INFOFLEX: FLEXIBLE AND DISTRIBUTED CONTENT MANAGEMENT

USING WEB SERVICES AND SEMANTIC WEB TO MANAGE CONTENT

Jesús Villamor-Lugo, Norberto Fernández-García, Luis Sánchez-Fernández, Jesús Arias-Fisteus
Department of Telematic Engineering, Universidad Carlos III de Madrid, Spain
Emails: jvl@it.uc3m.es, berto@it.uc3m.es, luis@it.uc3m.es, jaf@it.uc3m.es

Tomás Nogales-Flores, Antonio Hernández-Pérez, David Rodríguez-Mateos
Department of Librarianship and Information Science, Universidad Carlos III de Madrid, Spain
Emails: nogales@bib.uc3m.es, tony@bib.uc3m.es, pirio@bib.uc3m.es

Keywords: Content Management, Web Services, Semantic Web

Abstract: The development of information and communication technologies and the expansion of the Internet means that nowadays there are huge amounts of information available via these emergent media. The need to manage such information, which was in the past stored on paper media, has become apparent in different fields. A number of content management systems have appeared which aim to achieve this task. Most of these systems are oriented towards Web publishing on a central site, and they do not support collaboration among several, distributed sources of managed content. In this paper we present a proposal for an architecture for the efficient and flexible management of distributed content.

1 INTRODUCTION

At different application domains, there exists the need to offer a service that permits managing, producing and offering quality information. The entities devoted to the production and/or the register of contents must face several requirements. Some of them are:

- **Content creation and manipulation management**. A mechanism for the coordination of related work among different persons is needed.
- **User Profile Management**. At an environment in which the contents are created and accessed, we must be able to define mechanisms to identify who must create, modify or query some documents. It may also be necessary to provide some document review mechanisms, as previous step for the internal or public publishing.
- **Decision criteria on the contents to be published.** Two types of factors must be considered in the decision process: technical ones (i.e. which contents will be published on the Web site, links validation, etc) and political ones (i.e. trust degree on the information, confidentiality, etc).
- **Integration of several fragments in one document** (i.e. for constructing a web page from several content sources).
- **Validation**. The contents and functionalities offered by the system must be checked in order to detect errors or inconsistencies.

In this paper, we show an architecture (nowadays at development process) for the management of flexible and distributed contents: the **Infoflex System** (Sánchez et al., 2003). It offers the required facilities for a Content Management System (CMS). These facilities are currently being implemented using the following technologies:

- **XML** for the description (or encapsulation, if text data) of the information. The XML documents are stored in an XML Database

to ease the recovering. Likewise, the use of XML will permit to export the contents to different formats (using XSLT style sheets).

- **A light-weight workflow management system.**
- **A version management system.**
- **Web Services (WS),** to deal with collaboration among different distributed CMSs.
- **Business Process Execution Language for Web Services (BPEL4WS) (Andrews et al., 2003),** to define a complex interaction based on the coordination of different Web Services.
- **Semantic Web,** to deal with heterogeneity in data models.

The novelty of Infoflex is the use of Web Services and Semantic Web to allow a user (human or informatics system) to retrieve information from different distributed and heterogeneous management systems (even when their location is unknown) and to perform a unique query on several systems.

This paper is organized as follows. Section 2 introduces Web Services and how they are used in our framework. Section 3 introduces Semantic Web and its use in Infoflex. Section 4 describes the proposed working model and architecture. Concluding remarks complete this paper.

2 WEB SERVICES AND ITS APPLICATION IN INFOFLEX

Web Services (Web Services, 2003) have been defined as self-describing applications that can discover and engage other web applications to complete complex tasks over the Internet. This technology was designed to ease the cooperation among remote applications. Web Services technology is supported by a set of companies (IBM, HP, SUN, Microsoft, ...) and standardization organisms like the World Wide Web Consortium (W3C).

The Web Services architecture is a middleware technology based on XML data formats for encoding the exchanged messages (SOAP), service definitions (WSDL), Service registering-and-discovering facilities (UDDI) and Internet standard protocols (as HTTP or SMTP).

The interoperability model proposed by Web Services perfectly fits our distributed contents access requirements.

Initially, users don't know the location of the providers. So Infoflex must allow a user (human or software application) to query and retrieve information in several distributed CMSs in the desired domain (news agencies, content syndication sites, etc). In our framework, CMSs provide access to their functionalities through Web Services. This way, clients use UDDI repositories to look for them.

The choice of Web Services as communication mechanism provides two advantages:

- **Interoperability and platform independence,** because Web Services provide a standard text-based communication mechanism.
- **Data format compatibility,** because CMSs typically store their data as XML documents.

3 THE SEMANTIC WEB AND ITS APPLICATION IN INFOFLEX

The Web was born with the idea of making information accessible by humans. But with the quick Web growth, the amount of data available is so big, that, to accede it in an efficient manner, new automatic mechanisms are required. On the other hand, nowadays the Web is not only information but also services. For final users, it would be interesting that software agents could use such services in an automatic manner, freeing humans of tedious tasks.

This desired degree of automation could only be achieved if we allow machines (and not only humans) to understand resources and contents. This requires giving meaning (semantics) to data: we need a Semantic Web (Berners-Lee, 2001). With this purpose the Semantic Web Activity Group (Semantic Web, 2003) of W3C began its activities in 2001. Two are the main development areas:

- **Metadata:** the objective is the development of mechanisms that allow the description of contents, represented in web publishing formats like HTML or XML, in a machine-readable format. Perhaps the most important development in this field is the RDF (Resource Description Framework) recommendation (RDF, 2003).
- **Ontologies:** if we want contents to be interpreted by machines, it is not enough to provide machine-readable descriptions: we also need machine-understandable descriptions.

Ontologies are used with this purpose. An ontology is a model of a knowledge domain, that is, a set of concepts that are typically used in that domain and a set of relations among this concepts. These models are formal enough to allow machines to reason about them. This way, inferences can be made automatically and knowledge could be improved. When we want to describe a resource using metadata, we do not do it using natural language but the set of concepts included in the ontologies. So, machines could understand this metadata and the contents described by them. In this field the most important contribution of the W3C is OWL (Ontology Web Language) (OWL, 2003). This XML application is currently under development.

Infoflex uses these technologies to face the problem of accessing heterogeneous sources of information. A domain can be modelled with many different ontologies. To allow interoperation, we propose to use a general and consensual ontology. This ontology will be written in OWL. It will describe all the concepts and relations of interest. Every CMS must develope its own ontology, which is used to describe its contents. These CMS-specific ontologies must use the general ontology as a reference in one of the following ways:

The specific ontology can be just a subset of the general ontology.

- The specific ontology can be used for translating between the general ontology and another internal (and typically unknown) ontology used by the CMS to describe its information.
- This idea of using ontologies to deal with heterogeneity of information sources has been previously addressed in other papers like (Brisaboa, 2002).

To use a general ontology, it must be accessible for all CMSs, so we propose to store this ontology in a global service. This service is called Query Service (QS) and it will be implemented as a Web Service. All CMSs must access it, query the general ontology, and register their specific ontologies.

Clients must also query this QS because they typically do not know the CMS data-model and query language. So, they must query the QS, using a commonly-accepted language and data-model (the general ontology). As a result, they obtain a specific query adapted to the CMS data-model and language.

This way, heterogeneity in data models and formats is hidden to final clients. Next section explains this process in detail.

4 ARCHITECTURE AND WORKING MODEL

In this section we will introduce the proposed working model of our Infoflex framework.

The main actors involved in the full process are:

- **Clients**, who request information using the general query language. This language is not specified yet. It will be based on XML (XQuery, RDF, or similar).
- **Query Servers**, which translate queries from the model represented in the general ontology (used by the Client to make his queries) to the CMS information model.
- **Content Management Systems,** which store and manage the requested information.
- **UDDI repositories,** where all the Web Services involved on this scenario are registered.

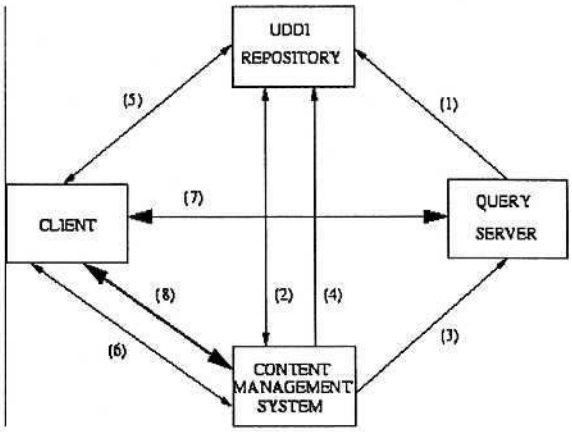

Figure 1: Architecture and Working Model

In figure 1, we can see the steps which are necessary to obtain information from one provider. Without loss of generality, we suppose in our exposition that only one component of each kind is available.

Basically, the steps are:

1. The Query Server registers two Web Services into a UDDI repository. One WS (**WSc**) is for Client access (to make queries). The other one (**WSp**) is for CMS access (to register its own information model ontology in the Query Server)
2. The CMS access the UDDI looking for the Query Server. It obtains a reference to the WSp of this Query Server.
3. The CMS registers its ontology in the Query Server, whose reference has been obtained in the previous step.
4. The CMS generates dinamically a **BPEL4WS** document. This document specifies the steps (WS invocations) that Clients must follow to query it. Two WS invocations are necessary at least: one for translating the query (operation done by the Query Server) and the other one to achieve the information (querying a Content Provider WS). The BPEL4WS document also contains information on how to compose the results of such invocations (use the responses of one service as the parameters of the other one) and how to react when a unexpected situation occurs (for example, the Query Server is unreachable). This document must be obtained by the Client to enact it and access the CMS, so this Provider must publish its BPEL4WS. In our proposal, the document is retrieved using a WS, which is registered by the CMS in a UDDI.
5. The Client access the UDDI looking for the CMS. It obtains the reference of its WS.
6. The Client invokes the Web Service of the CMS. It obtains as the result the BPEL4WS document. The Client executes this document to retrieve the desired information.
7. During the execution, the Client writes a query, using concepts in the general ontology. The BPEL4WS execution engine uses this query (as is indicated in the BPEL4WS document) to invoke the QS, which translates it to the CMS model.
8. The obtained traslated query is automatically used by the BPEL4WS engine to invoke the CMS and retrieve the desired information (if it is avaliable).

5 CONCLUSIONS

In this paper we have presented a proposal of an architecture for the management of contents in a flexible and distributed manner. This proposal is based on advanced Web technologies such as XML, Web Services or Semantic Web. We expect for this system to be useful at environments like business to business portals, news agencies or content syndication sites.

ACKNOWLEDGMENTS

This work has been partially financed by the "Ministerio de Ciencia y Tecnología de España" (Science and Technology Ministry of Spain) through the TIC2003-07208 "InfoFlex" Project.

REFERENCES

Berners-Lee, T., 2001; Hendler, J.; Lassila, O. "*The Semantic Web*". *"Scientific American*". Vol. 284, nº 5. 2001. pp. 34-43.

Brisaboa, N., 2002; Penabad, M.; Places, A.; Rodríguez, F. "*Ontologies for Database Federation*". "UPGRADE". Vol. III, n. 3. 2002. pp. 52-61.

World Wide Web Consortium. Web Services Activity. Consulted at: http://www.w3.org/2002/ws/ (28-10-2003).

World Wide Web Consortium. Semantic Web Activity. Consulted at: http://www.w3.org/2001/sw/ (28-10-2003).

World Wide Web Consortium. Resource Description Framework (RDF). Consulted at: http://www.w3.org/RDF/ (28-10-2003).

World Wide Web Consortium. OWL Web Ontology Language Reference. Consulted at: http://www.w3.org/TR/owl-ref/ (28-10-2003).

Sánchez L., Villamor J., Arias J., Fernández N., Nogales T., Hernández A., Rodríguez, D., 2003 "*Gestión Flexible de Contenidos Distribuidos*". Boletín de Redlris, diciembre 2003-enero 2004, Vol. 22.

Andrews T., Curbera F., Dholakia H., et al., 2003; *"Business Process Execution Language for Web Services. Version 1.1 Specification."* Available at: http://www-106.ibm.com/developerworks/webservices/library/ws-bpel (28-10-2003)

N. Guarino (ed.), 1998; "*Formal Ontology in Information Systems*". Proceedings of FOIS'98. Amsterdam, IOS Press, pp. 3-15.

Documentum, Enterprise Content Management. Consulted at: http://www.documentum.com/ (28-10-2003)

Microsoft Content Management Server. Avaliable at: http://www.microsoft.com/cmserver/ (28-10-2003)

Zhou N., Meliksetian D., Weitzman L., et al. 2001; "*XML Content Management: Challenges and Solutions*". Proceedings of XMLEurope 2001.