Metadata Interoperability and Meta-search on the Web

Enric Peig, Jaime Delgado, Ismael Pérez

Universitat Pompeu Fabra (UPF), Departament de Tecnologia
Pg. Circumval·lació 8, E-08003 Barcelona, Spain
{enric.peig, jaime.delgado, ismael.perez}@tecn.upf.es

Abstract

Several initiatives for establishing standards for metadata models are being carried out at the moment, but everyone focuses on their own requirements when defining metadata attributes, their possible values and the relation between them. From the point of view of someone who wants to seek and buy information (multimedia content in general) in different environments, this is a real problem, because he has to face different metadata sets, and so, must have different tools in order to deal with them.

In this paper, we present a model for the interoperability of different metadata communities, where neither the providers nor the users have to be aware that they all may be working with different metadata models. We are mapping the semantics of different metadata models with the objective of not loosing information when the user and the content provider use different metadata schemas. A "metadata agent" is used to carry out the interoperability functionality.

On the other hand, the use of the Internet as a tool for searching information and multimedia content is continuously growing, but the use of metadata in the Word Wide Web is very poor. As a result, many search engines have appeared, that help users to find information. These search engines are able to find information, but generally this information does not follow any metadata standard. Our objective here is to create a meta-search agent able to extract information from the Internet starting from server-independent queries, which are mapped to search engine specific queries. The results are then re-processed to provide users with the requested information, again in a server-independent way.

Keywords: *Metadata, interoperability, meta-search*

1 The need for interoperability

The usage of metadata schemes for referencing multimedia material is becoming more and more usual. But in the last years, many different schemes have been proposed. Some of them have very specific focus and their usage is circumscribed to particular environments, but other ones are of general purpose, and in some environments information providers that use different metadata schemes can be found together.

This situation forces applications to know all the schemes that may be found. Furthermore, it is also usual to find storage systems containing objects referred to following different metadata schemes at the same time. There is still another extra problem: we have to be aware of new metadata schemes that might appear. So, applications must be adapted to these new schemes.

Because all these reasons, there is a need to develop interoperability systems between metadata domains, with the purpose of simplifying the discovery and the access to the information, and to achieve a high level of automation in this access.

For our work, we are initially considering three metadata initiatives: Dublin Core [1], MPEG-7 [2] and IEEE LOM [3]. They are widely used and have different focus. Since we want to deal with metadata interoperability, these seem to be good reasons to select them.

In the next sub-sections we give a short overview of these three initiatives.

1.1 Dublin Core

Dublin Core is a standard that represents a metadata element set intended to facilitate the discovery of electronic resources. Although it was born in the bibliographic domain, it has turned out to be a de facto standard for metadata on the Web.

The metadata element set is formed by these 15 elements: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights. The simplicity and conciseness of the set is one of the keys that explain its success.

Besides the metadata element set, a list of qualifiers is formally recommended, intended to sharpen the semantics of the 15 original elements,

https://doi.org/10.23106/dcmi.952106664

and thus, to adjust to specific domains and local implementations.

1.2 MPEG-7

The Moving Picture Experts Group (MPEG) is a working group of ISO/IEC in charge of the development of standards for coded representation. Among many others, it is now working on the MPEG-7 standard, formally named "Multimedia Content Description Interface", whose aim is to create a standard for describing multimedia data, and to offer tools to create and manage their descriptors. Its natural scope is the description of audiovisual information, be it analogue or digital, and be it broadcasted in real time from some source or recorded in media such as film, magnetic tape, CD, etc.

The MPEG-7 tools will allow to create descriptors of content that may include information describing the creation and production processes of the content, information related to the usage of the content, information of the storage features of the content, structural information on spatial or temporal components of the content, conceptual information of the reality captured by the content, etc.

A description generated using MPEG-7 description tools will be associated with the content itself, to allow fast and efficient searching for, and filtering of material that is of interest to the user. MPEG-7 data may be physically located with the associated audiovisual (AV) material, in the same data stream, or in the same storage system, but the descriptions could also live somewhere else on the globe. When the content and its descriptions are not co-located, mechanisms that link AV material and their MPEG-7 descriptions are needed; these links will have to work in both directions.

The main tools used to implement MPEG-7 descriptions are the Description Definition Language (DDL), Description Schemes (DSs), and Descriptors (Ds). Descriptors bind a feature to a set of values. Description Schemes are models of the multimedia objects and of the universes that they represent; e.g. the data model of the description. They specify the types of the descriptors that can be used in a given description, and the relationships between these descriptors or between other Description Schemes.

1.3 IEEE LOM

The IEEE, through its Learning Technology Standards Committee, is working in a standard that aims to facilitate search, evaluation, acquisition, and use of learning objects, for instance by learners or instructors. Currently, this standard, called Learning Objects Metadata (LOM), is in the status of working draft.

The standard specifies a conceptual data scheme, formed by data elements that describe a learning object. Also, a Base Scheme is specified, which for each data element defines a name, an explanation, the size, the order, the value space, the data type and an illustrative example.

The data elements can be grouped into categories. The Base Scheme consists of nine categories: General, Lifecycle, Meta-metadata, Technical, Educational, Rights, Relation, Annotation and Classification.

2 A model for metadata interoperability

2.1 A first approach

Several approaches to interoperability have been carried out during the last years, but none of them has still got a relevant result. An example of this, mainly at European level is the work done by the CEN/ISSS (European Standardisation Committee / Information Society Standardisation System), where a Workshop was created to deal with these issues, mainly focussing on metadata for multimedia information.

The results from the Workshop, that was named MMI (Metadata for Multimedia Information), were a few CWA (CEN Workshop Agreement) specifying a model for metadata and business requirements [4].

The MMI Model proposes a conceptual model for metadata for multimedia information in terms of classes of metadata, the roles of the different actors involved and the actions performed by each role. At the conceptual level, the same concepts and life cycle model can be applied both to information resources and to metadata.

The MMI requirements do not attempt to produce a complete set of requirements for all uses of metadata, since this would be an endless task. On the contrary, the document is providing a metadata taxonomy and methodology to help identifying requirements for different sectors and applications.

The document can be used for different purposes in different ways. First, to have an overview about what are the general requirements for metadata; second, to check if some specific metadata requirements fit with the taxonomy; and, third (the most important use), to derive, from the described taxonomy, new specific requirements for new applications.

Finally, it should be noted that the MMI requirements was intended to be a "living" document that may be updated *1)* when new metadata requirements are developed for new applications; *2)* new applications discover their need for metadata; and *3)* if some new requirements are identified for the taxonomy.

As a conclusion, we could say that this work has been a good starting point and has identified the complexity of the problem, also showing that trying

https://doi.org/10.23106/dcmi.952106664

to cover all existing metadata schemes or trying to converge is not a feasible task. We should add that this Workshop has been disbanded once the CWAs were published, and has moved to a new Workshop, still running, focussing on Dublin Core.

2.2 Our proposal

The model we are proposing, with a totally different approach to the CEN/ISSS one, is oriented to the search and discovery of metadata referenced material. Hence, among all the aspects that are included in the definition of a metadata scheme, we are only interested in the element set and their meaning. The creation of the metadata or the relationship between attributes (if it exists) are of less concern to us.

Our model is based on two key aspects: Firstly, on a common vocabulary that gathers the metadata elements from the different schemes with similar meaning, and secondly on not imposing the knowledge of this vocabulary to any actor of our system. Instead of this, we propose to use a mediator, a kind of agent that will be in charge of the searches in the different information providers, at the request of the users of the system.

2.3 The common vocabulary

The common vocabulary has its origin in the analysis of the metadata schemes already mentioned. A first study reveals that there are semantic coincidences between some attributes of the different schemes. Then, these attributes will form the kernel of our common vocabulary, with the semantic mapping to the elements of the real metadata schemes.

From here, the system has to be able to incorporate new elements and the corresponding mapping in case of finding other metadata schemes with elements not yet considered in the common vocabulary. To make this inclusion it is clear that, on a first step human intervention is needed, since tools to deduce the semantics of these elements are not currently available. Using ontologies for this purpose is being considered.

Table 1 illustrates the kernel of the common vocabulary and its mapping to the mentioned metadata schemes.

This mechanism of semantic mapping from a general vocabulary to the different metadata schemes is easily scalable, since we do not need to maintain crossed mapping among all existing schemes. It is clear, as also stated in [5], that the idea of supporting a matrix for crossed mappings between all possible schemes is not a scalable one.

What we propose then in our interoperability model is to only consider the mapping between different schemes and our common vocabulary. Then, for every new scheme that we want to add to our system, we only need to fill a column in the previous table, where the attributes with a semantic relationship with our vocabulary would appear.

Taking into account that the objective of the model we are presenting is the search of content in heterogeneous sources, our approach is that it is not necessary to keep an exhaustive and complete mapping of all the attributes of the metadata schemes. Hence, we can forget about those attributes that only appear in only one scheme but not on the others. In this way, our vocabulary would be a kind of intersection of all available systems we could find.

Table 1. Common vocabulary and its mapping.

	Dublin Core	IEEE LOM	MPEG7
Identifier	Identifier	General.CatalogEntry	MediaInformation.MediaIdentification.Identifier
Title	Title	General.Title	CreationMetaInformation.Creation.Title.TitleText
Description	Description	General.Description	CreationMetaInformation.Creation.CreationDescription
Format	Format	Technical.Format	MediaInformation.MediaProfile.MediaFormat.FileFormat
Author	Creator	LifeCycle.Contribute.Entity	CreationMetaInformation.Creation.Creator
Creation_Date	Date	LifeCycle.Contribute.Date	CreationMetaInformation.Creation.CreationDate
Language	Language	General.Language	CreationMetaInformation.Classification.Language.LanguageCode
Rights	Rights	Rights	UsageMetaInformation.Rights.RightsID

https://doi.org/10.23106/dcmi.952106664

2.4 The metadata agent

As it has been said, our model is also based on not imposing our proposed metadata scheme (our common vocabulary) to information providers, but to use an agent that will be in charge of searching in the different information providers, at the request of the users of the system.

This *metadata agent* is the only element that knows about the common vocabulary and the mappings. In this way, the content providers and users searching for information are able to continue working with their own metadata schemes with the help of the agent.

Figure 1 shows how the different elements of our scenario are related.

We can see that the different elements (or actors) interchange two kinds of information. On one side, the user provides some keywords to the agent, so it can make the search in the provider system. As an example, we assume that users make queries such as "search for films from Director X" or "search for a painting from Artist Y between year Z1 and year Z2". Then, the agent has to map this information to the metadata schemes corresponding to the content providers where it will look for, in order to be able to deal with them, since we assume that they only understand queries following their metadata scheme.

On the other hand, we have the answers given by the content providers, which, in many cases, have the form of a metadata record following their own scheme. The task of the agent is then to provide the user with this information, that could follow their original scheme, the common vocabulary or the scheme requested by the user, if different.

With this approach, users are able to make queries to different content providers without the need of knowing their metadata scheme, both in the moment of producing the query and when receiving the answer with the requested information.

3 Application to meta-search in the Web

The most popular and the biggest information provider is the World Wide Web. There are ways of seeing the whole WWW as a unique, huge information provider. The most common approach is by using search-engines, which allow searching, normally by keyword-based queries, HTML pages in different Web servers.

Our approach to metadata interoperability considers that every information provider has its own metadata schema. We could consider WWW search engines as information providers with its own metadata schema each. Then, we could develop one of our interoperability agents as a meta-search engine, allowing users to query our agent to find any information on the Web.

However, the role of the agent is now at a different level. While the metadata agent translates metadata attributes between different schemes, the meta-search agent translates queries for different search engines.

Since trying to find any information on the Web is a rather ambitious objective, we are initially restricting ourselves to specific subject areas. The next subsections describe our meta-search agent.

3.1 Existing search engines

The enormous expansion of multimedia content and information through Internet has forced content providers to develop search engines to automatically find information. These search engines collect web pages and create databases for users to browse or search (see [7] for more information).

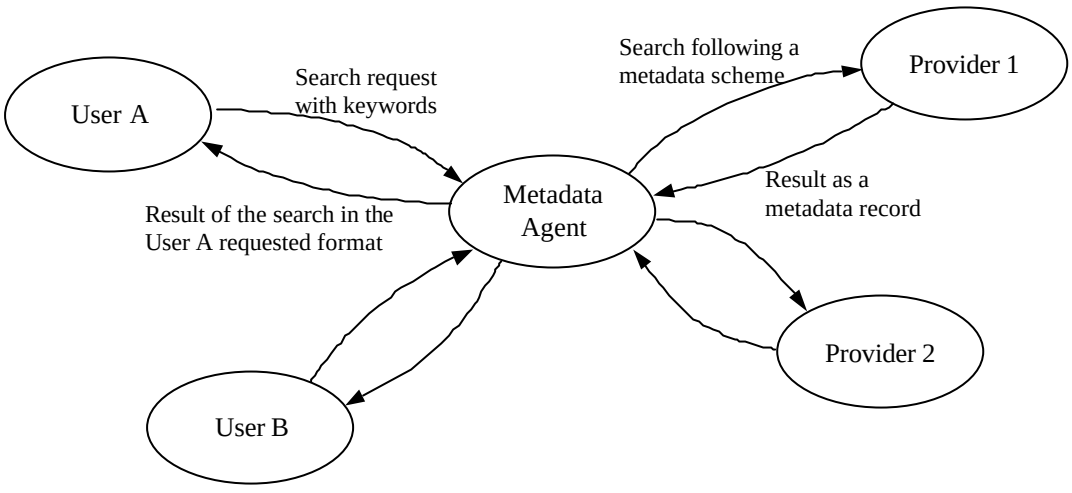

Figure 1. Relationship between elements.

https://doi.org/10.23106/dcmi.952106664

There are many different search engines for similar content and, in many cases, they are specialized in specific content subjects. Content attributes are usually defined, but generally this metadata attributes have no value. As a result of this, users must handle several engines in order to find the desired information, and, often, users must combine the information of different search engines to have full metadata information. Figure 3 illustrates this situation. See [8] for more information.

Moreover, each search engine:

- handles various content types,
- presents a specific user interface,
- requires its own set of rules for searching,
- has an ad-hoc database,
- has a set of indexed web pages,
- has not all content metadata attributes,
- presents search results to the user differently.

In addition, many meta-search engines already exist. A meta-search engine is a tool able to extract information of many search engines and present search results to the user as a simple search engine does.

Successful use of meta-search engines depends on the functionality of each individual search engine used. Some may search by a unique attribute and some require several attributes. For this reason, meta-search engines usually have only one word for users to search. Taking advantage of metadata-based retrieval techniques to allow the user to make searches based on several attributes may be a good solution to solve this problem. The retrieved metadata must be given back to the user with the resulting page. This is the approach we are taken in our work, as explained later.

3.2 The meta-search agent

Our main goal is to find a better way to obtain and manage this knowledge (information on the Web) in order to provide the user with a friendly and easy to use information access. We want to generalize the access way to the information.

Currently, users have to look for information in many search engines, and have to combine this information to obtain better information. The basic idea of our approach is to create a tool that replaces this hard work. This tool must have a uniform interface, where a query can be easily and quickly submitted, and where the search can be conducted to various search engines. Then, search results metadata information can be combined to return as much complete metadata information as possible and used to improve the user query and to discard some results not in accordance with the user demand. Therefore, the user can make a more powerful query. Figure 2 presents the basic structure of our search agent. It is also possible for our tool to change the user interface program in order to increase the flexibility of the application.

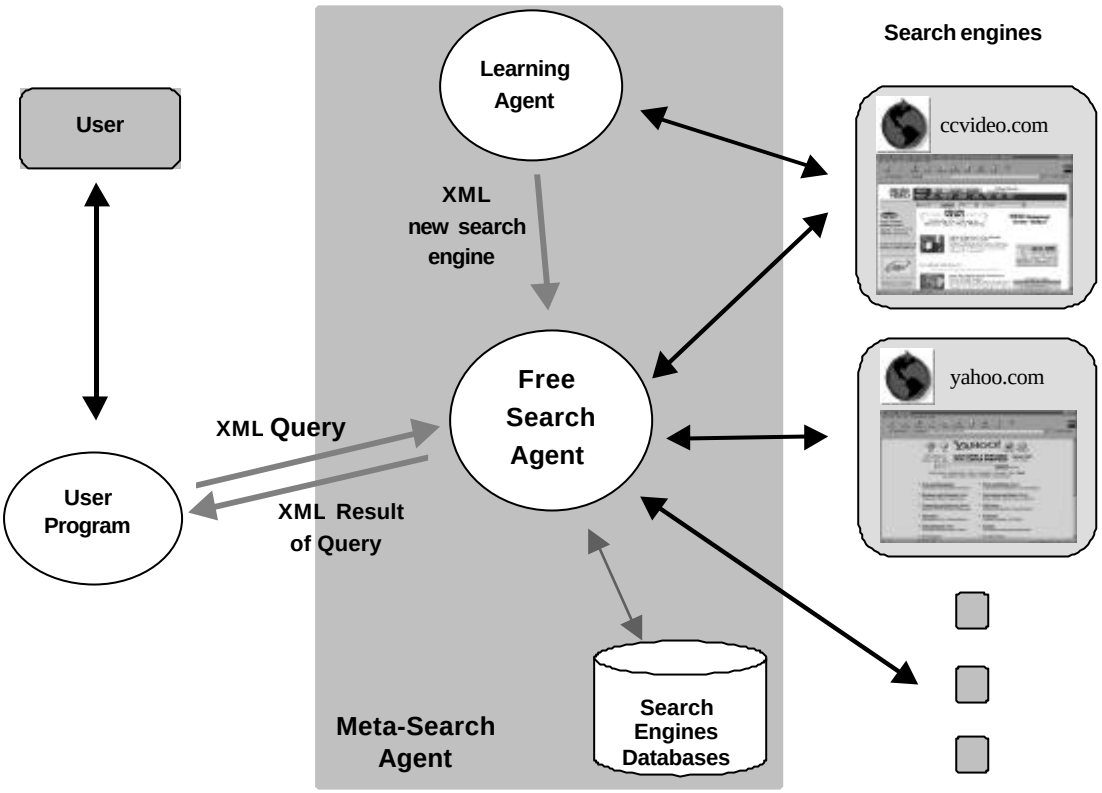

Figure 2. The meta-search agent.

https://doi.org/10.23106/dcmi.952106664

```
...
<FORM action="http://www.ccvideo.com/vid_s_new_search.cgi" method=get>
       <table border=0 cellpadding=0 cellspacing=1 width=80%>
       <td valign=bottom align=right>
              <img src="/imx/pixel.gif" alt="" width="1" height="5"><br>
              <input type=hidden name=search_type value=intersection>
              <input type=text NAME=key size=11>
              <input type=submit value=search><br>
       </td>
       <td align=center valign=bottom>
              <img src="/imx/pixel.gif" alt="" width="1" height="5"><br>
              <SELECT NAME=findby>
                   <option value=Title>Title
                   <option value=Actor>Actor
                   <option value=Director>Director
                   </SELECT><br>

       </td>
</FORM>
...
```

Figure 3. HTML source example.

In Figure 2 we can see the interaction of several elements. Our meta-search agent interacts with different search engines using standard HTTP, and with the user via an access program. In this second case we are using XML over CORBA for the communication. This meta-search tool is structured into two cooperating search agents and a database in order to store search engines information. The mentioned agents share information using XML; one agent: *the learning agent*, searches the web to find new search engines and to learn how to access them so as to incorporate them as new searchable engines; the other agent: *the free search agent*, uses this information to extract required information of many search engines.

An important feature is the automatic appending and classification of new search engines in order to query them. We want to develop a "learning agent" able to extract the communication method of each search engine so as to incorporate it as a new search engine.

There is information available in the Internet that may be used to learn search engine query methods for our agent. The information we are using to learn is obtained from the query information contained in the HTML page of the search engine program. By parsing the HTML source and retrieving information about HTML tables, forms and words, we can easily design a method to query this engine (see Figure 3 for an example), based on how the search engine query program works. Variables needed are sent through the open sockets communication (using HTTP protocol, of course) using the same attribute names and values of HTML fields. These search attribute names can be retrieved using an HTML parser and stored as a new search way to query this search engine.

In Figure 3, the HTML retrieved components (attribute names and associated values) are:

- search_type=intersection
- key=<word requested>
- findby={Title (Title), Actor (Actor), Director(Director)}

An example of query URL may be:
http://www.ccvideo.com/vid_s_new_search.cgi?search_type=intersection& key=matrix&findby=Title

One of the important features of our tool is the big amount of information that handles; we obtain this information from each search engine. Although this implies extra work, it has some advantages:

- To allow the user to make more specific consultation.
- The metadata information returned to the user is more complete.
- Our tool does not depend on the status of one search engine, but it depends on the status of the entire group.

Our meta-search engine aims to be a metadata access central place with a uniform interface, where a query can be entered by the user and the search can be conducted in as many search engines as necessary, and search results can be combined and returned to the user program in a consistent format, XML.

In our case we have chosen *a solution based on the use of an XML interface* between the program using the agent and the agent itself. It is therefore possible for the user to use any interface program with one premise: the communication between the user program and our tool must be XML [9]. As a result, the user program may be anything, a java applet in a web page, a specific program, etc.

Search and meta-search engines generally do not make powerful searches, but they only allow the user to look for a word or a list of words; using XML, users can make a complete query combining conditions with *and*s, *or*s, and parenthesis; in other words, a complex boolean query. Our meta-search engine allows the user to ask by metadata attributes and its values. Metadata attributes and values can be grouped using relationship conditions such as "greater than", "contains", and all the rest defined in the corresponding DTD we have developed, where we have tried to create a generic database query model.

https://doi.org/10.23106/dcmi.952106664

Search and meta-search engines generally do not make powerful searches. They only allow the user to look for a word or a list of words. By using XML, users can make a complete query combining conditions with *and*s, *or*s, and parenthesis; in other words, a complex boolean query. Our meta-search engine allows the user to ask by attributes and its values. Attributes and values can be grouped using relationship conditions such as "greater than", "contains", and all the rest defined in the corresponding DTD we have developed, where we have tried to create a generic database query and result model. These allow the user to query attributes that are not accepted for search engines using some metadata retrieval techniques.

The query following our DTD is a list of property names requested by the user program. It has also a list of constraints, that may be:

- A compound constraint: constraint *and* constraint, constraint *or* constraint, (constraint), *not* constraint.
- A simple constraint, that combines content specific attributes and values: attribute *equal* value, attribute *greater than* value, attribute *smaller than* value, attribute *greater or equal than* value, attribute *smaller or equal than* value and attribute *contains* value.

With this specification the user can make any kind of query. An example is shown in Figure 4.

Our agent extracts the HTML page of results from the associated search engine. The HTML result pages are different and currently may be difficult to extract metadata information from the query results.

Our objective is to extract as much information as possible of all resulting links with two purposes: to enhance the metadata information returned to the user and to allow the user to make a more powerful search query. With this meta-information retrieval operation, the user can request information of a multimedia element that no search engine can supply.

A search engine program constructs the resulting HTML page using a specific search engine interface that is different from other search engines. The interface of all the resulting links is different too. Because of this, the content attributes and values are structured in a different way. Some may be stored in horizontal or vertical tables, in a list of attributes and values, or using natural language. As a result, the metadata extraction may be difficult.

In some cases, there are some META tags in order to solve this problem, but the use of these elements is reduced to some tags in the HTML header which normally only gives information about the company or the title of the page [10]. Then, these meta-information tags do not help us to extract the meta-information needed to enhance the user query.

We are currently only interested in the textual meta-information retrieval, since non-textual meta-information such as images are heavy to process and extract, and the information that it contains is normally not very important for our purposes. Nevertheless, we leave this option open for the moment.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE QUERY SYSTEM "query.dtd">
<QUERY>
   <PROPERTYNAME>title</PROPERTYNAME>
   <CONSTRAINTS>
      <CONSTRAINTS>
         <CONTAINS propertyName="title" value="Matrix"/>
      </CONSTRAINTS>
      <AND/>
      <CONSTRAINTS>
         <PARENTHESISOPEN/>
         <CONSTRAINTS>
            <CONSTRAINTS>
               <NEGATIONOPEN/>
               <CONSTRAINTS>
                  <EQUALS propertyName="director" value="Steven Spielberg"/>
               </CONSTRAINTS>
               <NEGATIONCLOSE/>
            </CONSTRAINTS>
            <OR/>
            <CONSTRAINTS>
               <GREATER propertyName="duration" value="90"/>
            </CONSTRAINTS>
         </CONSTRAINTS>
         <PARENTHESISCLOSE/>
      </CONSTRAINTS>
   </CONSTRAINTS>
</QUERY>
```

Figure 4. XML query example implementation.

https://doi.org/10.23106/dcmi.952106664

In order to enhance the results of the query, we are trying with different techniques based on how the results are stored. Depending on it, we make new queries on this information to get the desired results.

Another mechanism, which we are also considering, is using ontologies, that is, a content-based access to the Web. An ontology provides the primitives needed to retrieve information about some categories of contents. These ontologies usually are keyword hierarchies according to metadata schemas specialized in a concrete type of content.

4 Conclusions

Using metadata to search for information on the Web is becoming more and more common. This is clearly good for users, since better search tools can be developed, but on the other hand is also creating disadvantages. A problem of interoperability is appearing, since different information providers use different metadata schemas that are normally incompatible.

We are working on a model to solve this problem, that is based on the idea of a central agent mapping metadata schemas between users and content providers. The initial ideas have been presented in the paper, but no implementation of a general metadata agent has been started yet.

However, a specific development of a meta-search engine is being made and has been presented. An agent that allows users to transparently access many search engines with only one interface is being implemented.

The implementation of the meta-search agent is being made according to the following premises:

- Adaptability: Our application must be easily moved to different systems.
- Interactivity: Our agent must easily intercommunicate with other programs and web pages. The use of XML technology grants this point.
- Functionality: With the use of metadata-based retrieval we enhance the search possibilities of current search-engines. Our agent is able to make more powerful queries, and able to extract more information from the individual results.
- Extensibility: Our tool must be able to incorporate new search engines, eliminate defunct ones and change itself in order to incorporate search engines new funcionality.

We have already implemented part of the meta-search agent in the MARS project [6], an application for the audiovisual sector in the context of the Internet 2 (broadband Internet) Pilot in Catalonia. The MARS application consists in a broker for multimedia content with some important and specific features: Use of a metadata database, XML and RDF interfaces, CORBA communication, watermarking of multimedia video content, etc. The application allows users to connect to the broker in order to make a query, and returns the multimedia content that is stored in the content distributors (shops, normally TV programs producers). The multimedia content is dynamically stored using all metadata attributes and lets the user make a heavy specific query. The application also allows content providers to incorporate new multimedia content to the broker's database and, in the future, to negotiate IPR conditions. When a user makes a query and the information requested is not found in the broker database, the broker starts a free search with the meta-search agent. The version implemented with this application only searches in various specific search engines and does not incorporate yet automatically indexed sites, but the first results have been very promising.

Apart from this implementation in the area of video content search, a second specific meta-search engine is being developed in the area of daily news [11]. The first results are showing that the approach we are taking is a valid one.

References

[1] Dublin Core Metadata Initiative, http://dublincore.org
[2] "Overview of the MPEG-7 Standard", ISO/IEC JTC1/SC29/WG11 N4031, http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm, March 2001.
[3] IEEE Learning Technology Standards Committee's Learning Object Meta-data Working Group, Approved LOM Working Draft 5 (WD5), http://ltsc.ieee.org/wg12
[4] Metadata for Multimedia Information Workshop, CEN Information Society Standardisation System, http://www.cenorm.be/isss/Workshop/delivered-WS/MMI/Default.htm
[5] J. Hunter. "MetaNet – A Meta data Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains", Journal of Digital Information, volume 1 issue 8, http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter, February 2001.
[6] MARS (Multimedia Advanced brokerage and Redistribution Surveillance) Project, http://www.upf.es/esup/dmag
[7] Agentes, Estructura de la web, http://www.dcc.uchile.cl/~rbaeza/inf/web.html (in Spanish).
[8] Guide to Meta-Search Engines, http://www.indiana.edu/~librcsd/search/meta.html
[9] XML (extensible Markup Language), http://www.w3.org/XML
[10] Metadata Retrieval, http://www.library.yale.edu/orbis2/public/wkgroup/FINALcatmeta.htm
[11] Distributed Multimedia Applications Group website, http://www.upf.es/esup/dmag

https://doi.org/10.23106/dcmi.952106664