Application of the Dublin Core format for automatic metadata generation and extraction

Ernesto Giralt Hernández
Sand (www.esand.net)
ernesto.giralt@esand.net

Joan Marc Piulachs
Sand (www.esand.net)
jmarc.piulachas@esand.net

Abstract:

This article describes a set of services and tools to be used by information systems to obtain metadata collections in a automated fashion from online content or other electronic repositories. This multi-module software uses a service-oriented schema based on the analysis of electronic resources published on the web or local networks. Through several algorithms is capable of generate and extract metadata elements from documents, explicitly declared or as a result the document's content analysis.

The adopted model makes it a scalable and distributed system, which may be extended to recognize more formats, types of meta-data and other sources.

Keywords:

Metadata, automatic meta-data generation and extraction, Dublin Core.

1. General Concepts.

Any planning of a new information system, oriented to extract and to process content from the chaotic repository that represents Internet today, run into the problem of deal with non catalogued contents, dissimilarity of formats, the media kaleidoscope, a lot of protocols and different standards needed to access to the final resources.

The born of the web search engines has brought certain light to the old problem of the content search and recovery, but the totality of the Web even is not fully covered by the combined work of the automatic agents and the human work cataloguing tasks [Aguillo, et al, 2005]; and a main part of the difficulty resides in the poor formalization of the contents or even no formalization at all.

Everybody knows that the content into the first established networks and services, which later became what actually is known as "the web", were not organized to be later easily recovered and processed automatically [Hully, 1997]. HTML and the most important originally designed protocols and formats (like FTP, Gopher, HTTP and others), was not included elements to mark and classify the content with a semantic meaning, or the included elements were not enough to accomplish a full work of tagging and describing, or even these are not extensively and regularly used by the publishers. It was later, when SGML and XML appear that content-oriented formats were introduced as an stable concept and more elaborated model.

This is not a kindly scenario to the documentation's professional world, where the most important strength resides in normalization and the control over the content itself, sources, authors and the information's life cycle.

After just a few years, the way to a fully automated access to information has born, and after many efforts, protocols like Open Archives Initiative (OAI-P MH) and Z.2950, languages and mark formats like METS, TEI, RDF and Dublin Core, are getting more and more relevance to the publishing world, the applications design and the design of new information networks.

2. Introduction.

Sand has developed Dublin Core Services (DCS) as a bundle of tools and services designed to find and generate meta-data. The goal is to process documents found in networks, local or remotes, and obtain any available information from the content itself or generated after the use of analysis algorithms, filter and conversions, described later in this document.

DCS differs from similar systems in it's ability to recognize and process any existing format and convert all found meta-data into a unique format: Dublin Core, which is the major world metadata initiative, the most popular standard (ISO 15863-2003) in description, recovery and management of digital information, and an infrastructure accepted by the whole community involved in the construction of the "semantic web".

By selecting Dublin Core as the key format for data generation, DCS has inherited the simplicity and interoperability of this format.

DCS has been designed to deliver services for library automation software, search engines, or database management systems that need to expand and associate metadata and registers in a automatic fashion; web content or web sites producers and managers, clients applications like web browsers, or any other system or application able to take advantage of the DCS 's potentialities to generate and "discover" the metadata. The results can be used as non-controlled vocabularies, full terms list, keywords index list, or other kind of terminological/linguistic resource inside of new applications

With this intention in mind it was developed as a web service, to act as a mediator between information providers and others users systems or applications.

For all the cases, the most known protocols and standards has been applied, to guarantee full and open connectivity with any other computer language or client system. The specifications to use the framework include:

- Dublin Core Metadata Initiative Specifications (http://www.dublincore.org)
- SOAP / WSDL
- Creative Commons
- RDF (http://www.w3.org/RDF/)
- XHTML1.0 (http://www.w3.org/TR/xhtml1/)

3. Model and Architecture

The engine is modularly designed to provide a collaborative interaction between the different parts to accomplish the work of translation and generation. En the whole cycle for a single request we can identify the following parts:

Queries
Information Sources
DCS Processes

1.1. Queries

The type and syntax of the query depends on the information source to which is desired to consult. After the engine receive the query, this is analyzed by a process that determine to which sub-system will have to be given. For example, for a group of keywords (ex: "metadata applications") the query analyzer will relation it with the harvester that use a web search engine (like Google) as an information source. (see Harvesters)

At the moment, two types of consultations are distinguished: open text or queries with operators. The open text consultations can be interpreted of diverse way by harvesters but it is typical that become to a list of keywords, as in the previous example. The queries with operator are more specific and directly determine the information source that is desired to use. For example, the query "isbn:0345339711" will interrogate the information sources registered by the engine to get the ISBN information about a book in particular.

The URL receives a special treatment, since they do not fall within the classification of the queries before mentioned. And as it is a resource already located, the URL is given directly to the correct harvesters without any other analysis process.

1.2. Information Sources

They are the resources, like servers, services, data bases, etc., that can contain/serve documents or information, susceptible to be analyzed by DCS. Examples of content sources are the web search engines that return online document collections from queries, and the relational databases management systems that serves the content organized in tables and registries.

1.3. DCS Processes

They make the work of analysis and treatment of the content.

- **Harvesters:** Processes that offer access to different content sources.
- **Crawlers:** Processes that are responsible to analyze the different contents types and formats that DCS is able to understand.

4. DCS capabilities

- DCS generate/extract metadata items for the following digital formats:
- RSS (all the versions)
- bibTex
- HTML/XHTML
- Images (PNG, GIF, JPG, BMP, TIFF, among others)
- Custom XML documents:
  - Amazon Web Services
  - Creative Common digital licenses (http://creativecommons.org/)
- RDF/XML
- Free format text
- RTF
- PDF
- Processes content in 11 different languages. They are included: Spanish, English, Italian, German and Arab.
- Dictionary of 5300 empty words and non-preferred terms. This dictionary is formed by words of the same languages that DCS is able to process.

– Recognizes more than 70 variants and alias of metadata elements used in documents HTML/XHTML expressed in 3 different languages (see the last update in http://www.describethis.com/xml/aliastable.xml)

5. Content Analysis

The results that receives a client application after a query, is created using different filtrate methods, conversion, extraction and generation of Dublin Core elements from the content of documents. These methods can be summarized as it follows:

- Statistical Content Analysis Methods (TF/IDF)
- Stylesheet processing (XSL)
- Regular expressions processing
- Direct formats translation and/or filtrate
- Empty words recognition lists

6. DCS's functionalities extension and improvement

The DCS design anticipates the incorporation of new elements and processes that can extend their capacities. The incorporation of new crawlers and harvesters is possible without having to change to the operation of the services and the lists or databases used by the processing engine to make the content analyses and the other processes.

The more important points in which improvements and capacity of extension have been anticipated are:

– **The system of analysis and processing of content**: Extending the capacities and details of this system it is possible to process more types of formats and resources.

– **The information sources**: This is a vital extension for the service development. Adding new information sources the system is able to have access to more documents and of varied types and therefore to increase its universal application. Some can be considered, such as:

Access to OAI servers (Open Archives Initiative)
Access to relational databases.
Directories and commercial online database.

7. Describethis.com

Describethis (www.describethis.com) is an example of the application of DCS to build a functional and usable software taking advantage of the automatic generation and metadata discovering process. Through the simple provided interface, the user is able to interact with the engine and get the Dublin Core registers in response to their queries.

The site makes available all the DCS's functionalities, except those ones that involve the processing of big files (video and audio stream files, for example), and can be used as example of the real capabilities and type of results that we can expect from the DCS.

8. Metadata Accuracy

The results of the content processing are delivered the final consumers (user or application client) in form of Dublin Core registers. The metadata elements that appear in these results are generated or translated from a correspondence mappings between the well-known formats and the Dublin Core format. In some cases these formats contain more elements and more complex than the corresponding ones in Dublin Core, or in the other hand, the format have missing elements, necessaries to obtain registers with a minimum level of document description, and these must be generated automatically

Anyway, the correspondence mappings, and the criteria of conversion of a format in another one decide the level of detail and accuracy of DCS. The participation of experts which validate or propose these mappings and criteria in form of program of joint work, forum, or any other initiative, it would be an essential step to really create a uniform and valuable system for the professionals of the information and documentation communities.

In the system these correspondence mappings are expressed in form of style sheets XSL and rules, so the capacity to change, to update or to improve is also guaranteed.

9. Conclusions

The possibility of having tools that are able to extract any type of metadata and that recognize different digital formats, increases the independence and integration capacity of the information systems. Using to DCS like a functional component in a more complex system, it can make the necessary tasks of translation and conversion of the dissimilar digital formats towards Dublin Core, improving this way the operation of the rest of modules.

Among the advantages to apply DCS, the developers are able to direct to all their efforts to more important parts of their informational systems and applications, leaving to DCS the heavy tasks of treatment of multiple and dissimilar documents formats, different databases, different engines and the all the drivers and protocols to reach them. In addition, all this is performed by DCS in a scheme of services and collaboration with other systems, so is possible to continually enhance and update their own systems with less costs and efforts.

In the same way, the professionals of the documentation and the information can be assisted in their cataloguing tasks, in aspects such as the document indization and the electronic classification.

DCS also proposes the creation of a professional program, or forum, to track and develop a bundle of recommendations and rules to effectively create the more accurate correspondence mappings between any digital format and Dublin Core, to be implemented and applied by DCS and that can be taken as a valuable resource, open and free, for professionals and information specialists.

References

1. Aguillo, I. F., et al, 2005. Análisis Cibermétrico de los principales motores de búsqueda. *In* 9^{a} Jornadas Españolas de Documentación (FESABID 2005), Madrid, 14-15 Abril, 2005.

2. Hully, 1997. Metadata: ELAG'97, Gdansk, June 18th 1997. BIBSYS; (http://www.bibsys.no/elag97/metadata.html)

3. Salton y McGill, 1983. Salton, G. & McGill, M. (1983). Introduction to Modern Information retrieval. New York: McGraw Hill.