “horses”

7 matches: Fotothek: 2, Language: 2, IMSS: 3

Complex Search “horses”

View Fotothek – object title: 3 in IMSS
View Fotothek – iconography: 2 in Fotothek
View Lineamenta – title: 3 in IMSS
View Lineamenta – keywords: 2 in Fotothek
View IMSS – title: 3 in IMSS
View IMSS –subject: 2 in Fotothek
View Language – title: 3 in IMSS
View Language – content: 2 in Fotothek

This example clearly indicates the strength of simple search and the weakness of complex search. The pattern of complex search is like a narrow path in the complex semantic space. If one looks at title one finds the IMSS hits, if one looks at content one finds the Fotothek hits. Both, however, are leading to useful hits where “horses” have an important role. The reason partly is that metadata in many cases is very sparsely encoded. In the case of IMSS the term horses is only mentioned in the title, but the content element is yet not used. In the language case thesaurus information is used to infer from the title content “spatial layout task, farm scenarios” to “horses”.

Further tests and examples will follow.

Yet, there is no clear statement whether simple or complex search are better. Simple search is good when one wants to be sure to get a large number of hits where the probability is very high that the documents looking for are included – even at the price of a large number of hits. Complex search is more selective and its matching operations are much more strict. In general complex search is excellent for those metadata elements that describe a more precise domain such as date, geographic location and authors. Content descriptions are done in very different ways and according to different categorization principles (thesauri, keywords). Any professional search on these elements requires a high degree of knowledge about the underlying category system and its semantics. If one wants to exploit the advantages a thesaurus such as IconClass can offer, one has to know its semantic construction principles.

One big advantage of simple search is that it uses all fields even if they contain prose text. However, it also increases the number of appropriate hits as was shown in the examples.

2.3 Ranking

Ranking is a possibility to satisfy the user in case of low precision. It is a general rule to offer more hits even if non-appropriate documents are included, since there is always a penalty between “recall” and “precision”. If the “recall” (ratio of appropriate documents found to total number of appropriate documents) shall be increased normally the precision (ratio of appropriate documents to in-appropriate) decreases. But the primary goal is to find all appropriate documents and offer them. A compromise then is to offer all appropriate documents first in case of clear evidence.

The implemented ranking is based on frequency of occurrence and not on semantic criteria. It makes sense to weight multiple occurrence of different terms higher than multiple occurrence of one term. The fact that more terms found in the query input are matching raises the probability that the found document is a useful hit. The results found are in general satisfying.

An implementation of a ranking based on semantic criteria requires much more experience and insight to the usage of all concepts. Since many metadata sets were offered at a very late moment within the project there was no chance to include semantics in rating. Including semantics also means to include a bias. It is obvious that people disagree on semantic relations and want to be able to tune the semantically related operations according to their wishes. Therefore, we refrained from making use of the “mapping quality” parameter which can be added to the mapping relations between the different metadata elements. It would require much more time to come with useful defaults.

At this stage of the DORA search engine ranking based on formal criteria is much more appropriate than including semantic criteria.

3. Conclusions

The final conclusions will be drawn when all evaluations have been done in June. Here some preliminary conclusions are made.

Creating an interoperable and interdisciplinary search space is a difficult task. So DORA is one of the first attempts to do this in a flexible and unbiased way without a specific goal in mind. It is not yet clear whether this approach is useful. A project approach – even if it includes a few disciplines – may have specific objectives in mind that will require a careful analysis of the included semantics and it may include strong biases.

DORA was intended to make it easy to integrate other domains into the search space. Integrating another discipline requires activities at the harvesting and data preprocessing level which will not be commented here. It was already described that most of the repositories are yet not so far to offer validated, correct and stable output. The OAI MHP protocol is important, but many repositories are not ready. Even the concept of metadata was new for some and a fair debate showed that some question the usefulness of keyword type of metadata. Here we can see a difference between institutions that hold large collections of multimedia objects and those that are more text oriented.

Discipline integration also requires various operations to integrate the semantics:

- The mappings to other metadata elements have to be added to support complex search.
- In the case of geographic descriptions one has to create a discipline specific list of terms and relate them to nodes in a geographic thesaurus.

- In the case of content descriptions one also has to create relations to concepts used in other domains.

Currently, the effort is very high, since there is no structural support and there are no existing knowledge documents one can refer to in the area of the humanities. What is needed to support such work and also allow individuals or groups to tune the semantics to their needs is as follows:

- Open Data Category Registries that contain ISO compliant concept definitions occurring in a discipline. Compliance to standards such as ISO 11179 would guarantee a certain degree of homogeneity and increase the re-usability. The definitions should be included in XML files that are associated with a schema. These definitions should contain only those relations that are part of the proper definitions of a concept, i.e., if for example the sub-class relation is important to define a concept than a relation to another concept could be included. However, it is wise to reduce this to a minimum, since relations often are a matter of disagreement even within domain.
- This also is valid for the thesauri. As far as is known to us, the big thesauri have their own definition style, come with a particular access interface and are not open available as an XML file[9].
- For the mappings we also need frameworks to easily create practical ontologies. These should be described in RDF and refer to concepts defined in open registries. It must be possible for users to easily create their own versions, i.e., to adapt existing relations or to add new ones.
- All these components must be machine-readable and inference engines must be available that can operate on them.
- Registration mechanisms have to be designed that allow to register knowledge components and to search for them.
- The RDF-S and OWL definitions are an excellent start to formalize relation types, however, in practical work we are often faced with fuzzy or unclear relations that cannot be described by RDF-S/OWL types.

Part of the work has been started in the area of Language Resources (ISO TC37/SC4). This can be seen as an example to start such work in other disciplines of the humanities. It will pave the way of the humanities towards the Semantic Web. DORA is an attempt to tackle some of the problems based on open and well-structured ontology components, yet, most of them are not based on established standards.

A key point for success of DORA like approaches with complex search based on selected metadata categories will be the flexibility for users and groups to tune the semantics. The above mentioned steps will help doing this, but smart and user-friendly tools have to be available.

From the experience it is obvious that the choice to not offer Dublin Core as the Gold Semantic Standard was appropriate. The success of selective search will depend on the knowledge about the vocabularies and the quality of the mappings. Dublin Core presents a rather reduced vocabulary with loosely defined concepts. It is not obvious how different disciplines will map their concepts on the Dublin Core ones and in general this mapping is not open. So the concept of a GOLD standard may be useful for cases like the domain of book descriptions where the concepts such as title, author, year of appearance and publisher developed for many years and are used by all libraries. For purposes such as DORA which want to go beyond these formalized elements, Dublin Core cannot be recommended. It may play a role for occasional users, but it can be questioned whether DC search is preferable compared to simple search.

An important aspect that restricts the quality of this evaluation is the lack of detailed metadata descriptions in many cases and the comparatively small number of objects in some of the repositories. Only the Fotothek and Language repositories have a large number of records. For repositories that offer about 100 records or less browsing is sufficient and then superior to

[9] To make IconClass useful in the DORA framework the database format used on the distributed CDROM had to be decoded with the help of scripts and some manual intervention to come to an appropriate XML structured file.

searching. However, it is obvious that this will change in all disciplines since the number of digital objects stored increases extremely fast.

The DORA technology has to be seen as one of the possible initiatives to indicate how difficult semantic integration is and how much has to be done in future. We need more of such attempts to build the infrastructures and tools to cope with the challenges of the Semantic Web and to prepare the disciplines of the humanities for these challenges.