XML WAREHOUSING MEETS SOCIOLOGY

François-Xavier Dudouet
Laboratoire d'Analyse des Systèmes Politiques
Université Paris X
foxd@club-internet.fr

Ioana Manolescu
Projet GEMO
INRIA Futurs
ioana.manolescu@inria.fr
W3C Member – XQuery WG

Benjamin Nguyen
Laboratoire PRiSM
Univ. Versailles St-Quentin
benjamin.nguyen@prism.uvsq.fr
W3C Member – Semantic Web BP

Pierre Senellart
Projet GEMO, INRIA Futurs
École normale supérieure
pierre@senellart.com

ABSTRACT

In this article, we describe a novel application of XML and Web based technologies: a sociological study of the W3C standardization process. We propose a new methodology and tools, to be used by sociologists to study the standardization process, illustrated by the W3C XQuery Working Group. The novelty of our approach has many facets. Information Technology (IT) has received little attention from sociologists, yet the standardization of the Web is a crucial issue, both economical and political, based on the use of a semi-structured content warehouse. We introduce a modeling and querying approach of an XML content warehouse, and show it produces high added-value information. This information is used to conduct a preliminary sociological analysis of the XQuery standardization process.

KEYWORDS

Political Science, Sociology, Standardization, Web Warehousing, XML, XQuery Working Group.

1. INTRODUCTION

Research work in social science needs to consult and analyze vast quantities of information. For instance, an analysis of unemployment in a given geographical area would require consulting census data, labor ministry data, independent surveys… A social scientist would issue hypotheses on his topic (e.g. correlation between immigration and employment), validate them on the collected data, and then issue new hypotheses. Nowadays, more and more human activities involve some Web technology. As a consequence, a tremendous amount of information documenting various human activities from business to culture, industry or information has moved online. While social science research could clearly benefit from Web data storage and analysis tools, these are currently not available in this domain. Some scientists do use general-purpose database management systems (DBMSs), however data is most often entered manually or by copy-paste from a Web browser, since such data does not fit the structured DBMS format.

The authors of this paper come from two fields: social science and database management, whishing to bridge the gap between these worlds, by analyzing the needs of sociologists through an example: the sociological analysis of the establishment of a W3C Recommendation. From the data management viewpoint, we aim at establishing the requirements for a data acquisition, storage, and analysis tool, to be used by social scientists taking advantage of the Web-based data. From the sociological point of view, the objectives are twofold. First, we analyze the workings of the W3C standardization groups, thus providing the intellectual tools to participate in the process and influence its outcome. Secondly, we extend analysis methods previously applied to the establishment of international regulation (e.g. for drug control [12]), to the process of IT standardization.

In this work, we focused on the process of establishing the XQuery W3C Recommendation, due to several factors. First, the standardization process is now close to the end, enabling us to reason over a full-blown process. Secondly, the acquaintance with XQuery of the computer scientists involved provides domain-specific knowledge to the project. We have performed an initial analysis on the public mailing list of the W3C XQuery working group (WG), and designed a set of interesting concepts for its sociological analysis, such as: individuals, organizations, discussion topics, etc. We extracted the mailing list content into a database, and performed a preliminary data analysis. Our main contribution is a reflection on how sociologists and computer scientists can collaborate, and produce tools and methods of Web data analysis, to complement the traditional statistical tools. Our approach innovates in sociology research, by using XML-

centered technologies, and by using database-style tools to analyze human interactions captured in mailing list content.

Section 2 briefly reviews the related work, in the field of Web data integration and warehousing, and in the field of sociological study of standardization bodies. We will introduce, in Section 3, the concepts crucial to the particular study of the XQuery standardization that we undertake. Section 4 describes our solution in order to model and query our specific problems, while Section 5 presents some example queries and results we obtained during our analysis. This work is part of a French government-sponsored project on the analysis of standardization processes in the area of Information Technology (IT) [3].

2. RELATED WORK

There has already been a lot of work on data warehousing, mediation and integration [22]; see [10] and [21] for a survey on OLAP, data warehousing and materialized views. However, these technologies only deal with highly quantifiable data, which is not the case for sociological data. The concept of *content warehousing* has been introduced in [1] and [2]. A content warehouse is a warehouse of qualitative information that has no trivial mathematical processing method, inappropriate for regular OLAP-style processing. This information, because of its high heterogeneity, can only be integrated by using a semi-structured data model.

Modern sociology was born at the end of the 19th century dealing with large amounts of statistical data. Since then, the methodology has been well improved. The latest important issues are certainly *factorial analysis* [4], [7], [14] (crossing large amount of personal information, aiming to build the typology of a social group) and *network analysis* [5], [8] (aiming to exhibit relationship structures). These otherwise helpful approaches face some limitations: *technical* (qualitative information such as personal views on facts is not accounted for) and *relative to the time dimension* (only possible representation of a social fact: a snapshot). Social sciences are interested by the standardization process, since defining technical standards is also choosing firms and countries which will control the technology, which has a clear economic and political impact. This may explain why, as observed by the OECD, many standards dominating the IT market, are not the best from the technical point of view [15] and [6]. So, the questions of *who*, *how* and *for what* standards are adopted become crucial. Answering these questions requires the use of social science tools. Despite the importance of understanding the standardization process, few social sciences have addressed this topic so far. [17] and [18] study the impact of standards in companies. The most advanced results on the international standardization process have been obtained by the Stockholm Center for Organizational Research ([9], [20]). However, the IT standardization processes remain vastly unexplored [19]. Our study is an attempt to fill this gap, exposing the actors and the mechanisms within the W3C standardization of XQuery.

3. TARGET APPLICATION: XQUERY STANDARDIZATION

The World Wide Web Consortium is central to the development of the Web; around 90% of the W3C Recommendations can be seen as de facto *standards*. However, the process of standardization is little understood. Even the people in the center of the process need a way to comprehend the *way it all works* [13]. Inside the W3C, discussions are usually held via *e-mail*. *Teleconferences* are also organized, but most of the time, they are to settle issues already dealt with on the mailing list. Just like live discussions, some e-mails are private, and are withheld to WG participants, but others are public, such as the final recommendations, or answers to questions that outsiders may direct to the experts. Thus, the arena that we are interested in is in fact quite accessible via [23]. At this URL, we will find not only all the participants, but also all their public statements and reactions of the last four years. Our study focuses on the public e-mails posted on the XQuery comment mailing list: about 5,000 e-mails that can be regrouped into threads, by determining which e-mails answer each other. Our goal is to build a *semi-structured data warehouse* model to store and process (by using XQuery!) this information corpus.

The social study of the standardization process must answer questions such as: Who are the *individuals* involved? What *relationships* do they have between each other? What *role* do they play in their organizations? Answering these questions should help expose links between individuals, the organizations they stand for, the context in which they act, and their final objectives with respect to a given standard. These

questions determine the conceptual structure of the database to be set up. Answers to these questions are crucial in understanding how and why a standard is built the way it is, which players succeed in influencing it, and how an organization could optimize its impact on the standardization process. Such information could help involve in the W3C user groups which are not currently well-invested, such as public institutions and universities in the case of the XQuery WG (see Section 5). Moreover, this kind of study could help the W3C itself improve their knowledge on the human and social interactions taking place. Our approach could also be applied to other collective Web-based negotiation and decision-making processes.

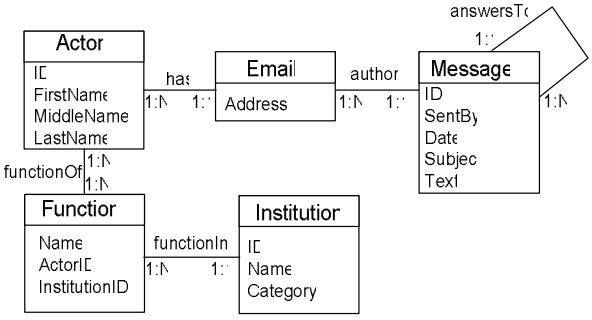

Figure 1: Conceptual model for the social analysis.

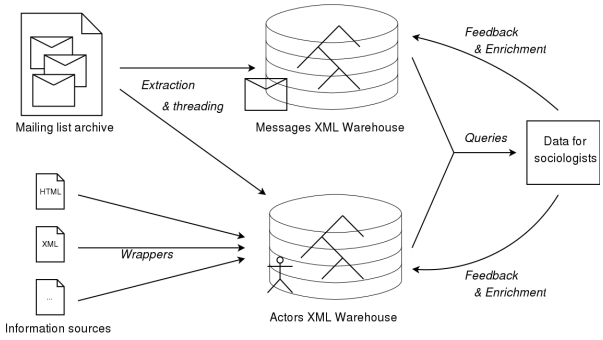

Figure 2: Warehouse construction process

4. METHODOLOGY

To exploit large data volume, a methodology which needs as little human intervention as possible is critical. However, human input and feedback can (and should) be used to tune and enrich the system.

We start with a conceptual modeling of the entities of interest to the sociological study, depicted in a standard Entity-Relationship diagram in. Standardization actors are the individuals that post messages on the mailing list. Each author has a unique ID, and first, middle and last name, and may have multiple e-mail addresses. Furthermore, an actor can have multiple roles within different institutions, e.g. be a university professor and a consultant for a company. Messages are posted from an e-mail address; we capture the date, author, subject, and text of each message. We then *map* data sources of interest to entities and relationships of this model, and *load* the data sources into our warehouse.

Figure 2 shows a diagram of the process used in the case of our mailing-list application. Content is extracted from a mailing list archive in a fully automatic way, into a semi-structured message warehouse. Information kept in this warehouse includes the thread structure of the mailing list (which message answers to which) as well as the author, date, subject and full text of each e-mail. Additionally, another warehouse is built to store information about actors of the mailing list and their institutions. This information comes first from the mailing list itself: names in the `From:` field can be used to identify actors; institutions are identified from the domain names in e-mail addresses and expeditions machines (`Received:` field). This could be complemented by other information sources, for instance found on the World Wide Web (HTML or XML data describing mailing list posters, actors' home pages, institution websites), using wrappers.

We have chosen to represent our content warehouse in XML, for the following reasons. First, XML represents *semi-structured* information, in which structured data (e.g. for each message, e-mails and dates) can be mixed with raw text (message body). XML is also *flexible*: new information can be added at will by adding new elements or attributes. Then, XML is intended to be the language of the Web, making it suitable to the writing of *wrappers* for Web pages or other data found on the Web. Moreover, a mailing list has an inherent *tree structure* (message A is the child of message B if B answers to A), which requires a nested representation format such as XML. Finally, XML remains *simple* to understand.

Therefore, for this and many similar applications, XML is a real step forward for quantitative sociological analysis, which has traditionally been carried out in the context of relational databases. The choice of XML naturally leads to using XQuery itself as an interrogation language: the expressive power of XQuery allows the formulation of the complex queries that we need and its declarative nature makes it much easier to use than alternative languages such as XSLT.

5. EXPERIMENTATION

Our experimentation dealt with the `public-qt-comments@w3.org` mailing list which is the W3C public list for submitting comments on the proposed XQuery, XSLT 2.0 and XPath 2.0 recommendations. A mailing list archive was obtained from the public mailing list server. The mailing list contained 5,626 messages at the time of extraction. A Perl script was written to convert this archive into the XML Data Model described in Section 2. This script uses the Perl `Mail::Thread` [11] module to build the thread structure, based on Jamie Zawinski's threading algorithm [24]. Wrappers for HTML and XHTML web pages (only available to W3C members) describing the list of members of the XQuery WG were also written, in order to add information about membership in the XQuery WG to the actors warehouse.

We now describe our data analysis process, and corresponding sociological interpretations. We issued a set of XQuery queries, which were processed by the QizX system [16]. Each query brings information that can be used to validate existing hypothesis and formulate new ones.

First, a complete list of institutions is extracted from the actor warehouse through a query. Each institution is then manually annotated with one of the following types: ***corp*** for IT companies, ***univ*** for academic institutions, ***org*** for not-for-profit organizations such as the ACM, ***prov*** for providers of Internet access and e-mail (obtained due to the extraction procedure, because actors send e-mails from accounts hosted by the providers), ***pers*** for personal domain names, and ***unknown*** for the remaining sites (about 5%). This typology is re-injected in the warehouse for further interrogation. We refine our categorization by devising a set of interesting **profiles**, where each profile consists of 1 or more types of institutions. Based on this categorization, the number of messages issued by actors of each profile is obtained (see Table 1).

Table 1: Distribution of actors' affiliations

Profile	# actors	# posted msgs.
Companies	135	2,689
Universities	39	112
Organizations	33	197
Companies & Universities	3	532
Companies & Organizations	22	1052
Universities & Organizations	6	36
Non specified	65	681
Total	**303**	**5299**

An analysis of e-mail addresses of users posting on the mailing list shows that most of them (37%) come from IT **companies**. This can be explained by the impact of W3C recommendation on the success of a technology commercialized by a company, thus the economic interest of companies in the making of recommendations. The mail distribution confirms the companies' domination. Out of 5,299 mails, 4,273 (81%) come from people connected to at least one company. A cademics (individuals connected with universities but not with companies) have a low participation rating: 3 messages on average posted by an "University" actor, and 6 for the "University and Organization". This is low when compared with a global average of 17 postings per individual, and an average of 20 postings per individual with a "Company" profile. Are academics less interested in the standardization process? Further analysis on the private list would likely answer this question.

Another interesting observation is that the most active participants have a mixed profile which includes a company affiliation. These results confirm the observations made in a previous study on international regulation [11]: the most active, and often most influential actors in regulation/standardization processes belong to several social contexts (such as companies and universities), especially when one such context involves economic interests (e.g. a company). We call such individuals *key actors* because they are at the interface of different social arenas, and bridge communities which were not directly connected.

A last set of interesting results is in the distribution of the answers of most frequent posters. We can distinguish three different trends in these posters answering "habits": **Balanced answers** (actors do not seem to privilege any specific person in their answers), **Unbalanced answers** (most common, a few individuals represent a majority of the answers of actors in this category), **Highly unbalanced answers** (an actor with a large number of messages only replies to 3 different people. His position is even more peculiar, since he hardly ever replies to anyone). These three different profiles show various attitudes that important posters

have on this *public* mailing list. We can already issue some hypotheses on the posters. The second profile shows important posters, who tend to continue discussions amongst themselves, or with other important posters, on the public mailing list. Is this the indication that they see the standardization process discussions continuing through the questions of "outsiders"? This would mean that non WG members (such as the posters asking questions) can have a big impact on the standard itself. Finally, the third attitude is explained by the fact that the actor is in fact not posting answers to the public mailing group, but rather comments, and explanations on certain parts of the standard. This also denotes a specific attitude: a sort of FAQ poster, who gives global answers and precisions, rather that replying to specific individuals' questions.

6. CONCLUSION

The goal of this experimentation was to show the feasibility of studying mailing lists, using XML technologies. However, we only studied public information, and although our results are convincing, they are somewhat limited. No doubt the sociological interpretation could be improved by using the private XQuery WG mailing list, and related information found on HTML pages. To this end we have contacted the W3C soliciting the right to include them in our analysis, since this information is confidential. Our tool can directly be reused in the context of other W3C WGs. Other user communities may also be interested in the tool, and might lead to different social patterns and interactions. We plan to work on the Web Content Accessibility WG, the MathML WG, or more generally the Linux Kernel mailing lists and IEEE standardization working groups. Other foreseeable applications may include the analysis of exchanges within a given corporation.

REFERENCES

[1]Abiteboul S., 2003. Managing an XML Warehouse in a P2P Context. *CAiSE Conference*.

[2]Abiteboul S. et al., 2002. Sets of Pages of Interest. *Bases de Données Avancées*.

[3]*ACI Normes des Politiques Publiques*. Available at http://www-rocq.inria.fr/gemo/Gemo/Prokects/npp/.

[4]Benzecri J.P. et al., 1973. *L'Analyse des données*. Dunod.

[5]Berkowitz S. D., 1982. *An Introduction to structural analysis*. Toronto, Butterworth.

[6]Besen S.M. and Farrell J., 1991. The Role of the ITU in Standardisation. *Telecom. Policy*, 15:4 , pp. 311–321.

[7]Bourdieu P., 1979. *La Distinction: critique sociale du jugement*. Les Editions de Minuit.

[8]Breiger R. L. et al., 1975. An Algorithm for Clustering Relational Data with Application to social Network Analysis and Comparison with Multidimensional Scaling. *Journal of Mathematical Psychology*, 12.

[9]Brunsson N. and Jacobsson B., 2002. *A World of Standards*. Oxford University Press.

[10]Chaudhuri S. and Dayal U., 1997. An overview of Data Warehousing and OLAP Technology. *SIGMOD Record*.

[11]Cozens, S. Mail::Thread Perl Module. Available at http://search.cpan.org/~simon/Mail-Thread.

[12]Dudouet F-X., 2002. *International drug legislation: 1921-1999*. PhD dissertation, Univ. Paris X Nanterre.

[13]Fernandez M., 2004. The Statesman, The General, His Lieutenant, and Her Sentry. *Keynote speech at the 1st Int'l Workshop on XQuery Implementation, Experience and Perspectives (XIME-P)*.

[14]Lazega E. and Vari S., 1992. Acteurs cibles et leviers : analyse factorielle de réseaux dans une firme américaine d'avocats d'affaires. *Bulletin de méthodologie sociologique*, 37.

[15]OECD, 1991. *La dimension économique des normes en matière de technologies de l'information.*

[16]*QizX Open: a free-source XQuery Engine*. Available at http://www.xfra.net/qizxopen.

[17]Segrestin D., 1996. La normalisation de la qualité et l'évolution de la relation de production. *Revue d'économie industrielle*, n° 75, janvier.

[18]Segrestin D., 1997. L'entreprise à l'épreuve des normes de marché : Les paradoxes des nouveaux standards de gestion dans l'industrie. *Revue française de sociologie*, 38:3.

[19]Special Issue on Standard Making: A Critical Research Frontier for Information Systems. 2003. Pre-Conference Workshop, *International Conference on Information Systems*, Seattle, Washington, December 12–14.

[20]Tamm-Hallström K., 2004. *Organizing International Standardization — ISO and the IASC in Quest of Authority*. Cheltenham, United Kingdom, 2004.

[21]Vaisman A.A., 1998. *OLAP, Data Warehousing, and Materialized Views: A Survey*.

[22]Widom J., 1995. Research problems in Data Warehousing. *Int. Conf. on Information and Knowledge Management*.

[23]W3C. *The W3C XQuery Working Group*. Available at http://www.w3.org/XML/Query.

[24]Zawinski, J. *Message threading*. Available at http://www.jwz.org/doc/threading.html.