Sociological Analysis of the W3C Standardization Process

XML Warehouse Meets Sociology

François-Xavier DUDOUET
Laboratoire d'Analyse des Systèmes Politiques
Univ. Paris X, 200, av. de la République, 92001 Nanterre Cedex, FRANCE
tel : (33) 1 40 97 76 52
fxdudouet@u-paris10.fr

Ioana MANOLESCU
Projet GEMO
INRIA Futurs
10, r. J. Monod, 91893 Orsay Cedex, FRANCE
tel: (33) 1 72 92 59 20
ioana.manolescu@inria.fr

Benjamin NGUYEN
Laboratoire PRiSM
Univ. Versailles St-Quentin
43, av. des Etats-Unis, 78000 Versailles, FRANCE
tel: (33) 1 39 25 40 49
benjamin.nguyen@prism.uvsq.fr

Pierre SENELLART
Projet GEMO, INRIA Futurs and École normale supérieure
45, r. d'Ulm, 75230 Paris Cedex 05, FRANCE
tel: (33) 1 72 92 59 29
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. We propose methods to analyze the standardization process, based on the usage of a semi-structured (XML) content warehouse. We introduce a modeling and querying approach of an XML warehouse, and show it produces high value-added information. This information is used to conduct a preliminary sociological analysis of the XQuery standardization process.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Computer applications and Sociology.

General Terms

Experimentation, Human Factors, Standardization.

Keywords

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

1. INTRODUCTION

Research work in social sciences needs to consult and analyze vast quantities of information, relevant to the topic of the particular research work performed. Thus an analysis of, say, the unemployed population in a given geographical area would require consulting census data, labor ministry data, independent surveys and studies and standard economic indicators which both measure and predict the employment figures for a given time period. A social scientist would issue hypotheses on his research topic (such as the correlation between immigration and employment in a given area), validate them on the data he has collected then issue new hypotheses. A database of relevant information is a precious tool for a social scientist, but it is not the only one. Another important ingredient of social sciences research are face-to-face interviews; during interviews, the scientist is able to direct the questions in order to obtain information from viewpoints that can not be assessed otherwise.

Traditionally, data gathering in social sciences takes the form of visiting a large number of libraries and photocopying useful information. But 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, in the form of HTML, XML, and PDF documents. For instance, national and international organizations are gradually publishing information online. A raw Google estimate of the number of Web pages under europa.eu.int (the European Union official website) is one million, those under gouv.fr (French government site) are estimated at 600.000, and those under .gov (US government sites, including the Library of Congress and the National Institute of Health) at 15.600.000.

While social science research could clearly benefit from Web data storage and analysis tools, these are currently not available to the social scientist. Some scientists do use general-purpose database management systems (DBMSs) to store their information. However, such data collection is typically done by manual insertion or copy-paste from a screen to a structural database format, since Web document formats do not fit the typically relational DBMSs used. Furthermore, such DBMSs are poorly suited to the social scientist's needs, since they do not support the inherent heterogeneity of various sources, and do not assist the scientist in the modeling and conception of these very specific applications.

The authors of this paper come from two fields; social science and database management. Our goal is to bridge the gap between these worlds, by analyzing the needs of sociologists through a concrete example: the sociological analysis of the establishment of a W3C Recommendation. The choice of this topic is due to our participation in a French government-sponsored project on the analysis of standardization processes in the area of Information Technology [50]; however, we see a wide range of applications and possible extensions of this work, as described in Section 7.

The study of information technology is an emerging hot topic for today's sociologists and remains a domain vastly unexplored. The success of innovative technologies depends on their widespread adoption, which depends on their recognition as a standard. Let us stress that recommendations such as those issued by the W3C, although not formal standards technically

speaking, are *de facto* standards once they are adopted by a worldwide user or industrial community. Henceforth, we will only use the term *standard* to refer to such technical documents. The role of the sociologist in standard-setting bodies such as the W3C is obvious: such bodies are concerned about the usability and accessibility of Web technologies to the greatest possible number. Understanding the processes of communication, technical initiatives, and standard production is useful for any organization (academic or corporation) with a focus on Web technology. Such understanding is equally useful for the standardization body itself, as it can lead to improved or better explained procedures.

The need for social analysis of IT, however, goes well beyond standard-setting bodies; even major companies, such as Microsoft, are hiring sociologists to analyze interactions on UseNet message boards (although their numbers are still thin for now [23]).

The W3C public working group pages capture communication, interaction between different people, called *actors*, and trace their actions, positions and declarations *through time*. These Web-based information sources inform the sociologists about the interactions between actors of a given process. In the particular case of elaborating IT standards, *mailing lists* are becoming the prevalent means of interaction between participants scattered around the globe, and working in different time zones; their archived content is typically published in Web pages. Moreover, when the participants do meet physically or attend teleconferences, written notes of the meeting or teleconference are taken and typically published on the Web shortly afterwards. A social scientist studying the standardization process must adapt to this situation by developing and using computerized data management tools, since techniques such as manual interviews and information collection become powerless to apprehend the sheer size of the corpus.

1.1 Goals

From our dual viewpoint on this application, the goals can be divided into two categories: *data management support for social sciences*, and a *social analysis of the W3C XQuery Working Group*.

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's tremendous potential of information. To this end, we propose the architecture for such a tool, based on readily available Web tools, models, and languages; rather than being a database researcher's "pet project", this architecture specifically targets ease of use and good support for the needs of the sociologist, as the authors' experience describes them. This tool should be extensible, based on Web standards, and automated to the greatest possible extent.

From the sociological point of view, the objectives are twofold. First, this analysis must demonstrate the workings of the W3C standardization groups, thus providing the intellectual tools to participate in the process and influence its outcome. Knowing who participates and how decisions are taken leads to a better understanding of the successes and failure of the consortium, and more generally the Internet phenomenon. Advertised as the result of uncontrolled individual initiatives, the weaving of the Web nevertheless respects standards issued by organizations such as the W3C, that are vital to its large scale interoperability. Secondly, we would like extend analysis methods previously applied to the establishment of international regulation (e.g. for drug control [15], or civil aviation [29]), to the process of IT standardization, and see how they carry over.

1.2 Roadmap and contributions

The work we report on here proceeded along the following path. We focused on the process of establishing the XQuery W3C Recommendation (to be issued very soon). We picked XQuery due to several factors. First, the standardization process has spanned over 4 years, and is now close to the end, enabling a sociologist 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 XML Query working group; this list is archived at http://lists.w3.org/Archives/Public/public-qt-comments. We have designed a set of interesting concepts and dimensions for the sociological analysis of this list, such as: individuals posting on the list, their organizations, discussion topics etc. Based on this conceptual model, we extract the mailing list's contents in a database, and perform a preliminary data analysis using a set of queries and a simple graphical tool.

The main contribution of our work is a reflection on how sociologists and computer science researchers can collaborate, and produce specific application tools and methods of Web data analysis, to complement the traditional interviews and statistical analysis. Our approach innovates over the state of the art in sociology research, by using Web technologies centered on XML, and by providing database-style tools to analyze human interactions as captured in mailing list content available on the Web. The second contribution of our work is a generic architecture, based on standards such as XML, XSL, and most importantly XQuery, for a Web data management tool to be used by social scientists collecting and analyzing Web-based data sources.

1.3 Organization

This article will read as follows. Section 2 briefly reviews the related work, in the field of data integration and Web 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. Section 5 presents some example queries and results we obtained during our analysis, as well as their preliminary sociological interpretation; the full sociological study is beyond the scope of this paper. Section 6 describes the generic architecture of a sociologist's data management tool, derived from our experience in this work. We finish with conclusions and perspectives for future collaboration and research in the field.

2. RELATED WORK

2.1 Web warehousing

There has already been a lot of work on the topics of data warehousing, mediation and integration of data [33]. We refer to

[13] and [19] for a survey on OLAP, data warehousing and materialized views, and of methodologies and tools for constructing classical data warehouses. However, all these technologies only deal with highly quantifiable data, in general integers or floats.

In the case of the construction of a sociological warehouse, the approach is radically different. The concept of *content warehousing* has been introduced in [1] and [3]. Let us give a rapid overview of this new approach. A content warehouse is a warehouse of qualitative information such as sociological data that has no trivial mathematical processing method. Representing the relations between participants or their exact roles does not lead to information that can be processed in a regular OLAP approach. On the contrary, this information, quite often available from various sources on the web, is highly heterogeneous, and can only be integrated by using a semi-structured data model. We developed in [3] a methodology for the design and construction of a content warehouse focused on food risk, the *e.dot* project [37] based on the Active XML development platform [35].

Central to the task of retrieving information from the Web are the Crawlers. In our case, we do not want to retrieve *all* the pages of the Web but retrieve specific information found on public web pages pertinent to Internet standardization.

2.2 Data integration and XML

Data integration systems offer the possibility to query heterogeneous and distributed data sources as if they were contained in a single database. The classical architecture of a data integration system comprises a *mediator* [34] offering the integrated database view to the user, and a set of *wrappers*, which make the connection between individual data sources and the mediator. Research in this area has produced several data integration prototypes (among them we cite Disco [32] and Garlic [20]), some of which have been transferred into industrial products (respectively, the KelKoo comparative shopping portal [39] and the IBM DB2 product suite). Such systems could solve some of the problems faced by the sociologist seeking to exploit diverse data sources. However, they have several disadvantages. First, writing wrappers is generally a very tedious and error-prone task; in contrast, in our work, we take advantage of the acceptance of standard Web data formats (in particular XML) to automate as much as possible the process of data extraction. Secondly, they are oriented mainly on relational data, while the very nature of Web documents requires handling XML. Thirdly, such systems are typically *query-intensive*, that is, they focus on the efficient execution of complex queries on large distributed data sources, and give little or no support for the *modeling* stage, where the concepts relevant to the sociological study are defined; this stage is crucial in order to make the data analysis queries relevant for the social study.

2.3 XML query languages and databases

To manipulate XML documents, the XPath [43] and XSL [48] languages have been widely used in the context of application development, where the focus is on fast extraction and transformation of relatively small volumes of data.

The first query language proposed for XML in the database community was XML-QL [14]. XML-QL was based on two simple concepts: a pattern had to be matched against the XML document, and, for each match encountered, a certain output had to be constructed. Closer in spirit to object-oriented query languages is the X-OQL language [1]. In parallel with X-OQL, the XQL query language [46] is independently proposed. Finally, the Quilt language [11] forms the basis of the current W3C standardized XML query language, XQuery [47]. Kweelt [26] is the earliest complete implementation of the language (with respect to the language's features at that time).

All these XML query languages have become inputs to the W3C XML query standardization; thus, the XQuery language [47] borrows from each of them. An excellent resource for understanding XQuery is [12]. We rely on XQuery in our work for its expressive power, which allows complex XML manipulations, and for its increasing acceptance as a standard.

Several prototypes and industrial products currently allow to store XML documents and query them using XQuery; a comprehensive list is provided on the page of the W3C working group on XQuery [45]. In this work, we do not attempt to build an XML data management system; rather, we show how such a system could be integrated in a generic tool architecture, to be used by social scientists analyzing Web documents. Any XQuery-compliant management system fit into this architecture.

2.4 Sociology and database use

Modern sociology was born at the end of the 19th Century dealing with large amounts of statistical data. For Emile Durkheim, one of the founding fathers of sociology, the use of statistics was necessary to establish sociology as a science. For him the sociological phenomenon should be studied as an *objective reality* not as an *abstract idea*. This stood in opposition to the prevalent approach at that time, which favored ideology over experimentation. Durkheim defines a social fact as *"a way of acting, thinking and feeling, external to the individual, endowed with a coercive power by which it imposes itself upon him"* [16]. For instance, religious practices, observance of civil duties, or even suicide are social facts. When individuals collaborate in a standardization process which will determine the use of Internet for millions of people, they are not merely writing specifications, but are also participating in a well determined interaction process governed by its own set of rules.

In order to undertake a scientific analysis of social facts, one must set aside all their preconceived ideas, such as moral judgements or common sense truths. Statistical reasoning should be used to distinguish personal experience from global tendencies. Scientifically establishing a social fact means verifying it statistically. For instance, women go to church more often than men, the suicide rate is higher in the cities than in the countryside. Co-occurrence of two social facts likely means that there is an explanation connecting one to the other. For instance, Durkheim has shown that suicide was not a personal or psychological question, but a sociological phenomenon connected to global economic trends and social integration [17].

Since the 19th century, Durkheim's methodology has been well developed. The latest important issues are certainly *factorial analysis* and *network sociology*. Developed since 1960, these approaches deal with various individual data in order to exhibit relationship structures [5], [9] (network sociology), social properties in a specific social configuration, or to build the typology of a social group (factorial analysis) [21]. On the one hand, factorial analysis crosses large amounts of personal information (sex, profession, education, salary) [4] and [7] in order to build the typology of the social group (used to examine

the structure of opinion). On the other hand, the sociological network approach is very helpful in discovering relations among large numbers of people.

Nevertheless these approaches face three limitations. The first one is *conceptual*. These methods deal with society taken as a whole. Therefore it is difficult to retrieve individual behavior through them. The second one is *technical*. To be processed, the data needs to be quantified. Thus, qualitative information such as personal views on actual facts is not accounted for. These methods are able to highlight global dynamics, but can not describe the individuals' impact on them. The third limitation is that these methods do not capture the *time* dimension. Their only possible representation of a social fact is a snapshot. It is possible to compare an analysis at time t_1 with another at time t_2, but it is very difficult to explain the temporal evolution.

In contrast, our analysis of Web documents capturing actor interactions allows both to identify individuals and particular topics, and to practice a global analysis. The analysis of written messages allows a qualitative interpretation, besides the quantitative interpretation enabled by the usage of database methods. Finally, the analysis of mail threads, naturally ordered by the time dimension, allows us to highlight time-based evolution of the actors' interactions.

Marc Smith [23] has studied the way that people act on news groups, viewing these as a new interaction and communication medium. His work involves the development of a tool named NetScan, representing the participation of people in discussion, and extracting this data from the news group Web page. However, his analysis has a different focus than ours: he is mainly interested in filtering trustworthy information and identifying spammers, while we study a different problem, namely, IT standardization. Most importantly, our approach is based on XML, a semi-structured data model, which is much more appropriate for our target application (Section 5.2).

2.5 Sociological studies on standardization

One could ask why social sciences are interested by the standardization process. Standard negotiations are very technical and reserved to a small number of experts. Nevertheless standards do not only deal with technical issues. 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[1], many standards dominating the IT market, are not the best from the technical point of view [24] 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. The works described in [27] and [28] 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 [10], [30] and [31]. However, the IT standardization processes remain vastly unexplored [22]:

> *Slightly more than 2% of the published* [social sciences] *journal articles in the Information Systems field have dealt with standards over the past 10 years, and most of this work has reported on newly established IT standards rather than examining the processes and importance of standard setting processes. Notably absent are studies that analyze different standardization concepts, standardization processes, industrial coordination and strategy, and economics of standards.*

Our study is an attempt to fill this gap, by exposing the actors and the mechanisms within the W3C standardization of XQuery.

3. TARGET APPLICATION

In this section, we introduce all the concepts necessary in order to comprehend the final goals of the study, but first we start by motivating our interest on the standardization topic, both from the sociological and computer science point of view.

3.1 Why study the W3C?

Internet is everywhere. This is, in fact, its defining characteristic: to be turned towards the world, so that all those with a computer and a means of connecting may discuss and communicate, and the World Wide Web Consortium is in the center of this network. Around 90% of the W3C Recommendations can be seen as *standards* in so far as they clearly define i.e., standardize various aspects of Internet life. However, the whole process of standardization is somewhat vague. Even the people in the center of the process, although able to catch a glimpse of the overall picture, need a way to comprehend the *way it all works* [18], and this is what a thorough and methodological study of the W3C standardization process can offer. But let us start by defining our objects. Each W3C specification is the product of a group of individuals called *working group*.

We have chosen to study more specifically the XQuery Working Group, for several reasons. First of all, in any sociological study, it is important to have "experts" of the field that can analyze the sometimes technical issues and serve as interpreters and referees. The author's acquaintances with XQuery pointed this topic out as relevant. The second, and more profound reason, is that the XQuery working draft is now very close to a recommendation status [47]. It is therefore possible for us to view the working group's results, and initial objectives, and those of individuals. By analyzing this working group, from its beginning to its achievement, we are able to cover all the situations that could arise in such a context.

3.2 The XQuery standardization scene

One may imagine that standards are drawn up in closed rooms full of boring people talking about vapid topics. Is this how it really takes place? Where are the discussion rooms packed with *standardizers*? Should we take a trip to the M.I.T. in order to meet with all these experts, and spend countless hours interviewing them, in order to analyze the standardization process?

In fact, these discussions only rarely (once or twice a year) take the form of face to face meetings. The rest of the time, discussions are held via *email*. *Teleconferences* are also

[1] Organisation for Economic Co-operation and Development. http://www.oecd.org

organized, but most of the time, they are to settle issues already dealt with on the mailing list. Just like live discussions, some emails are private, and are withheld to working group 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 [47]. At this site, we will find not only all the participants, but also all the public statements and reactions that they will have had, during these last four years.

Our study focuses on the public emails posted on the XQuery comment mailing list, during the last four years: about 5.000 emails that can be divided into threads, by determining which emails answer each other. Our goal is to build a *semi-structured data warehouse* model in order to store and process (by using XQuery!) this information corpus.

3.3 Sociological approach of the W3C

The social study of the standardization process must answer seemingly simple questions: Who are the *individuals* involved? What are their *motivations*? What *relationships* do they have between each other? What *role* do they play in their organizations?

Answering these questions should draw a preliminary sociological map of the standardization process. The purpose of this approach is to 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. For instance, the interaction between two individuals can be inferred by the content of the mailing list, and the roles played by the individuals within their organizations (corporations or academia) can be extracted from the W3C web pages. A corporation's goal might be to reuse existing technology to implement a new standard. These questions determine the conceptual structure of the database to be set up.

4. THE METHODOLOGY

This section describes our methodology for the social analysis of the W3C public comments mailing list. We present our approach for data modeling and extraction (Section 4.1), storage and querying (Section 4.2), filtering and enrichment (Section 4.3). After these data-oriented methods and tools used in our project, we outline complementary sociological tools in Section 4.4.

4.1 As automatic as possible!

In order to exploit very large quantities of information, it is critical to design a methodology which needs as little human intervention as possible. However, human input and feedback can (and should) be used to tune and enrich the system.

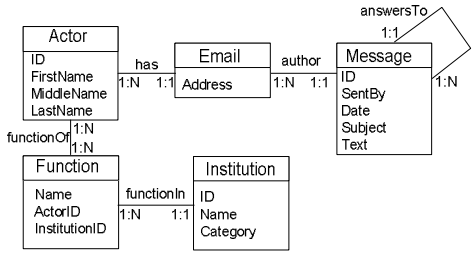

Figure 1: Conceptual model for the social analysis.

In our case, the process starts with a conceptual modelling of the entities of interest to the sociological study, depicted in a standard Entity-Relationship diagram in Figure 1. We are interested in identifying actors: the individuals that post messages on the mailing list. Each author has an unique ID, and first, middle and last name. An actor can 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. This model is the starting point of our analysis; the result of the social scientist's queries may lead to identifying other interesting entities and relationships (we exemplify this in Section 5.2). Once the conceptual model is established, we *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, home pages of the more important actors, institutions websites), using wrappers.

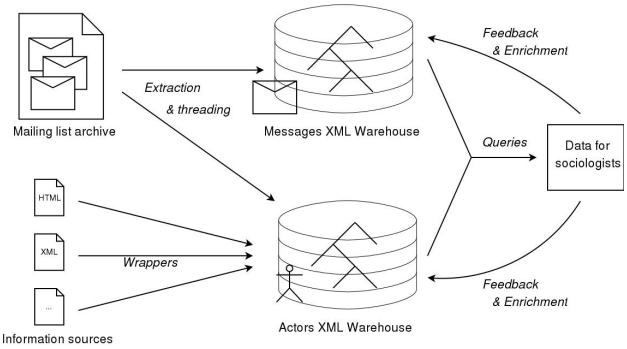

Figure 2: Outline of our warehouse construction process.

The content of these two warehouses is the "raw" data; we may enrich it with extra information on actors and institutions. This information is gathered by fetching information from the Web, by wrapping specific HTML pages into XML. Let us stress that this procedure is semi-automatic. The system generates propositions, and these need to be manually confirmed.

4.2 XML storage and querying

We have chosen to represent our content warehouse in XML, for the following reasons:

- 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 *flexible*: new information can be added at will by adding new elements or attributes
- 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.
- 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.
- XML remains *simple* to understand. Sociologists can grasp the extent and the kind of the information stored in the warehouses just by reading the XML data.

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 (see Section 5.2), and its declarative nature makes it much easier to use than alternative languages such as XSLT.

4.3 Data filtering and enrichment

A number of practical issues required the use of automatic and semi-automatic filtering and enrichment:

- The name of an institution may appear written in many different ways. We devised generic patterns to recognize that "Sun Microsystems Inc." is the same institution as "Sun" and "Massachusetts Institute of Technology (MIT)" is the same as "MIT". Further identification can be done manually (e.g. to understand that "cerisent.com" is a Web site belonging to "Mark Logic Corporation").
- A person can have several different e-mail addresses; identification of persons must thus be done on their names, and not on their addresses (assuming that two persons do not have the same name). First names, optional middle name and last name are extracted from the full name description, making possible the assimilation of "FirstName LastName" and "LastName, FirstName" for instance.
- There can be "holes" in the thread structure of the mailing list (lost e-mails, corrupted mail headers...) for which fake messages must be added.

4.4 Complementary sociological tools

The data automatically collected will be complemented by more qualitative data sources, such as interviews, technical documents, outsiders' surveys, in order to understand how individual strategies participate to the larger picture. Individuals involved in the standardization process do not only represent the viewpoint of their institutions; they also have their own expertise and beliefs which impact their judgment and statements. Besides, personal relationships (both friendship and animosity) influence the decision making process. Clearly, such relationships are not recorded in the mailing list. Finally, the identification of high status individuals with recognized authority on the technical subject is not straightforward. This status is often implicit, and although it is known by the specialists knowledgeable in the field, it may remain hidden to an outsider.

In general, the hierarchy structuring the standardization social space is confirmed by the qualitative analysis of the actors' perception.

5. EXPERIMENTATION

5.1 Data acquisition

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 answering to e-mails to public-qt-comments-request@w3.org and then converted to a Unix *mbox* format. The mailing list contained 5626 messages at the time of extraction.

A Perl script was written to convert this archive into the XML Data Model described in Section 4.1. This script uses the Perl **Mail::Thread** [49] module to build the thread structure, based on Jamie Zawinski's threading algorithm [53]. Actors and institutions information were also generated by this script. Some actors were not extracted, when it was impossible to extract automatically a full name from the **"From:"** field. This was the case, for instance, of a number of spams.

Wrappers for HTML and XHTML web pages (only available to W3C members) describing the list of members of the XQuery W3C working group were also written, in order to add information about membership in the XQuery working group to the actors warehouse.

5.2 Model, queries and results

Once the data is extracted according to the conceptual data model into the messages, respectively, actor XML warehouses, we get a first level of information on this data by applying our XSum graphical tool [52] on the XML documents. For instance, the graphical representation derived by XSum from the actor warehouse is depicted in Figure 3.

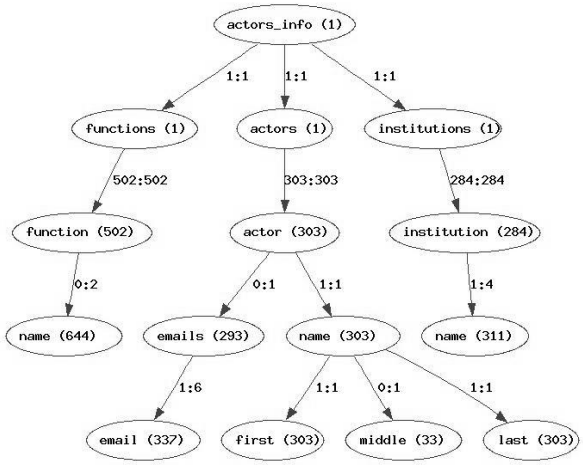

Figure 3: Graphical summary of the actors warehouse.

XSum [52] computes a structural summary of the XML document, and a set of simple statistics. For instance, in Figure 3, there are 284 institution elements, each of which has at least 1, and at most 4 names, as indicated by the "1:4" label on the edge between the institution node and its child labeled name.

We now describe our data analysis process. We issued a set of XQuery queries, which were processed by the QizX system [25]. Each query brings information that can be used to validate existing hypothesis and formulate new ones. The sociological interpretation of the results appears in Section 5.3.

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, the W3C etc.[2], ***prov*** for providers of Internet access and email (due to the extraction procedure, we collected quite a lot of such providers, simply 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, allowing us to use it for further interrogation. Then, the number of institutions for each type was computed.

The natural question would then be to compute the number of messages per actor belonging to each type of institutions. However, this is not really useful because an actor may belong to many institutions, potentially of several types. Moreover, ***prov***, ***pers*** and ***unknown*** organizations were not considered interesting for the sociological analysis.

Therefore, we refined our categorization by devising a set of interesting **profiles**, where each profile consists of 1, 2, or more types of institutions. Query 1 (below) computes *for each actor the profile derived from the institutions he belongs to*:

```
let $actors:=doc("actors_info.xml")/actors_info/actors
let $institutions:=doc("actors_info.xml")/actors_info/institutions
let $functions:=doc("actors_info.xml")/actors_info/functions
let $ac:=element actor_categories{
 for $a in $actors/*
 let $c:=distinct-values(
   for $f in $functions/*[@actor_id=$a/@id and @institution_id !='xquery']
   let $i:=$institutions/*[@id=$f/@institution_id]
   where $i/@category='org' or $i/@category='corp' or $i/@category='univ'
   return $i/@category)
 return element actor {
   $a/@id,
   attribute category {string-join((for $t in $c order by $t return string($t)),'_')}
 }
}
return element actor_categories {
 for $t in distinct-values($ac/*/@category)
 let $a:=( for $actor in $ac/*[@category=$t]
   return <actor>{$actor/@id}</actor>)
 order by $t
 return <category cat="{$t}">{$a}</category>
}
```

Query 1: Actors regrouped by institution categories

Based on this categorization, *the number of messages issued by actors of each profile* is obtained by applying Query 2 (below):

```
let $threads:=doc("public-qt/public-qt.xml")/threads
let $actors:=doc("actors_info.xml")/actors_info/actors
let $ac:=doc("actor_categories.xml")/actor_categories
return element messages_count {
 for $t in $ac/category
 let $c:=sum(
   for $ac in $t/actor/@id
   let $a:=$actors/actor[@id=$ac]
   return count($threads//message[@author=$a/emails/email])
 )
 return <count category="{$t/@cat}">{$c}</count>
}
```

Query 2: Number of messages for actors of each institution categories

Other queries we performed concern the depth of threads, and their length in messages by. The repartition of threads by their length is shown in Figure 4.

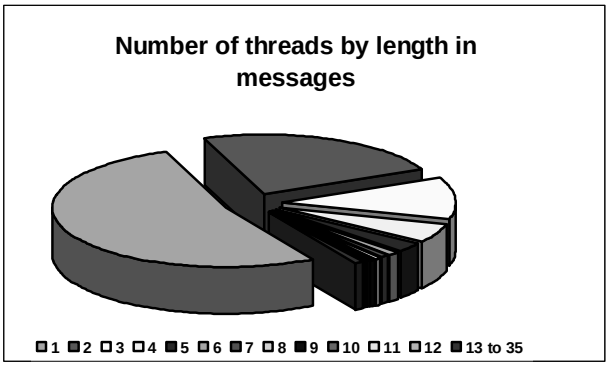

Figure 4: Thread distribution by length.

We notice that a large number of threads consist of just one message (no answer). By inspecting the messages themselves and interviewing participants of the Work group, we conclude that these messages are split in four roughly equal classes:

1. Spam (messages completely unrelated to the topic)
2. Messages left without an answer.
3. Messages answered, but where the answer was posted as a new message, due to the user's choice.
4. Messages answered, but where the answer appears as a new message due to mailer (mis-)configuration.

The presence of "real" no-answer messages (class 2) can be explained as follows:

a. Most of these messages refer to Last Call Working Draft i.e. result of public review process; they are taken as input by the WG and do not need to be explicitly answered.

b. Some of them are posted from participants in the Work Group itself, and the discussion may continue on the private mailing list (thus, outside the scope of our analysis).

[2] We have eliminated the XQuery working group itself from the "organization" category. Keeping it would skew our analysis, since it would artificially boast the organizations' participation in the mailing list.

c. In some cases, the WG may have decided decided not to reply to all the comments – especially editorial ones – individually.[3]

Among the longer threads, we notice that 2 and 3-message threads are quite important; thread length may reach 35, but from 13 to 35 messages, there is just one thread in the respective category. This corresponds to the fact that the mailing list is for *comments*, which often require punctual answers, and rarely yield lengthy exchanges.

We have also followed the evolution, in time, of the e-mail addresses of a few actors. These queries, and all other documents and queries we used, can be found at [50].

5.3 Sociological interpretation

The first step of sociological analysis consists of identifying the institutions involved in the messages on the public mailing list.

An analysis of e-mail addresses of users posting on the mailing list shows that most of them come from IT **companies**: 37%. The next most represented group contains e-mail addresses hosted at various Internet providers, which do not give us useful information about the organizations of their owners: 34%. **University** e-mail addresses represent 16%; independent (freelance) IT **professionals** make up 7%, while 6% of e-mails come from non-profit **organizations**.

We have analyzed the number of actors, and the messages they issued, for a group of seven profiles. We obtained the following distribution:

Profile	# actors	#posted mssgs.
Companies	135	2689
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**

From this distribution, we see that from about 300 people having posted at least one mail on the list, 160 are connected to at least a company. These first results show that companies are dominating the institutional landscape. Their extremely visible domination in terms of actors present in the mailing list highlights the interest of such companies in the W3C standardization process. 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 second step of our analysis is to verify if the large participation of company employees is reflected in the volume of their postings in the mailing list threads. Thus, we need to analyze the message thread structure and its cutout with respect to the various user profiles.

When analyzing the distribution of mails posted by users for each interesting profile, we see that the commercial companies' domination is confirmed. From the 5299 mails we analyzed, 4273 (81%) come from people connected to at least one company. On the contrary, academics (individuals connected with universities but not with companies) have a low participation rating: 3 messages on average posted by an actor with "University" profile, and 6 for the "University and Organization" profile, which 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, than individuals connected to companies? Are they more present in private list? Further analysis on the private list would answer this question.

Another interesting observation is that the most active participants to the mailing list have a mixed profile, which furthermore includes a company affiliation. The most active participants are those whose profile is "Companies & Universities" (177 postings per person) and those with profile "Companies and Organizations" (48 postings per person). These first results confirm the observations made in a previous study on the international regulation process [15]: the most active, and often most influential actors in regulation/standardization processes are those belonging 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.

Nevertheless, we noticed two kinds of mixed profile individuals: those having several affiliations *at the same time* and those affiliated to successive institutions *along the time*. For instance, one actor is connected to two very large software corporations C_1 and C_2, and one university U. An XQuery query on the mailing list archive can extract all addresses associated to this actor, in the order of the e-mails, thus, *in time order*. Comparing successive addresses allows detecting address changes. In the case of our sample actor, we detected exactly two address changes, from C_1 to U (using an alumnus e-mail address) and a few months later from U to C_2. We can conclude that this actor has successively been connected (probably as an employee) to two very important companies, and he is or was connected with an important IT-oriented university. More frequently, we noticed the other kind of mixed profile individuals, simultaneously affiliated to two different types of organizations.

Further queries development would make systematic distinction between these two kind of *key actors*. One should note, that what is important is not necessary simultaneous affiliation, but also *social networks developed* during successive affiliations.

This example is already a very interesting output for sociologist: it permits to identify one potential important actor for deeper research.

Qualitative analysis (interviews and text analysis) should continue our study, in order to verify the real influence of *key actors* in the mailing list. Further analysis on the message thread structure (who asks questions and who is answering them) will also bring more information about the social configuration of the

[3] The authors are grateful to Liam Quin (W3C liaison for the XQuery working group) for his explanations of the messages left without an answer.

list. Finally, we expect that this methodology could highlight other kinds of information, such as actors' functions, companies' interests and actors' networks (the personal and professional connections between them).

6. ARCHITECTURE OF A GENERIC TOOL FOR THE SOCIAL SCIENTIST

From our joint work in this project, we have learned a set of lessons which we crystallize in a generic architecture for a social scientist's data management tool. This architecture is outlined in Figure 5.

A necessary input to the tool is the set of concepts that are relevant to the social analysis; these concepts (e.g. actors, institutions, messages etc.) are established by the sociologists prior to actual data acquisition. Ideally, these should be described via a convivial interface as in [3]; they can be then serialized in an XML format (transparent to the scientist) and stored in the XML database.

The scientist must then single out Web data sources to load in the tool; these are extracted and loaded via wrapper (Web data acquisition) modules, shown at the bottom of Figure 5. The data may require cleaning and filtering, as we have seen in Section 4.3. Thus, the tool must be easy to extend with new filtering capabilities, perhaps under the form of user-defined functions.

Once the data is loaded, the sociologists will start by browsing/getting acquainted to this data; we illustrated the use of XSum [52], and there are many others, such as XMLSpy [51].

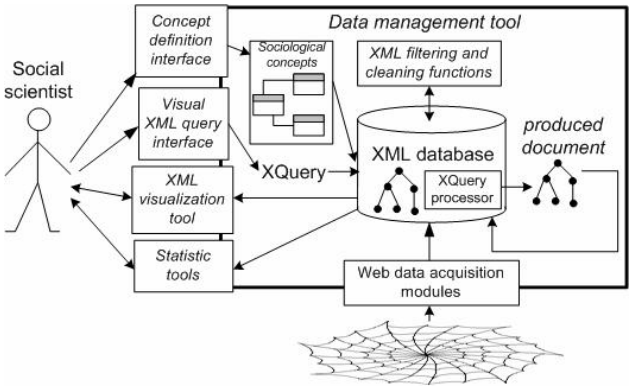

Figure 5: Generic architecture for a social scientist's data management tool.

Based on this knowledge of the sources, the scientist can start querying them. XQuery queries in general, and in particular those used in our analysis, tend to be quite complex. A way out of this complexity is to formulate *one simple query at a time*, *materialize* the XML document it produces, *visualize* it, formulate another query etc. The author of this paper with a social sciences background found it much easier to issue XQueries based on several intermediary results. Furthermore, the very nature of his work is exploratory and multi-stage: sociologists formulate hypotheses, which they attempt verify by querying the database. Positive (or negative) results naturally lead him to formulate other hypotheses; for this reason, the XML database must make it easy to materialize the result of a query as a refined document, to be added to the database for further querying. Simple XQuery queries can be formulated with the help of a graphical tool such as XQBE [8].

Finally, sociological analysis requires statistical tools, such as spreadsheets, and graph production tools, in order to judge of the statistical correlation between two facts etc. Such tools are usually included as utilities in various operating systems (e.g. Office on Windows and Open Office on Linux); interfacing them smoothly with the data management tool would be very helpful. We stress that such tools are used by today's sociology research, but in our case, the very production of the numbers on which they apply requires the previous usage of semi-structured (that is, XML) databases.

7. CONCLUSION AND PERSPECTIVES

The study presented here is conducted primarily based on publicly available information on the XQuery standardization process: HTML Web pages and publicly-archived mailing lists.

Obviously, the interpretation and analysis could be improved by using the private mailing list of the XQuery working group [44], on which the W3C discussions on the standard itself take place. However, we feel confident that the results obtained by using public information are both interesting and respectful to privacy policies. We have contacted the W3C soliciting the right to include in our analysis the messages from this list, and are currently waiting for the response.

More generally, the issue of credentials to access a set of Web documents (including archived message boards) is orthogonal to our approach: our purpose is to support, and provide tools for social analysis on the basis of a given Web dataset.

In the near future, we plan to include in our analysis: the e-mails from the private XQuery mailing list (about 10.000 messages), the W3C Web pages listing the previous and current XQuery working group members, as well as the HTML-based meeting and teleconference notes, posted on the W3C site. Access to these documents is restricted to W3C members; getting the permission to use them will enable us to analyze the interactions between W3C working group members which shaped the definition of the language.

Applying our tool to other W3C working group sites is technically speaking straightforward, and the sociological method could be the same. However, potential user communities, issues, players, and interactions patterns may be very different for, say, the Web Content Accessibility Working group [42], which focuses on making the Web accessible to all regardless of disabilities, and the Math ML working group [41] which is dedicated to the inclusion of mathematical expressions in Web pages, and thus very domain-specific. Thus, a social analysis of these different groups is likely to highlight different types of actors, interactions etc.

More generally, the tool we have developed could be used to acquire and organize the data contained in any mailing list archived on the Web, allowing thus to analyze the interactions taking place between the participants. This includes e.g. the Linux kernel mailing list, archived at [40], mailing lists for various Linux distributions (e.g. Debian and RedHat), for the IEEE standardization working groups [38] etc. Other foreseeable applications may include the analysis of exchanges within a given corporation or one of its department/taskforce, scientific

mailing lists such as DBWorld [36] for database research and many others.

8. BIBLIOGRAPHY

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

[2] Aguilera V., Boiscuvier F., Cluet S., and Koechlin B., *Pattern tree matching for XML queries*. Gemo Technical Report number 211, 2002. Available at http://www-rocq.inria.fr/gemo/Publication

[3] Abiteboul S., Cobena G., Nguyen B., Poggi A., *Sets of Pages of Interest*. In Bases de Données Avancées, 2002

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

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

[6] Besen S.M. and Farrell J., *The Role of the ITU in Standardisation*. Telecommunications Policy, 15 (4), 1991, 311-321.

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

[8] Braga D., Campi A., Ceri S. *XQBE: A Graphical Interface for XQuery Engines*. EDBT Conference 2004: 848-850

[9] Breiger R. L., Boorman S. A. and Arabie P., *An Algorithm for Clustering Relational Data with Application to social Network Analysis and Compariason with Multidimensional Scaling*. In Journal of Mathematical Psychology, 12, 1975

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

[11] Chamberlin D., Robie J., and Florescu D., *Quilt: An XML query language for heterogeneous data sources*. In Proceedings of the International WebDB workshop, Houston, USA, 2000

[12] Chamberlin D.. *XQuery: A query language for XML*. In SIGMOD, page 682, 2003. Slides available at http://www.almaden.ibm.com/cs/people/chamberlin

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

[14] Deutsch A., Fernandez M., Florescu D., Levy A., and Suciu D., *A query language for XML*. In Proc. of the Int. WWW Conf., volume 31(11-16), pages 1155-1169, 1999.

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

[16] Durkheim E., *Les règles de la méthode sociologioque*. Flammarion, 1988. p. 97.

[17] Durkheim E., *Le Suicide*. Presses Universitaires de France, 1997.

[18] Fernandez M., *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), 2004.

[19] Vaisman A.A., *OLAP, Data Warehousing, and Materialized Views: A Survey*. Available at : citeseer.nj.nec.com/vaisman98olap.html

[20] Haas L., Kossmann D., Wimmers E., Yang J., *Optimizing Queries Across Diverse Data Sources*. VLDB 1997: 276-285

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

[22] *Special Issue on Standard Making: A Critical Research Frontier for Information Systems: Pre-Conference Workshop International Conference on Information Systems*, TIC Meeting on Information System Quarterly, Seattle, Washington, December 12-14, 2003

[23] *Interview with Marc Smith, Microsoft Corp*. Available at http://www.microsoft.com/presspass/events/svspeaker/07-29MSmith.asp

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

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

[26] Sahuguet A.. *The Kweelt system*. Available at http://sourceforge.net/projects/kweelt.

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

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

[29] Sochor E. *The politics of international aviation*, Basingstoke; London, Mac Millan, 1991.

[30] Tamm-Hallström K., *In Quest of Authority and Power: Standardization Organizations at Work*. Scancor Workshop : Transnational regulation and the transformation of states, California, USA, 22-23 June 2001

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

[32] Tomasic A., Raschid L, Valduriez P., *Scaling Heterogeneous Databases and the Design of Disco*. ICDCS 1996: 449-457

[33] Widom J., *Research problems in Data Warehousing*. International Conference on Information and Knowledge Management, 1995

[34] Wiederhold G., *Mediators in the Architecture of Future Information Systems*. IEEE Computer 25(3): 38-49 (1992)

[35] Active XML reference: http://www.axml.net/

[36] The DBWorld mailing list. Available at http://www.cs.wisc.edu/dbworld

[37] Projet e.dot : http://www-rocq.inria.fr/gemo/Projects/edot/

[38] IEEE Standardization Working Groups Areas. Available at http://grouper.ieee.org/groups/index.html.

[39] The KelKoo comparative shopping engine. Available at http://www.kelkoo.com

[40] The Linux Kernel mailing list archive. Available at http://www.uwsg.indiana.edu/hypermail/linux/kernel

[41] The W3C Math Home Page. Available at http://www.w3.org/Math.

[42] The Web Content Accessibility Guidelines Working Group. Available at http://www.w3.org/WAI/GL.

[43] XML Path Language. Available at http://www.w3.org/TR/xpath.

[44] The W3C XQuery mailing list (access restricted to W3C members). Available at http://lists.w3.org/Archives/Member/w3c-xml-query-wg.

[45] XQuery products and prototypes. Available at http://www.w3.org/XML/Query#Products.

[46] The XQL query language. Available at http://www.w3.org/TandS/QL/QL98/pp/xql.html.

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

[48] The Extensible Stylesheet Language Family. Available at http://www.w3.org/Style/XSL

[49] Available at http://search.cpan.org/~simon/Mail-Thread

[50] Action Concertée Incitative *Normes Pratiques et Régulations des Politiques Publiques* . Available at : http://www-rocq.inria.fr/gemo/Gemo/Projects/npp/index

[51] *XML Spy*. Available at: www.altova.com.

[52] *XSum: the XML Summary Drawer*. Available at http://www-rocq.inria.fr/gemo/Gemo/Projects/SUMMARY

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