Webstand, une plateforme de gestion de données web pour applications sociologiques

Benjamin Nguyen, Antoine Vion, François-Xavier Dudouet, Dario Colazzo, Ioana Manolescu

WebStand, une plateforme de gestion de données Web pour applications sociologiques

Benjamin Nguyen* — Antoine Vion — François-Xavier Dudouet*** — Dario Colazzo**** — Ioana Manolescu*******

** Laboratoire Parallélisme, Réseaux, Systèmes et Modélisation (PRiSM)*
Université de Versailles Saint-Quentin, CNR UMR 8144

benjamin.nguyen@prism.uvsq.fr

*** Laboratoire d'Economie et Sociologie du Travail (LEST)*
Université d'Aix-Marseille II, CNRS UMR 6123

antoine.vion@univmed.fr

**** Institut de Recherches Interdisciplinaire de Sociologie, Economie, Science Politique (IRISES)*
Université de Paris Dauphine, CNRS UMR 7170

dudouet@dauphine.fr

***** Laboratoire de Recherches en Informatique (LRI)*
Université de Paris XI, CNRS UMR 8623

dario.colazzo@lri.fr

****** Institut National de Recherches en Informatique et Automatique*
Projet GEMO

ioana.manolescu@inria.fr

RÉSUMÉ. *WebStand est un projet ANR pluri-disciplinaire débuté en 2006, regroupant des informaticiens spécialistes de bases de données et d'internet et des sociologues spécialistes dans l'étude des processus de normalisation et des nouvelles technologies. Dans le cadre du projet a été réalisée une plate-forme permettant certaines analyses sociales de personnes postant sur des listes des discussions. Nous décrivons l'exemple des listes de discussions techniques du W3C dans le cadre de la normalisation de XQuery.*

ABSTRACT. *WebStand is a multidisciplinary ANR project that started in 2006, involving database computer scientists and sociologists specializing in the study of the standardization of new technologies. Its main results are a platform the enables the social analysis of people posting on mailing lists. We illustrate our results by applying it to the W3C public mailing lists, discussion groups on specific technical issues, in the context of XQuery.*

MOTS-CLÉS : *Normalisation, W3C, XQuery, plateforme sociologique.*

KEYWORDS: *Standardization, W3C, XQuery, sociological platform.*

1. Introduction

Etudier les expressions publiques sur le web est devenu un enjeu d'études important pour les sciences sociales, qui restent pourtant assez mal outillées du point de vue informatique. Le projet Webstand, dont nous présentons ici les résultats, a visé à répondre à un certain nombre de problèmes méthodologiques posés dans le traitement des données sociologiques, afin de faire progresser ces méthodes d'enquêtes.

1.1. Un domaine d'analyse en plein développement : étudier les communications sociales sur le web

L'étude de la communication sur le web progresse de jour en jour. Certains travaux se focalisent sur les nouvelles formes de socialisation et de civilité qui s'y développent (Beaudouin, *et al.*, 1999, Papacharissi, 2004, Auray *et al.*, 2007). De son côté, l'étude des mouvements sociaux essaye de suivre les nouvelles formes de mobilisation sociale (Diani, 2001 ; Pickerill, 2001 ; Van Aelst *et al.*, 2004 ; Juris, 2004 ; Della Porta *et al.*, 2005 ; Calderaro, 2007). De même, la science politique s'intéresse de plus en plus aux pratiques de campagne en ligne et de blogs politiques (Hacker *et al*., 1996 ; Drezner *et al.*, 2004 ; Adamic *et al.*, 2005 ; King *et al.*, 2007, etc.), ou aux enjeux démocratiques de la délibération sur le web (Hague *et al.*, 1999 ; Trechsel *et al.*, 2003 ; Pavan *et al.*, 2008).

Malgré l'intérêt de tous ces travaux pour les sciences sociales, force est de constater que l'outillage des chercheurs ne leur permet généralement pas de surmonter un certain nombre de difficultés inhérentes à la structure des données web. Premièrement, l'hétérogénéité des données et des formats (HTML, PDF, JPEG, TXT, Word, etc.) rend difficile l'usage des méthodes traditionnelles de management des bases de données relationnelles (Microsoft Access, IBM DB2, Oracle), qui imposent des logiques de copier-coller coûteuses et difficiles à appliquer à des grands volumes. Deuxièmement, si ces méthodes traditionnelles peuvent s'appliquer pour effectuer des traitements statistiques probabilistes sur la base de sondages (King *et al.*, 2007), elles deviennent beaucoup plus handicapantes lorsqu'il s'agit d'entreposer un corpus exhaustif de données temporalisées (comme l'activité sur plusieurs années d'une mailing-list, par exemple). Troisièmement, la lourdeur des opérations de stockage peut provoquer un défaut de contrôle des modifications successives, dont même les plus marginales peuvent affecter sévèrement le sens des messages et les relations entre les objets – comme c'est le cas pour les pages wiki, par exemple (Chateauraynaud, 2007).

Pour toutes ces raisons, il nous a semblé utile d'approfondir les collaborations entre informaticiens et sociologues, afin d'outiller les investigations sociologiques et de progresser en informatique sur la modélisation des architectures de bases de données, notamment sur les schémas de mailing-lists.

1.2. Webstand : une collaboration originale entre informaticiens et sociologues

La collaboration entre certains membres informaticiens et sociologues de notre équipe remonte à un travail sur le contrôle international des drogues et la réalisation d'une base de données relationnelle sur les personnes impliquées dans ces arènes (Dudouet, 2002). A l'époque, les données étaient issues d'archives papier de plusieurs sources différentes (SDN, ONU, etc.) des années 20 aux années 90. Nous avons pu mesurer la difficulté d'une part pour réussir à stocker ces informations dans un format unique, un problème de médiation en somme, ainsi que la relative non adéquation de SQL pour interroger des données parcellaires ou incomplètes, de structure hétérogène. Nous avons donc changé d'approche, et adopté une représentation XML, dans le projet que nous décrivons dans cet article, *WebStand*.

WebStand est un projet ANR Jeunes Chercheurs, débuté en 2006 pluridisciplinaire, regroupant des informaticiens spécialistes de bases de données et d'internet avec des sociologues spécialistes dans l'étude des processus de normalisation et des nouvelles technologies, possédant deux objectifs principaux : a) la réalisation d'une plateforme prototype basée sur un entrepôt de données XML pour la gestion de données récupérées sur le web dans le cadre d'une application sociologique, b) l'étude, à l'aide de cette plateforme du cas du World Wide Web Consortium (W3C). Notre plateforme possède plusieurs modules tels que l'extraction d'informations sur internet, le nettoyage de données. Son intérêt principal est de permettre le traitement de listes de discussions d'emails, ou de forums sur lesquels s'expriment les normalisateurs. Notre plateforme a vocation à être utilisée pour des applications sociologiques plus larges, mettant en jeu des personnes s'exprimant sur le web (blogs ou autres).

1.3. Spécificité des données sociologiques : entretiens et archives

Au cours de notre projet, nous avons eu à traiter essentiellement deux classes de données : des données issues d'entretiens avec des « humains », et des données issues d'archives (papier ou informatiques). Nous nous intéressons principalement aux documents informatiques trouvés sur le web, même si notre application permet la gestion de données manuelles, par l'édition de la base de données.

Les données sociologiques peuvent être quantitatives (par exemple l'âge d'une personne) et dans ce cas un traitement mathématique simple peut être appliqué, mais la plupart du temps, et en particulier lorsqu'il s'agit de données récupérées sur le web, elles sont incertaines, inexactes, ou encore floues : « *ça s'est produit courant 2008…* ». Ces caractéristiques nous ont conduit au choix d'une base de données XML, et subséquemment XQuery comme langage d'interrogation pour notre plateforme. L'un des premiers problèmes de notre collaboration a donc été de valider l'utilisation du meta-modèle semi-structuré pour la réalisation du modèle de données de l'entrepôt.

Notons que la définition d'un modèle général de la personne, au sens sociologique s'apparente à la création d'une ontologie générale de la sociologie, et

dépasse largement le cadre de notre article. Nous proposons ici un modèle spécifiquement dédié à l'analyse de personnes participant à des discussions par le biais d'emails, et nous l'explicitons dans les sections 3.2. et 3.3. Ce modèle est néanmoins générique, et peut être réutilisé dans toute application gérant des auteurs de courriels. Nous travaillons actuellement à son extension à toute forme de personnalités « en-ligne » comme des bloggers, participants à des forums, membres de réseaux sociaux tels que Facebook, etc. Les travaux présentés ici montrent comment l'utilisation du modèle de données XML permet très simplement d'établir un schéma pour l'entrepôt, mais nous sommes conscients que d'autres applications sociologiques pourraient donner lieu à des modèles radicalement différents.

Le cœur de notre application est le traitement de listes de discussions sur lesquelles s'expriment les normalisateurs du W3C : les listes publiques des *Working Groups* (WG) du W3C et en particulier celle que nous allons détailler, le XQuery Working Group. Précisons que nous avons choisi ce WG pour notre bonne connaissance de son fonctionnement, et qu'ainsi nous étions à même de répondre en tant qu' « experts » à certaines questions des sociologues sur divers points de fonctionnement ou encore d'apporter des commentaires sur des débats techniques, pour indiquer par exemple si deux discussions qui semblent porter à controverse sont sur le même thème. Notons que les données stockées dans l'entrepôt, et les informations produites sont des informations du domaine public.

Le traitement des listes de discussions doit permettre de répondre aux questions suivantes : qui sont les normalisateurs, et qui sont les plus influents ? Quelles sont les stratégies des entreprises participant à ce groupe de travail ? Quel est le type de gouvernance qui régit un organisme de standardisation tel que le W3C ? Nous apportons des éléments de réponse à ces questions dans la Section 4.3 basé sur l'étude que nous avons faite sur les listes de discussion publiques. Pour obtenir les réponses exactes à ces questions, il aurait fallu traiter les listes de discussion privées, ce qui est malheureusement impossible pour des raisons de confidentialité. Notons néanmoins que la méthodologie reste identique, et que depuis quelques années, la politique du W3C est de n'utiliser plus que des listes publiques.

1.4. Plan

Dans cet article, après un bref état de l'art, nous allons présenter l'architecture de notre plateforme, puis décrire notre application sociologique : l'étude des acteurs de listes de discussion du W3C, et nous conclurons sur les perspectives de notre travail. Tout au long de cet article, nous nous efforcerons d'insister sur les spécificités du cadre socio/informatique.

2. Etat de l'art

Nous présentons ici un bref état de l'art sur les thèmes liés à notre projet, que ce soit sur le plan informatique ou sociologique.

2.1. Entrepôts de données du web

Il y a déjà eu beaucoup de travaux sur le thème des entrepôts de données (data warehousing), la médiation et l'intégration de données (Widom, 1995). Nous référencerons simplement (Chaudhuri, *et al.* 1997) et (Vaisman, 1998) pour une revue d'OLAP, les entrepôts de données et les vues matérialisées. La différence entre ces technologies et le travail du sociologue est que classiquement les entrepôts de données s'intéressent à des variables hautement quantitatives.

En effet, le cas de la construction d'un entrepôt de données sociologique se doit d'adopter une approche radicalement différente, plus proche du concept de *content warehouse* (entrepôt de contenu), introduit dans (Abiteboul, *et al.*, 2002) et (Abiteboul, 2003). Un entrepôt de contenu est un entrepôt d'informations quantitatives **et** qualitatives, comme la plupart des données sociologiques, qui ne possède pas de méthode de traitement mathématique triviale. Le fait de représenter des relations entre personnes, ou bien leur rôle, n'est pas quelque chose qui s'analyse facilement avec l'OLAP traditionnel. Cette information, obtenue de sources parfois concurrentes sur le web, est hautement hétérogène et ne peut être modélisée de manière satisfaisante qu'en utilisant un format flexible et riche comme XML. Une fois le choix du format de donnée fixé, nous nous appuyons sur une méthodologie ayant déjà fait ses preuves lors de la construction d'un entrepôt de données biologiques sur le risque alimentaire, le projet *e.dot*, méthode également poursuivie dans le cadre du projet ANR WebContent.

2.2. Intégration de données et XML

Les systèmes d'intégration de données offrent la possibilité de poser des requêtes à des sources hétérogènes et distribuées, comme s'il s'agissait d'une base de données unique. L'architecture classique d'un système d'intégration de données inclut un *médiateur* (Wiederhold, 1992) qui offre une vue intégrée à l'utilisateur, et un ensemble de *wrappers*, qui permettent la connexion entre les sources de données individuelles et le médiateur. Les travaux dans ce domaine ont produit divers prototypes, nous citons par exemple XLive (Dang Ngoc, *et al.*, 2005) car il possède l'avantage d'utiliser XML et XQuery, et de permettre de connecter des sources relationnelles ou XML. Notre objectif n'étant pas de réaliser un nouveau moteur de base de données, nous avons testé divers moteurs (eXist, Qizx, MonetDB-XQ) qui peuvent tous être simplement intégrés par un médiateur tel que XLive.

Bien que l'utilisation d'un médiateur puisse nous simplifier un peu la tâche, les systèmes de médiation sont en général axés sur l'exécution rapide de requêtes relativement complexes sur des données distribuées, mais n'aident pas forcément l'utilisateur lors de la réalisation de son modèle de données. De plus, la réalisation des wrappers est une tâche assez fastidieuse qui doit être réalisée par l'utilisateur du médiateur. Un simple médiateur ne pourra donc pas être livré 'clés en main' à un sociologue.

2.3. Langages de requête XML et store

La manipulation de documents XML est souvent faite en utilisant XPath et XSLT (Recommandations de novembre 1999), lorsque le but de l'application est l'extraction et la transformation de quantités de données relativement réduites. Depuis janvier 2007, XQuery a passé le cap de la recommandation, même si le langage était déjà utilisé. Nous ne détaillons pas ici les langages d'interrogation ayant précédé XQuery. Nous avons choisi XQuery comme langage de requêtes pour notre plateforme pour son pouvoir expressif puissant, et parce qu'il est standard. Notre interface graphique génère ainsi du XQuery compatible 1.0. Il existe diverses implémentations gratuites ou commerciales de XQuery, notre application fonctionne avec MonetDB et Saxon-B, mais nous avons également testé Qizx et eXist[1]. Le souci principal est qu'il n'y a pas encore de standard pour les interfaces dans les langages de programmation comme Java pour les bases de données XML, contrairement à ce qui peut exister avec JDBC pour le relationnel par exemple.

2.4. Sciences sociales et utilisation des bases de données

La sociologie moderne est née à la fin du 19e siècle, traitant notamment de larges quantités d'informations statistiques. Pour Emile Durkheim, l'un des pères fondateurs de la sociologie, l'utilisation de statistiques était nécessaire pour assoir la sociologie en tant que science. Pour lui, un phénomène sociologique devait être étudié comme une *réalité objective* et non pas une *idée abstraite*. Cette approche s'est développée en opposition à d'autres penseurs que Durkheim accusait de préférer le sens commun à l'expérimentation. Durkheim définit un fait social comme « *toute manière de faire, fixée ou non, susceptible d'exercer sur l'individu une contrainte extérieure* » (Durkheim, 1988). Par exemple les pratiques religieuses, le respect des devoirs civils, ou même le suicide sont des faits sociaux. Lorsque des individus collaborent à un processus de standardisation qui va déterminer l'utilisation du web pour des milliards de personnes, il font plus que simplement écrire des spécifications techniques : ils participent également d'un fait social gouverné par son propre jeu de règles.

Pour pouvoir réaliser une analyse scientifique de faits sociaux, il faut mettre de côté ses idées préconçues, comme des jugements moraux, ou vérités de bon sens. Le raisonnement statistique doit être utilisé pour distinguer l'expérience personnelle des tendances globales. Etablir un fait social de manière scientifique se traduit par sa vérification statistique : i.e. les femmes vont davantage à la messe que les hommes, le taux de suicide est plus élevé à la ville qu'à la campagne. Un peu comme en physique avec le principe de causalité, on estime que la cooccurrence de deux faits sociaux signifie qu'il y a une explication qui connecte l'un à l'autre. Par exemple, Durkheim a montré que le suicide n'était pas seulement une question personnelle ou

[1] La plupart des produits et prototypes XQuery sont listés sur http://www.w3.org/XML/Query#Products.

psychologique, mais aussi un phénomène sociologique entre autres relié à la tendance économique générale et à l'intégration sociale (Durkheim, 1997).

Bien entendu, la méthodologie de Durkheim a évolué depuis le 19e siècle. Les derniers développements sont *l'analyse factorielle* et la *sociologie des réseaux*. Etudiées depuis les années 1960, ces approches traitent des données individuelles pour en extraire des relations (sociologie des réseaux), voir chez (Granovetter, 1973, Berkowitz, 1982, Breigner, *et al.*, 1975, Lazega, 1992), des propriétés sociales dans une configuration sociale particulière, ou pour construire la typologie d'un groupe social (analyse factorielle), voir chez (Bourdieu, 1979, Benzecri, 1973). D'un côté l'analyse factorielle croise de larges quantités d'information personnelle (sexe, profession, cursus scolaire, salaire) pour construire la distribution des propriétés sociales et leur covariance, très utilisées pour examiner la structure de l'opinion, et d'autre part l'analyse des réseaux est très utile pour découvrir la structure des relations sociales parmi un grand nombre de personnes.

Néanmoins, ces approches ont trois grandes limitations. La première est *conceptuelle* : ces méthodes analysent la société comme un tout. Il est donc difficile de retrouver un comportement individuel particulier au milieu. La deuxième est *technique* : les données doivent nécessairement être quantifiées pour être traitées. De l'information qualitative comme des points de vue personnels marginaux n'est pas prise en compte. Ainsi, ces méthodes peuvent éclairer les dynamiques globales, mais ne peuvent pas véritablement décrire l'impact d'un individu sur celles-ci. La dernière limitation est que ces méthodes ne capturent pas la dimension *temporelle* autrement que par le *snapshot* : il est possible de comparer une analyse à l'instant t_1 avec une autre d'instant t_2 mais il est difficile d'expliquer ou d'explorer l'évolution temporelle. Il est important de souligner que ces aspects temporels représentent une partie non négligeable des problèmes intéressants soulevés par notre collaboration, essentiellement lorsqu'on les relie au problème du *sourcing* (i.e., qui dit quoi et quand ?). En effet, notre analyse de données du web capture ces interactions par le biais de messages échangés à des instants biens précis, entre des individus bien identifiés. L'analyse subséquente du contenu des messages permet une interprétation qualitative, en plus de l'analyse quantitative rendue possible par l'utilisation de technologies de bases de données. L'analyse des sujets de discussion sur une liste de mail, canoniquement ordonnée selon la date d'envoi nous permet d'illustrer l'évolution temporelle à la fois des acteurs, mais aussi de leurs interactions.

Citons tout de même pour finir les travaux divers de Marc Smith (http://www.connectedaction.net/), sociologue chez Microsoft jusqu'en décembre 2008, qui a étudié et surtout produit des logiciels permettant la mesure du comportement de personnes sur des listes de discussion. L'objectif principal de son outil était le filtrage d'information de qualité et l'identification de spammers. Notre approche, basée sur XML, est plus adaptée à notre application cible.

2.5. Etudes sociologiques sur la standardisation

Les sciences sociales s'intéressent depuis un certain temps déjà aux standards techniques que ce soit pour étudier leur impact dans l'entreprise (Segrestin 1997), leurs effets sur les dynamiques compétitives (Matutes et Regibeau, 1996) ou les politiques comparées des agences de normalisation (Daudigeos, 2004). Plus récemment, un certain nombre de travaux se sont concentrés sur les acteurs de la normalisation internationale et les processus concrets de production des standards (Brunsson, *et al.*, 2002, Tamm-Hallström, 2004, Graz 2006). C'est dans cette perspective que les sociologues associés au projet développent une approche qui associe les producteurs de normes aux enjeux politiques et économiques de la standardisation considérant que la définition de standards techniques signifie également choisir les entreprises et les pays qui vont contrôler cette technologie (Dudouet et al. 2006).,. Cette démarche pourrait expliquer pourquoi, comme observé par l'OECD[2], de nombreux standards qui dominent le marché des TIC ne sont pas forcément les meilleurs du point de vue technique (Besen *et al.*, 1991, OECD, 1991). Ainsi la question de *qui*, *comment* et *dans quel but* les standards sont ils adoptés devient cruciale.

Malgré l'importance de la compréhension du processus de normalisation, peu d'études ont abordé le sujet de la normalisation d'Internet, et aucun en ce qui concerne le W3C, les travaux portant sur le cadre global de la gouvernance d'Internet (Marzouki, 2008).

L'un des objectifs de notre projet était donc de combler ce manque, en apportant une étude du phénomène de normalisation du W3C.

3. L'application WebStand

Nous présentons dans cette section l'architecture de notre application, et nous détaillons les modules que nous avons développés.

3.1. Architecture globale

La Figure 1 donne une vue globale de l'architecture de notre système, centré sur la base de données XML contenant les données du web que l'on souhaite traiter. Le prototype est constitué de quatre modules : un module de définition des concepts (dans notre contexte, un éditeur de schémas XML), un module d'acquisition des données (acquisition des listes de discussion et de données sur les personnes), un module d'interrogation XQuery basé sur une interface graphique simple et un module d'exportation de ces données pour une utilisation par les logiciels de sociologie.

[2] Organisation for Economic Cooperation and Development. http://www.oecd.org/

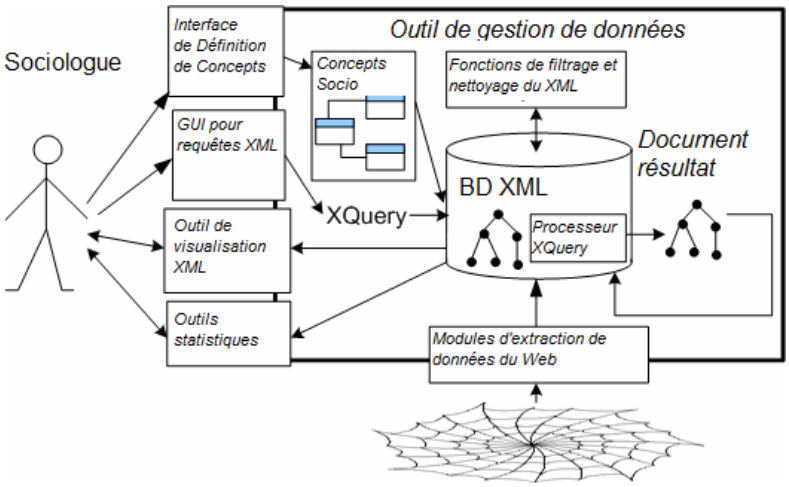

Figure 1- Architecture de WebStand

Nous présentons dans la suite de cette section le rôle de ces modules plus en détail.

3.2. Module de définition des concepts

Ce module permet la construction du modèle de données sociologique XML. Elle débute par une étude préalable permettant d'identifier les concepts pertinents pour le sociologue, dans notre exemple des personnes (appelées *acteurs*) des *institutions*, des *messages* etc, en utilisant par exemple une méthodologie et une interface décrite dans (Abiteboul *et al.*, 2006). Ces concepts sont traduits en un ensemble de schémas qui vont structurer l'entrepôt, en utilisant notre éditeur graphique de schémas.

Dans le cas de notre application qui a pour cadre les listes de discussion sur le W3C, le schéma que nous utilisons est décrit dans la Figure 2, sous forme de modèle entité-association pour plus de compacité. Ce qui est produit par notre outil est en réalité un XML Schema Descriptor (XSD), dont la partie représentant l'acteur est donnée dans l'exemple de la Figure 3.

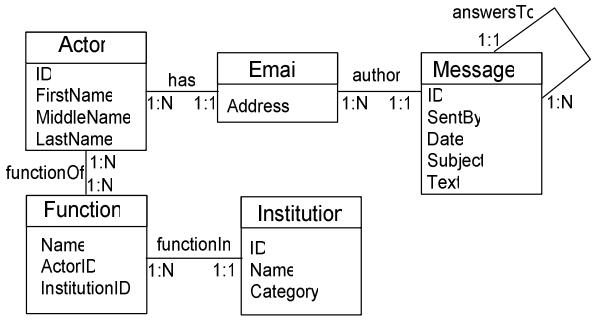

Figure 2- Modèle de données pour le cas d'utilisation W3C

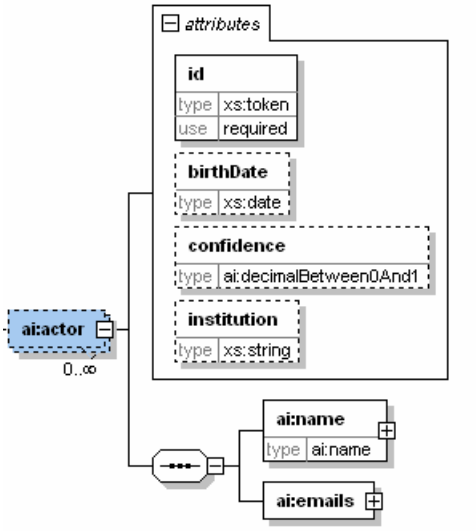

Figure 3- Le type *ai:actor*

Nous sommes intéressés à identifier des *acteurs*, qui sont les individus qui interagissent sur la liste de discussion. Nous ne détaillons pas tout le schéma ici, mais insistons sur certaines particularités : la possibilité pour un acteur d'avoir de multiples adresses mail, et d'occuper plusieurs fonctions. Notons que l'aspect temporel est présent ici uniquement dans les messages, car c'est une information concrète dont on dispose en chargeant les informations. La gestion de la temporalité d'autres informations, en particulier des informations moins sûres, ou floues comme par exemple les dates d'appartenance d'un individu à une institution est géré par le modèle de *sourcing temporel* que nous ne discutons pas ici (Nguyen *et al.*, 2009).

La Figure 3 en elle-même ne suffit pas pour définir le type *ai :actors*, par exemple on ne peut pas voir ici les références (clés étrangères en quelque sorte) que nous utilisons.

3.3. Module d'acquisition de données

Ce modèle est le point de départ de l'analyse sociologique qui va avoir lieu. Les requêtes nous amèneront peut-être à identifier d'autres entités intéressantes. Une fois ce modèle établi, nous mappons les sources de données qui nous intéressent (par exemple l'archive de mails du W3C) sur notre schéma, et nous *chargeons* ces données dans la base. Le contenu de l'archive est traité dans un premier temps de manière totalement automatique. Les informations conservées dans l'entrepôt incluent donc toute la structure des discussions qui ont eu lieu, les auteurs et bien entendu le contenu de tous les messages. Les informations sur les acteurs eux-mêmes sont complétées par des techniques de web mining, en particulier l'extraction semi-automatique de CVs des acteurs (Pinchedez, 2007).

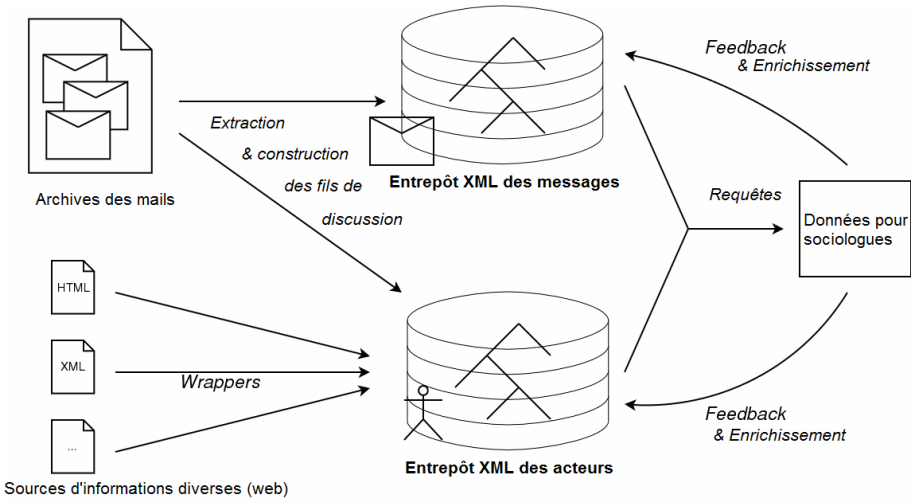

Figure 4- Processus de construction de l'entrepôt

La Figure 4 montre un diagramme du processus d'extraction d'information et de construction de l'entrepôt utilisé dans l'application W3C. Dans le dessin nous avons divisé en deux parties notre entrepôt pour montrer que ces informations étaient indépendantes, et pourraient par exemple être gérés par des serveurs différents, dans le cadre de la construction d'un entrepôt de données pair-à-pair (Abiteboul *et al.*, 2008). Notons qu'il est également possible pour les sociologues de remplir l'entrepôt « à la main » pour rajouter des informations obtenues lors d'entretiens, de recherches manuelles sur le web, ou tout simplement de correction d'erreurs.

3.4. Module d'interrogation des données (aXess)

Basé sur cette connaissance des sources, le sociologue peut commencer ses requêtes. Toutefois, les requêtes que nous utilisons, comme les requêtes XQuery en général, peuvent être assez complexes, et le fait de « vulgariser » l'utilisation de XQuery a été l'une des difficultés de ce projet. Nous avons résolu le problème en nous basant sur une analyse du travail du sociologue : Les sociologues ont tendance à vouloir effectuer des requêtes exploratoires, souvent à base d'agrégats, pour trouver des catégories, et pour classifier des personnes (e.g., une personne postant plus de 50 messages). Puis pour ces personnes, ils veulent effectuer un traitement supplémentaire (e.g., trouver sur combien de listes différentes écrit une certaine personne). Ecrire la requête globale est un travail trop compliqué pour le sociologue. Ce n'est pas seulement dû à la syntaxe de XQuery mais tout simplement à la requête elle-même, et pour faciliter son utilisation par les sociologues nous avons introduit un système de vues matérialisées. Le domaine des vues matérialisées XML est assez vaste (Abiteboul *et al.,*1998, Cluet *et al.,* 2001 Dang-Ngoc *et al.,* 2005). Nous avons donc introduit un système et une interface simple, qui permettent au sociologue de construire un ensemble de personnes qui l'intéresse, puis de les réutiliser, par le moyen de requêtes qui restent relativement simples. Nous travaillons actuellement sur la réécriture et l'optimisation de requêtes de ce style, mais ces travaux sont en dehors du contexte de cet article.

Nous avons développé une interface graphique permettant la manipulation de ces entités, appelée *aXess,* puisqu'il ressemble au système MS Access pour la gestion graphique de BDR, dont le fonctionnement est assez bien maîtrisé par les sociologues. D'autres systèmes graphiques de XQuery existent, tel que XQBE (Braga *et al.,* 2004), qui bien que plus complets sont un peu plus durs à appréhender pour le sociologue. Nous donnons dans la Figure 5 une capture d'écran de notre interface (Saint-Ghislain, *et al.,* 2008), réalisée en Java. Notons qu'aXess permet également l'édition des données XML de la base MonetDB-XQuery que nous utilisons comme moteur de stockage.

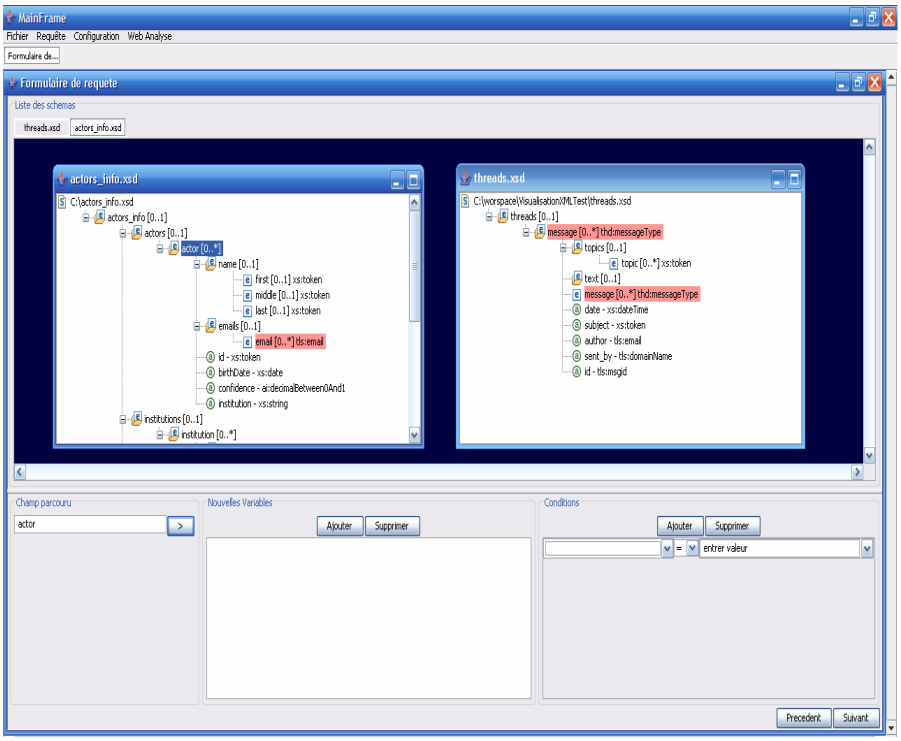

Figure 5- Interface graphique aXess

3.5. Module d'exportation des données

L'analyse sociologique requiert des outils statistiques, des tableurs, ou encore des logiciels de production de graphes, qui illustrent ou prouvent la corrélation statistique de certains événements. Nous proposons des exports dans les formats les plus usités (csv, pajek, etc.) par le biais de feuilles de transformation XSLT, mais bien entendu dans cette approche à base de XSLT, n'importe quel format peut être généré. Par exemple, les figures 6 et 8 ont été obtenues en exploitant un export de nos données dans pajek.

3.6. Appropriation des outils par les sociologues

Tout au long du projet, nous avons noué une collaboration forte entre informaticiens et sociologues, incluant un ingénieur de recherches d'un des laboratoires de sociologie, qui possédait des compétences en informatique, mais débutait en XML. Nous avons réalisé plusieurs séminaires, d'une durée totale d'une semaine, pour présenter aux sociologues les concepts des entrepôts de données XML, qui ont débouché sur la réalisation en commun du schéma de l'entrepôt : les concepts d'un schéma XML sont compris par les sociologues, mais l'écriture du

schéma lui-même a été réalisée par un informaticien en utilisant les outils graphiques. Nous avons également présenté en plusieurs sessions la réalisation logicielle du prototype aXess, sur une durée de trois jours. Par la suite, nous avons passé un certain temps, qu'on peut évaluer globalement à environ un mois, segmenté sur un an environ à travailler conjointement entre informaticiens et sociologues et à écrire des requêtes en coopération.

Notons que l'avantage majeur que nous avons trouvé lors de l'utilisation des outils graphiques, que ce soit pour la conception du schéma, ou l'écriture des requêtes réside dans le fait que c'est une manière bien plus simple de *dialoguer* entre sociologues et informaticiens. Si un informaticien peut sans problèmes comprendre la requête que souhaite poser un sociologue et peut la traduire dans une requête textuelle, les sociologues ont souvent du mal à « valider » le fait que la requête soit correcte. Avec l'outil graphique, il est beaucoup plus facile pour eux de vérifier que la requête fait bien ce qu'ils demandent, bien qu'il leur soit encore difficile d'être autonomes dans l'écriture des requêtes, même avec l'outil graphique. Compte tenu de nos expériences passées avec Microsoft Access et SQL, où nous avons noté le même comportement au départ, à savoir une compréhension de la requête (graphique) écrite par un informaticien, puis une écriture autonome en utilisant l'interface graphique, puis à terme une modification du code SQL des requêtes, voir une écriture de code SQL, nous espérons que l'utilisation de XQuery suivra un cheminement semblable, même s'il est évident que les requêtes XQuery sont plus complexes que les requêtes SQL.

Nous présentons dans le tableau ci-dessous les concepts importants que nous avons introduits auprès des sociologues, ainsi que les outils que nous avons mis à leur disposition, et nous notons leur degré d'appropriation, selon les critères suivants :

- compréhension : le sociologue a compris le concept et est capable d'expliquer la requête/schéma écrit par un informaticien.

- testé : le sociologue a testé le logiciel ou manipulé le concept, supervisé par un informaticien, et a contribué par des remarques qui ont été prises en compte.

- adopté : le sociologue utilise de manière autonome le concept ou le logiciel.

- nous indiquons enfin le degré exact de compétences dont disposent désormais les sociologues qui ont participé au projet.

Concept	Compris	Testé	Adopté	Compétence
Modèle XML	OUI	OUI	OUI	Peut lire et écrire du XML en utilisant un éditeur graphique.
XML Schema	OUI	OUI	OUI	Peut construire un schéma XML en utilisant un éditeur graphique.
Navigation dans un entrepôt de données XML	OUI	OUI	OUI	Peut consulter les données présentes dans un entrepôt XML en utilisant l'interface graphique.
Ecriture de requêtes simples (sans jointures)	OUI	OUI	OUI	Peut écrire des requêtes, incluant des agrégats simples (count) sur un document XML en utilisant l'interface graphique.
Ecriture de requêtes multi-documents (avec jointures)	OUI	OUI	NON	Comprend le concept de jointures, arrive a écrire des jointures simples entre deux attributs, mais a du mal à construire des jointures complexes mettant en jeu des expressions de chemin avec des conditions.
Réutilisation de résultats intermédiaires pour construire des requêtes complexes	OUI	OUI	NON	Arrive à matérialiser des requêtes intermédiaires, mais a du mal à utiliser plusieurs résultats intermédiaires pour arriver à un résultat complexe.

Tableau 1- Bilan de compétences des sociologues

Comme indiqué dans le tableau, les résultats de l'enquête actuelle ont été obtenus par des requêtes programmées par les informaticiens, et qui ont été validés par les sociologues, qui en ont compris la sémantique dans la vaste majorité des cas, y compris pour des requêtes temporelles mettant en jeu des opérations de regroupement par granularités différentes, ainsi que des requêtes avec des expressions régulières de recherche « plein texte ».

4. Application de WebStand à l'étude des acteurs de listes de discussion du W3C

Nous référençons (Dudouet *et al.*, 2005, Colazzo *et al.*, 2007) pour plus de résultats sociologiques de notre étude, et nous insistons sur les travaux encore en cours.

4.1. Hypothèses

Les hypothèses principales que nous émettons, et que nous souhaitons valider sont les suivantes : 1) les normes sont élaborées par un petit réseau d'experts qui ont développé un savoir faire rare 2) les normes structurent les nouveaux marchés, ainsi le processus de normalisation est un enjeu économique pour les entreprises 3) le processus de normalisation consiste à laisser un groupe de personnes ou d'entreprises monopoliser les formats des applications industrielles (Dudouet *et al.*, 2006). Ces hypothèses ont également été testées par les auteurs dans le domaine de la téléphonie mobile, mais nous nous concentrons sur les normes web dans cet article.

Afin de valider ces hypothèses, nous avons décidé d'effectuer une analyse multi niveaux en observant concrètement l'activité des experts du W3C puis en examinant leurs liens avec l'industrie ou d'autres institutions concernées par l'innovation. Une telle méthode mène à une analyse plus profonde et à des mesures systématiques des dynamiques structurelles du travail de normalisation comme étudié par Tamm-Hallsström (2001, 2004) ou Graz (2006).

4.2. Le W3C : une arène de normalisation

Le W3C est une organisation dirigée par Tim Berners-Lee, l'inventeur du web. Le rôle de cette organisation est d' « amener le web à son plein potentiel », ce qu'elle essaie de faire en publiant des *Recommandations* qui sont à quasiment tout point de vue des standards. Etre membre du W3C, et avoir un droit de vote est payant, mais tout le monde peut participer aux discussions publiques.

L'activité des experts du W3C consiste ainsi essentiellement à argumenter et débattre sur des listes de discussions, puis de rédiger des standards, au sein d'entités appelées *Working Groups* (Groupes de Travail). Les discussions publiques sont accessibles librement via le site web du W3C, ainsi que les archives mail, et nous pouvons ainsi tracer les actions, déclarations, positions de ces *acteurs* au fil du temps. Ces sources d'information web nous informent des interactions entre les acteurs d'un processus donné. Dans le cas des standards web, les listes de discussions sont extrêmement pertinentes, puisqu'elles sont en train de devenir la manière principale d'interaction entre des participants éparpillés tout autour du globe, et travaillant dans des fuseaux horaires différents. De plus, même lors de réunions téléphoniques (hebdomadaires la plupart du temps dans le cas des groupes de travail du W3C) ou de réunions physiques (Face-to-face meetings) des minutes sont toujours prises et envoyées à la liste, qui sert d'archive. Il est important

d'insister sur le fait que les techniques classiques des sociologues qui consiste à lire toutes les archives et les traiter à la main ne passent tout simplement pas à l'échelle pour ce qui est du W3C par exemple (environ 4Go de mails).

Les résultats attendus de notre analyse sont d'une part de découvrir la structure des interactions sur les listes de discussion, afin de voir qui dirige les débats, de manière quantitative (e.g, en comptant les emails et leurs réponses) et qualitative (e.g., en faisant des entretiens). Nous présentons donc les personnes et les institutions les plus impliquées dans le domaine de la normalisation de XQuery, en nous basant sur les listes publiques. Nos outils permettent bien entendu de traiter également les listes privées, mais pour des raisons de confidentialité nous ne pouvons pas présenter de tels résultats. Les données que nous avons traitées correspondent à environ 20.000 emails postés sur 8 listes différentes par plus de 3.000 personnes sur une durée de près de 7 ans.

Un certain nombre d'aspects pratiques nécessitent l'utilisation d'enrichissements sémantiques automatiques ou semi-automatiques. Typiquement la manière dont le nom d'une institution apparaît nécessite des traitements de type détection d'entités nommées. Dans tout le travail réalisé ici, l'extraction a été faite automatiquement, mais jusqu'à un certain seuil, les données ont été validées à la main : par exemple nous avons généré automatiquement les affiliations des personnes ayant posté plus de 4 messages, et nous les avons validées manuellement. Nous n'avons pas fait de validation sur les personnes ayant posté moins de 4 messages (environ 30% des acteurs). Nous obtenons globalement un taux d'erreur de 10% lors de la phase d'extraction automatique d'information.

Un autre problème important de la fouille d'une liste de discussion est le fait qu'une même personne physique peut utiliser plusieurs emails différents, avec potentiellement une signification : le fait de changer d'extension d'email signifiera en général un changement d'affiliation de la personne concernée. Nos travaux actuels prennent en charge ce problème de trajectoire, qui est ici calculée à partir de données brutes, et donne des résultats convaincants, que nous ne présentons pas ici pour des raisons de place.

4.3. Mesure d'activisme et mesure d'influence : des individus clé aux institutions clé

Les résultats présentés ici ne sont pas encore finaux, car nous travaillons encore sur les aspects temporels de données. Toutefois ils sont assez satisfaisants et convaincants pour illustrer le bon fonctionnement de notre démarche et de notre prototype.

Le corpus est constitué de 8 listes de discussion publiques du W3C, actives de 1999 à 2006 que nous avons extraites automatiquement. Nous avons également extrait les listes de recommandations techniques produites en relation avec ces listes. Nous avons choisi le thème de XML et XQuery car c'est un thème que nous connaissons bien, et ainsi nous pouvons être familiers avec les discussions techniques, et repérer les enjeux plus facilement. Nos premiers résultats concernent

la mesure de l'activisme sur ces listes. Nous considérons comme actifs les participants ayant posté au moins 20 messages sur une liste de discussion, ce qui nous a donné 72 acteurs, ayant posté 10619 messages, ce qui correspond à environ 61% des interactions, si l'on exclut les mails de « bugzilla » (3944 messages), qui nécessitent un traitement de contenu plus approfondi que nous ne traitons pas ici[3].

La Figure 6 illustre l'importance et les particularités de chaque participant, en termes de messages envoyés, et des personnes actives sur plusieurs listes de discussion. Nous représentons les listes de discussions par des losanges et les acteurs par des cercles. Plus ils sont grands, plus la liste contient des messages ou plus l'acteur a posté de messages. Afin de donner une vue globale nous avons été obligés de réduire la taille du graphe, mais des versions plus détaillées sont disponibles à la demande.

Notre première observation est que les listes de discussion ne sont pas équivalentes en termes de flux (nombre de messages) mais qu'elles sont toujours liées par au moins deux acteurs. Ces acteurs « clés » sont des multi-posteurs (14 sur 72). L'activisme en termes de participation est donc dû en particulier à ce genre de participants, même si nous trouvons également des activistes qui se cantonnent à une seule liste.

[3] Bugzilla est le système de gestion des bugs dans les spécifications du W3C. Lorsqu'une personne rapporte un bug dans ce système, un message est automatiquement envoyé à la liste de discussion, avec comme auteur bugzilla@w3.org. Afin de déterminer qui est le véritable auteur du message, il faut passer par une phase de recherche plein texte dans ces messages. Nous effectuons cette phase avec plusieurs requêtes XQ-FullText.

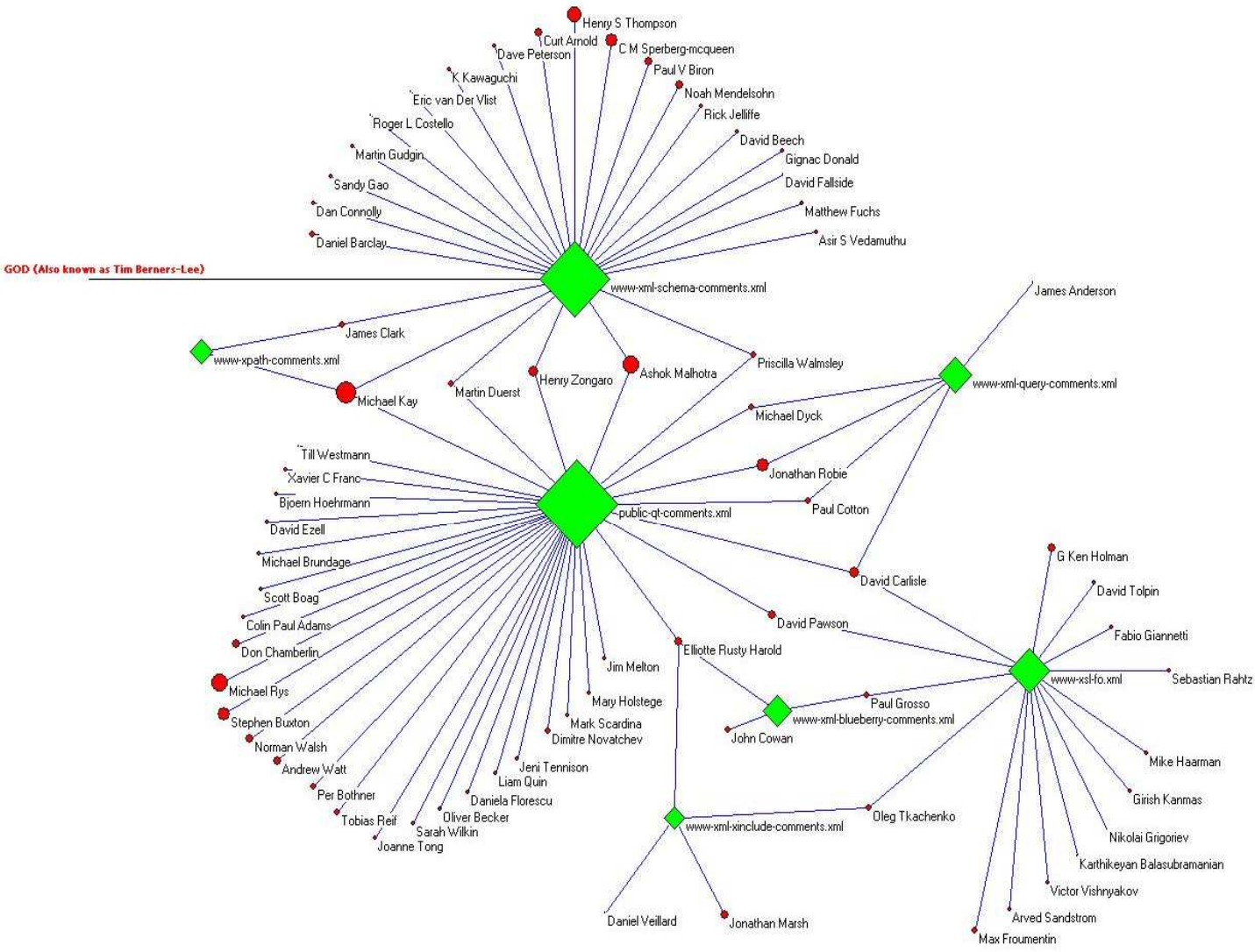

Figure 6 - Activisme des individus postant sur les listes de discussion publiques du W3C sur le domaine XML (en utilisant Pajek 1.18)

Un des résultats intéressants de notre méthode est qu'elle capture toutes les institutions clé qui sont impliquées dans le processus de normalisation du web dans

une image claire. Les études sociologiques inspirées par le *néo-institutionnalisme* ont toutes été concernées par les manières dont les entreprises réussissent à construire des réseaux durables et les utilisent pour gagner de l'influence et de la légitimité (Stark *et al.*, 2006).

Observons la Figure 8. Les losanges représentent toujours les listes de discussion, et les carrés représentent les institutions. La taille représente encore le nombre de messages. Le graphe montre que les entreprises sont très actives : les employés de Microsoft, Software AG, IBM, Oracle, Saxonica ont posté plus de 1000 messages. On voit que les institutions les plus importantes, c'est-à-dire connectant plusieurs listes de discussion sont des entreprises privées (14 vs. 8 d'un autre type – recherche publique, associations, ONGs). On pourra noter l'absence d'entreprises non américaines à part Software AG, ainsi que celle de Google, Il apparaît clairement que les instituts de recherche et les entreprises non américaines jouent un rôle secondaire.

Cette analyse ne donne pas de résultats définitifs, mais est utile pour visualiser les acteurs dans leur ensemble et pour pouvoir émettre de nouvelles hypothèses sur leur influence (Marsden, *et al.*, 1993). Il faut donc se poser la question de savoir si l'activisme de ces personnes a des conséquences sur le processus de normalisation, et pour ce faire, nous devons effectuer une comparaison avec les éditeurs de recommandations, ce qui va nous donner une corrélation entre l'activisme (le fait de poster) et l'influence (le fait de participer à la rédaction de la norme). Précisons qu'il est acquis que l'éditeur d'une norme possède une très grande influence sur les choix techniques qui sont faits.

Ainsi, afin de comprendre qui a un impact sur la rédaction des normes, nous observons désormais la Figure 7.

La table montre que la corrélation entre l'activisme et l'influence n'est pas absolue. Certaines institutions qui ont énormément participé aux Recommandations ne sont pas très actives sur les listes de discussion. Cela peut s'expliquer de trois manières complémentaires :

- Nous avons travaillé avec des listes publiques, mais les Recommandations sont discutées aussi sur les listes privées. Nous ne traitons pas les listes privées pour des raisons légales.

- Certains individus ont des fonctions multiples, et vont ainsi signer avec toutes leurs affiliations.

- En dépit de leur activisme certaines entreprises ne parviennent pas à imposer leurs choix technologiques alors que d'autres y réussissent. L'écart s'expliquerait par les conséquences de la compétition entre firmes qui ferait apparaître des gagnants et des perdants. Nous détaillons le cas de Microsoft dans (Diaz *et al.*, 2009).

En tous cas, ce sont les dix institutions les plus actives qui prennent directement part à l'écriture des Recommandations finales, si on exclut Michael Kay dont le cas

est assez spécial (programmeur indépendant, ayant travaillé chez Software AG). Nous tirons la conclusion, surtout en regardant les nombres de personnes impliquées dans l'écriture des normes, que ce sont les entreprises qui réussissent à mobiliser davantage de personnes sur ces tâches particulières et techniques, qui prennent le dessus. La tâche de normalisation, dans le domaine académique, n'étant pas reconnue. Il est intéressant de constater, et c'est l'un des résultats de notre recherche que dans le domaine de XML Microsoft s'est posé en tant que challenger, étant très actif sur les listes de discussion, les 20 employés de Microsoft ayant posté environ deux fois plus que ceux d'IBM, et pourtant avec une rentabilité moindre en terme d'écriture de Recommandations.

INSTITUTION	TYPE	# INDIV	TOTAL TEXTS	REC.	W3C WG NOTES	DRAFTS
IBM	Corp	11	13	8	2	3
Oracle	Corp	8	13	6	1	6
AT&T	Corp	2	7	4		3
Microsoft	Corp	5	6	4		2
Unknown	n.a.	2	3	3		
Sun Microsystems	Corp	1	3	3		
DataDirect	Corp	1	6	2	2	2
Univ. of Edimbourg	Uni	2	3	2	1	
Saxonica	Corp	1	2	2		
Infonyte GmbH	Corp	1	3	1	2	
Brown University	Uni	1	1	1		
CommerceOne	Corp	1	1	1		
Inso	Corp	1	1	1		
Kaiser Permanente	Org	1	1	1		
SIAC	Corp	1	1	1		

Figure 7- Corrélation entre activisme et influence sur les Recommandations

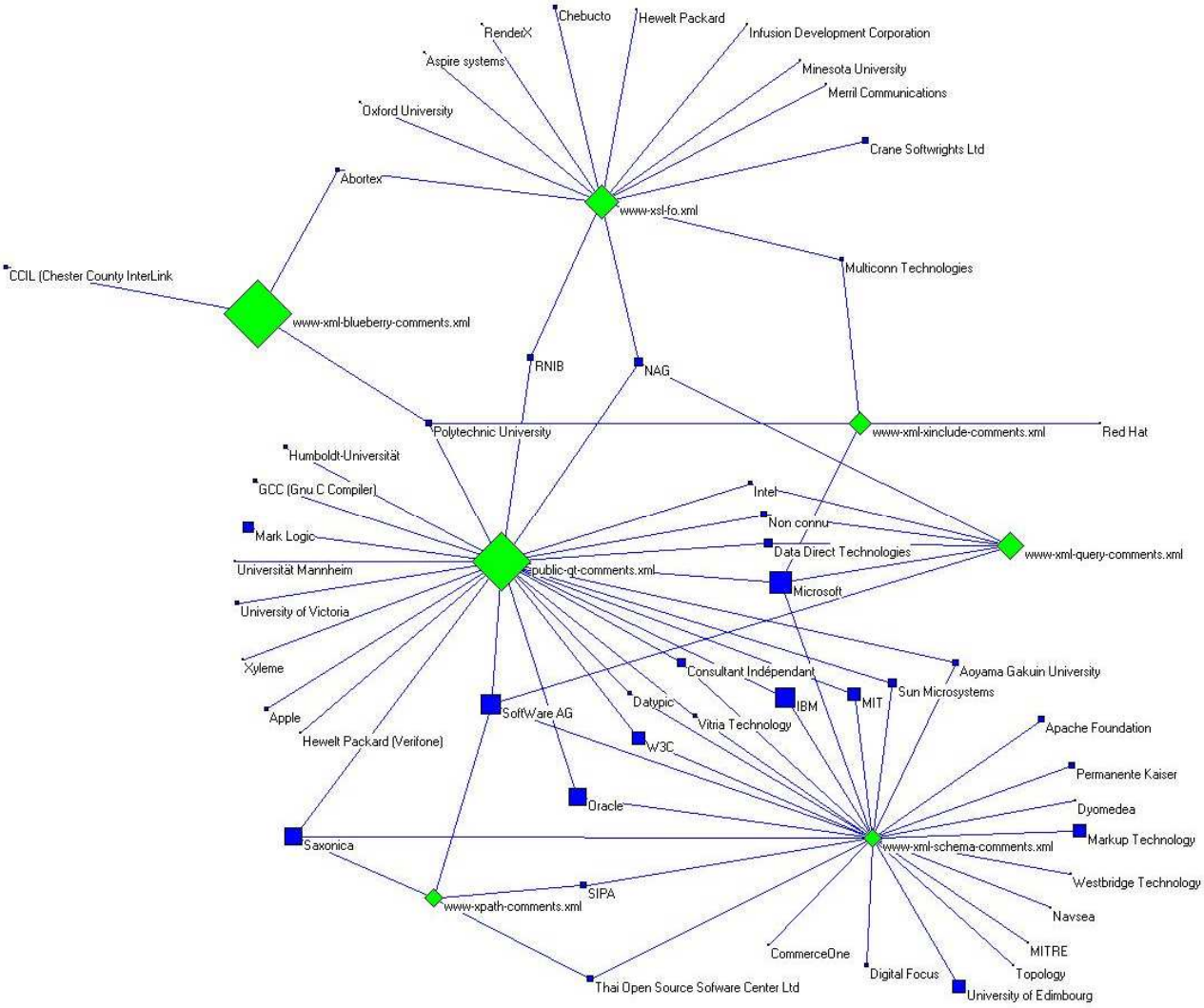

Figure 8- Activisme des institutions sur les listes de discussions publiques XML (avec Pajek 1.18)

5. Conclusions et Perspectives

Dans cet article, nous avons montré comment l'utilisation d'un entrepôt XML permet de réaliser des applications à vocation sociologique, tout en donnant accès à une approche quantitative des données, nouvelle pour les sciences sociales. Bien que nos résultats actuels soient déjà convaincants, et validés par la communauté française de sciences politiques, nous travaillons actuellement sur la prise en compte de la dimension temporelle de toutes ces informations : caractériser la trajectoire d'une personne au cours du temps, compter à une date précise les personnes travaillant pour une entreprise etc.

La difficulté majeure de ce point est la mise en place d'interfaces permettant aux sociologues de poser eux-mêmes leurs requêtes. Pour le moment, nous avons construit un outil graphique permettant d'écrire des requêtes XML en collaboration avec les sociologues, et qu'ils réussissent à lire, mais ils ne sont pas encore capables de s'en servir seuls.

Dans notre projet, la collecte et le traitement des données accessibles à partir du web ont été dans la mesure du possible automatisées, et elles sont stockées dans une base de données XML. En s'appuyant sur cet entrepôt, nous avons utilisé deux types de techniques d'analyse : des requêtes XQuery, et l'outil graphique d'analyse de réseaux Pajek. En général, tout autre outil d'analyse et/ou de visualisation qui accepte des données XML en entrée peut bien évidemment être employé.

Nous souhaitons également pérenniser notre plateforme, en trouvant d'autres applications sociologiques basées sur l'étude de listes de discussions, ou de forums afin de valider nos outils dans le cadre d'une seconde application.

6. Bibliographie

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

Abiteboul, S., Allard, T., Chatalic, P., Gardarin, G., Ghitescu, A., Goasdoué, F., Manolescu, I., Nguyen, B., Ouazara, M., Somani, A., Travers, N., Vasile, G., Zoupanos, S., *WebContent: Efficient P2P Warehousing of Web Data*. In *Proceedings of the 34th International Conference on Very Large Data Bases* (VLDB), 2008. (Demonstration Track)

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

Abiteboul, S., McHugh, J., Rys, M., Vassalos, V., Wiener, J.L., *Incremental maintenance for materialized views over semistructured* in *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB), 1998*

Abiteboul, S., Nguyen, B., Ruberg, G., *Building an Active Content Warehouse*. Chapter in *Processing and Managing Complex Data for Decision support*, Jérôme Darmont, Omar Boussaïd editors, IDEA Group Publishing, 2006

Adamic L.A., Glance N., *The political blogosphere and the 2004 US elections: divided they blog*, In *Proceedings of the 3rd International Workshop on Link Discovery*, pp 36-43, 2005.

Auray N., Conein B., Dorat R., Latapy M. *Multi-level analysis of an interaction network between individuals in a mailing-list*, in *Annals of Telecommunications*, Vol. 62, n°3-4, March-April 2007.

Beaudouin V. Velkovska J. *Constitution d'un espace de communication sur internet*, in *Réseaux*, n° 97, 123-177, 1999.

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

Benzecri J.P., *L'Analyse des données*. Dunod, 1973.

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

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

Braga D., Campi A., Ceri S. *XQBE: A Graphical Interface for XQuery Engines*. In *Proceedings of the Extending Database Technology Conference* (EDBT) pp848-850, 2004.

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

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

Calderaro A., *Empirical analysis of political spaces on the Internet. The role of mailing-lists in the organization of the global justice movements*, European Univ. Institute, 2007

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

Chateauraynaud F. *Wikipedia et les controverses de neutralité*, GSPR, Séminaire socio-informatique, concepts et méthodes pour l'analyse des dossiers complexes, EHESS Paris, 22 janvier 2007.

Colazzo, D., Dudouet, F-X., Manolescu, I., Nguyen, B., Vion, A., *Analysing Web Databases*. In *Proceedings of the French Political Science Association's Congress* (AFSP), 2007

Cluet,S., Veltri, P., Vodislav, D., *Views in a large scale XML repository*, in *Proceedings of the 27th International Conference on Very Large Databases* (VLDB), 2001

Dang Ngoc, T.T., Jamard, C., Travers, N., XLive *: An XML Light Integration Virtual Engine.* In *Base de Données Avancées* (BDA), 2005

Dang Ngoc, T.T., Sans, V., Laurent, D., *Classifying XML Materialized View for their Maintenance on Distributed Web Sources*, In *Revue des Nouvelles Technologies et de l'Information* (RNTI), 2005

Daudigeos T., *La RSE : un nouveau front pionnier pour les institiuts nationaux de normalisation : Comparaison des travaux de normalisation français et anglais*, Instruments d'action publique et technologie de gouvernement, colloque de la section d'études internationales de l'AFSP, Paris, 2004

Della-Porta D., Mosca L., *Global Net for Global Movements ? A network of networks for a movement of movements*, Journal of Public Policy, vol 25, n°1, pp165-190, 2005.

Diani M. *Social movement networks: virtual and real*, in Webster F. (ed), Culture and Politics in the Information Age, Routledge, London, pp117-128, 2001.

Diaz P.A., Dudouet F-X., Graz J-C., Nguyen B., Vion A., *Gouverner la standardisation par les changements d'arène. Le cas du XML*. In Congrès de l'Assotiation Française de Sciences Politiques, Grenoble, 2009

Drezner D.W., Farrell H., *The Power and Politics of Blogs*, In *American Political Science Association*, Chicago, USA, 2004

Dudouet F-X., *Le Contrôle international des drogues, 1921-1999*, Université Paris-X Nanterre, Thèse de troisième cycle, 2002.

Dudouet F-X., Manolescu I., Nguyen B., Senellart P., *XML Warehousing Meets Sociology*, *Proceedings of the IADIS International Conference on the Web and Internet*, 2005.

Dudouet F-X., Mercier D., Vion A. *Politiques de normalisation. Jalons pour la recherche empirique*, in *Revue Française de science politique*, vol 56, n° 3, 367-392, 2006.

Durkheim E., *Le Suicide : étude de sociologie*, Presses Universitaires de France, Quadrige (Coll.), 9e édition, 1997.

Durkheim E., *Les règles de la méthode sociologique*, Flammarion, Champs Flammarion (Coll.), 1998.

Granovetter M., *The Strengh of Weak Ties* , In *American Journal of Socio*logy, 78, 1973

Graz J-C., *Les hybrides de la mondialisation*, in *Revue Française de Science Politique*, vol 56, n° 6, 2006.

Hacker K.L., Howl L., Scott M., Steiner R., *Uses of Computer-mediated Political Communication in the 1992 Presidential Campaign: A content analysis of the Bush, Clinton and Perot Computer Lists*, Communication Research Reports 13, 1996, 138-146.

Hague B., Loader B., *Digital democracy: discourse and decision-making in the information age*, Routledge, London, 1999.

Juris J., *Networked social movements* in Castells M. (ed) *Network Society*, Edward Elgar Publishing Limited, Northampton, MA, pp 341-362, 2004.

King G., Hopkins D., *Extracting social science meaning from Text*, In *Proceedings of the French Political Science Association's Congress* (AFSP), 2007. Available online at http://www.congres-afsp.fr/tr1sess3kinghopkins-2.pdf

Lazega E., 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.

Marsden P., Friedkin N., *Network studies of social influence*, in *Sociological Methods and Research*, vol 22, n°1, 127-151, 1993.

Matutes C., Regibeau P., *A selective Review of the Economics of Syandardization : Entry Deterence, Technological Progress and International Competition* , in *European Journal of Political Economy*, 12, 1996.

Marzouki, M., *From Internet regulation to Internet governance : losses in translation*, In *Proceedings of the 2nd European Communication Research and Education Association* (ECREA) Conference, 2008.

Nguyen, B., Vion, A., Dudouet, F-X., Saint-Ghislain, L., *Applying an XML Warehouse to Social Network Analysis*, In *W3C Workshop on the Future of Social Networ*king, 2009. Available online at : http://www.w3.org/2008/09/msnws/papers/WebStand.pdf

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

Papacharissi Z., *Democracy online : civility, politeness, and the democratic potential of online political discussion groups*, New Media and Society, Vol 6(2), pp254-283, 2004.

Pavan E., Diani M., *Structuring online and offline discursive spaces on Internet governance. Insights from a new network approach to map an emergent field*, Giganet, 3rd Annual Meeting, Internet Governance Forum, Hyderabad, 2008. Available online at : http://giganet.igloogroups.org/annualsymp

Pickerill J., *Weaving a green web: environmental protest and the computer mediated communication*, in in Webster F. (ed), Culture and Politics in the Information Age, Routledge, London, pp142-166, 2001.

Pinchedez, C., *Module d'extraction de CV pour l'entrepôt WebStand*, mémoire de stage Ecole de Mines de Nancy, 2007.

Saint-Ghislain, L., Vincent, R., *Le prototype aXess, une interface grahpique pour l'entrepôt WebStand*, mémoire de stage Ecole de Mines de Nancy, 2008.

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

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.

Stark D., Vedres B., *Social Times of Network Spaces: Network Sequences and Foreign Investment in Hungary*, in *American Journal of Sociology*, Volume 111, Number 5, 2006

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

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

Trechsel A., Kies R., Mendez F., Schmitter C.P., *Evaluation of the use of new technology in order to facilitate democracy in Europe*, WP, STOA, 2003. Available online at : http://www.erepresentative.org/docs/6_Main_Report_eDemocracy-inEurope-2004.pdf

Vaisman A.A., *OLAP, Data Warehousing, and Materialized Views: A Survey*. Available at : citeseer.nj.nec.com/vaisman98olap.html, 1998

Van Aelst P., Walgrave S., *New media, new movements ? The role of internet in shaping the antiglobalization movement*, in Van de Wonk W., Loader B.D., Nixon G.P., Rucht D

(eds) Cyberprotest. New media, citizens and social movements, Routledge, London, pp 97-122, 2004.

Widom J., *Research problems in Data Warehousing.* In *Proceedings of the International Conference on Information and Knowledge Management* (CIKM), 1995.

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

The XQL query language. Available at http://www.w3.org/TR/xquery/

The W3C website. Available at http://www.w3.org/

The WebContent Project website. Available at http://www.webcontent.fr/