De l'index nominum à l'ontologie. Comment mettre en lumière les réseaux sociaux dans les corpus historiques numériques ?

Gautier Poupeau

De l'index nominum à l'ontologie.

Comment mettre en lumière les réseaux sociaux dans les corpus historiques numériques ?

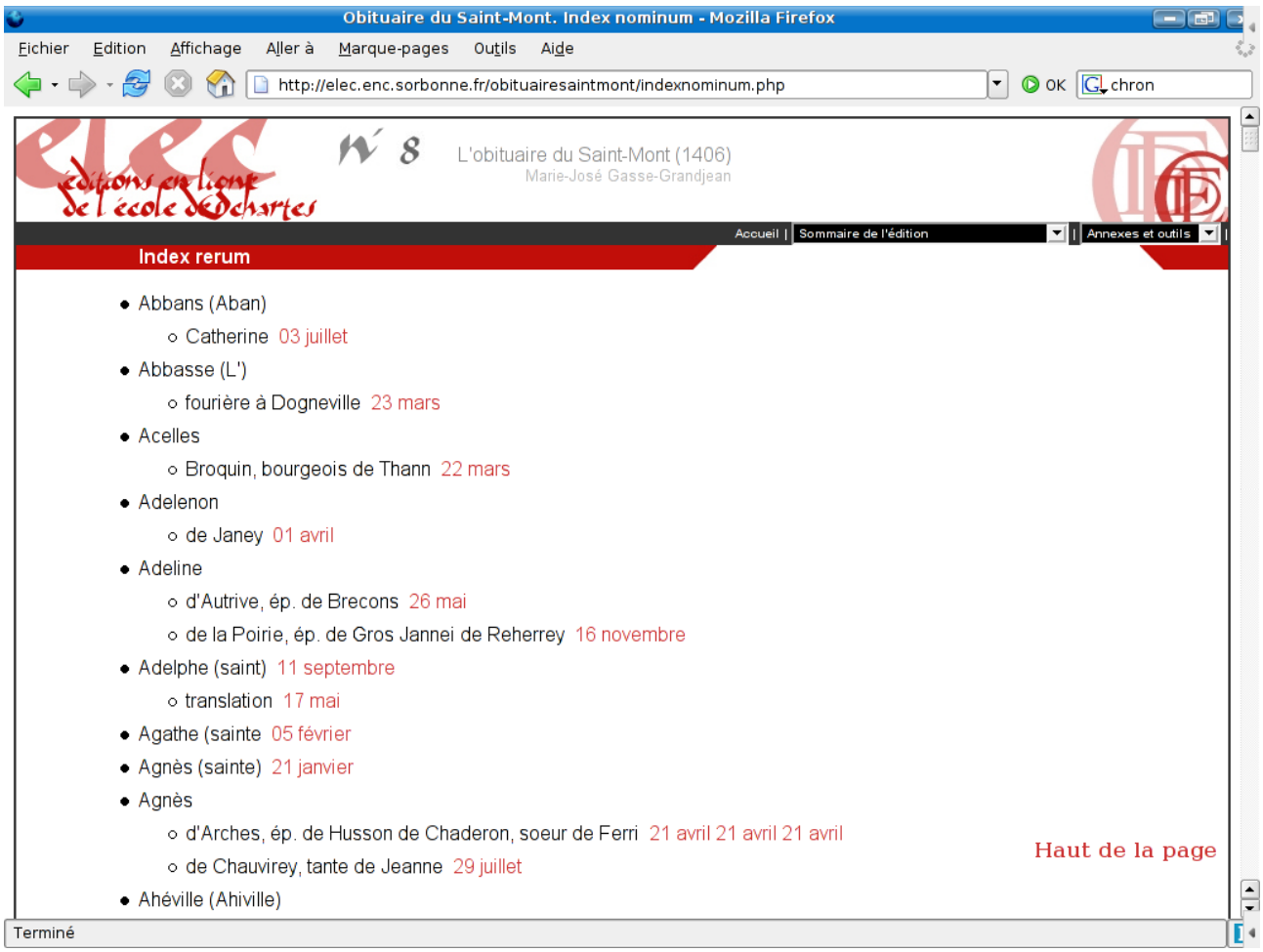
Dresser les index fait partie intégrante du travail d'élaboration d'une monographie historique ou d'une édition critique de sources pour laquelle son absence est même considérée comme une erreur scientifique. L'index constitue une liste alphabétique, sous une forme normalisée, de noms de personnes ou de lieux cités dans l'ouvrage dans le cas de l'*index nominum* ou de sujets traités dans le cas de l'*index rerum*, accompagnés de références permettant de les localiser. Il se veut exhaustif dans le cas de l'*index nominum* et se révèle forcément sélectif dans le cas de l'index des matières, les entrées étant choisies selon l'intérêt de l'éditeur scientifique. Dans le cadre de l'édition traditionnelle sur le support papier, sa présence est indispensable à plus d'un titre. En l'absence de possibilité de recherche en plein texte, il permet de retrouver une référence précise à l'intérieur du texte. De plus, il offre aux lecteurs un panorama rapide du contenu du texte forcément parcellaire en plus des tables, mais qui peut se révéler fort utile à certains stades d'une recherche, comme la mise au point du corpus. Même s'ils répondent au même but, les index peuvent présenter de nombreuses disparités de présentation. Les références peuvent renvoyer aux pages physiques de l'ouvrage ou à ses structures logiques ; il peut exister des différences typographiques, l'utilisation de l'italique, des petites majuscules ou même du gras n'étant pas normalisée, de mêmes les renvois entre les différentes entrées sont laissés à l'appréciation de l'éditeur. Cette absence de règles strictes ainsi que la difficulté de l'exercice font de l'index un travail long et fastidieux qui peut prendre jusqu'à un quart du temps consacré à l'établissement de l'édition scientifique. Toutes ces raisons expliquent en partie l'absence des index de la grande majorité des éditions électroniques aussi bien sur cédéroms que sur le Web, alors même que l'index peut rendre des services dans ce contexte. Représentant une première couche d'analyse, il pourrait sans problème constituer une matière première au traitement informatique de la source voire même créer une connaissance sur le corpus impossible à dégager par des méthodes manuelles. A travers cette communication, nous voudrions faire des propositions pour exploiter et dépasser la notion traditionnelle de l'*index nominum*[1] grâce aux outils de gestion des connaissances, en particulier ceux mis au point au W3C dans le cadre des travaux sur le Web sémantique.

1 Dans la suite de cette communication, nous entendrons le mot « index » dans le sens « index nominum ».

I- Index et édition électronique

Ces quinze dernières années ont vu la multiplication des éditions électroniques de sources historiques, d'abord sur cédéroms et aujourd'hui sur le Web. Dans l'immense majorité des cas, ces éditions ne contenaient pas d'index. Plusieurs raisons peuvent expliquer cette absence, à commencer par la forme prise par ces publications. Il ne s'agit pas à proprement parler d'éditions critiques mais plutôt de bases de données textuelles. A l'image du CETEDOC de Brepols ou de la patrologie latine de Chadwyck-Healey, ces éditions ne proposent en guise de point d'accès à l'information qu'un formulaire de recherche et ne proposent que le texte de la source, rarement des apparats critiques. L'argument économique n'est pas non plus à négliger, il y a fort à parier que le produit n'aurait pas été rentable dans le cas de l'élaboration manuelle d'un index. De plus, la recherche en texte intégral permise par l'indexation automatique du texte a pu/peut faire croire que l'index se révèle inutile surtout si le corpus fait l'objet d'une lemmatisation. Ces publications sont clairement destinées à la recherche précise de termes, de formes ou d'informations, mais en aucun cas au « grapillage » ou feuilletage de l'information. Dans ce cas, l'*index nominum* au sens traditionnel trouve difficilement sa place. Le travail éditorial et l'apport de telles publications ne se situent pas dans l'ajout de contenu scientifique, mais dans les outils de recherche permis par l'environnement numérique.

Mais, à y regarder de plus près, l'index peut rendre des services auxquels la recherche en texte intégral ne peut pas répondre. Au niveau de la mise au point de l'édition, l'index constitue une couche d'annotations permettant à l'éditeur une compréhension plus aboutie de la source qu'il édite ; il l'oblige à travailler en profondeur sur la source et le balayage exhaustif peut lui permettre de résoudre des problèmes d'appréhension, des difficultés d'identification voire même de transcription. Pour les lecteurs, tout comme pour le papier, l'index constituant un panorama, il offre un moyen de prendre connaissance rapidement du contenu de la source. Outil assez peu maniable sur le papier, cantonné à son rôle de repérage, il devient grâce à l'hypertexte un puissant et simple outil de navigation et de feuilletage du corpus. Un simple clic de souris renvoie à l'endroit exact de la référence et chaque entrée de l'index devient ainsi un parcours de lecture, le lecteur rebondissant de référence en référence.



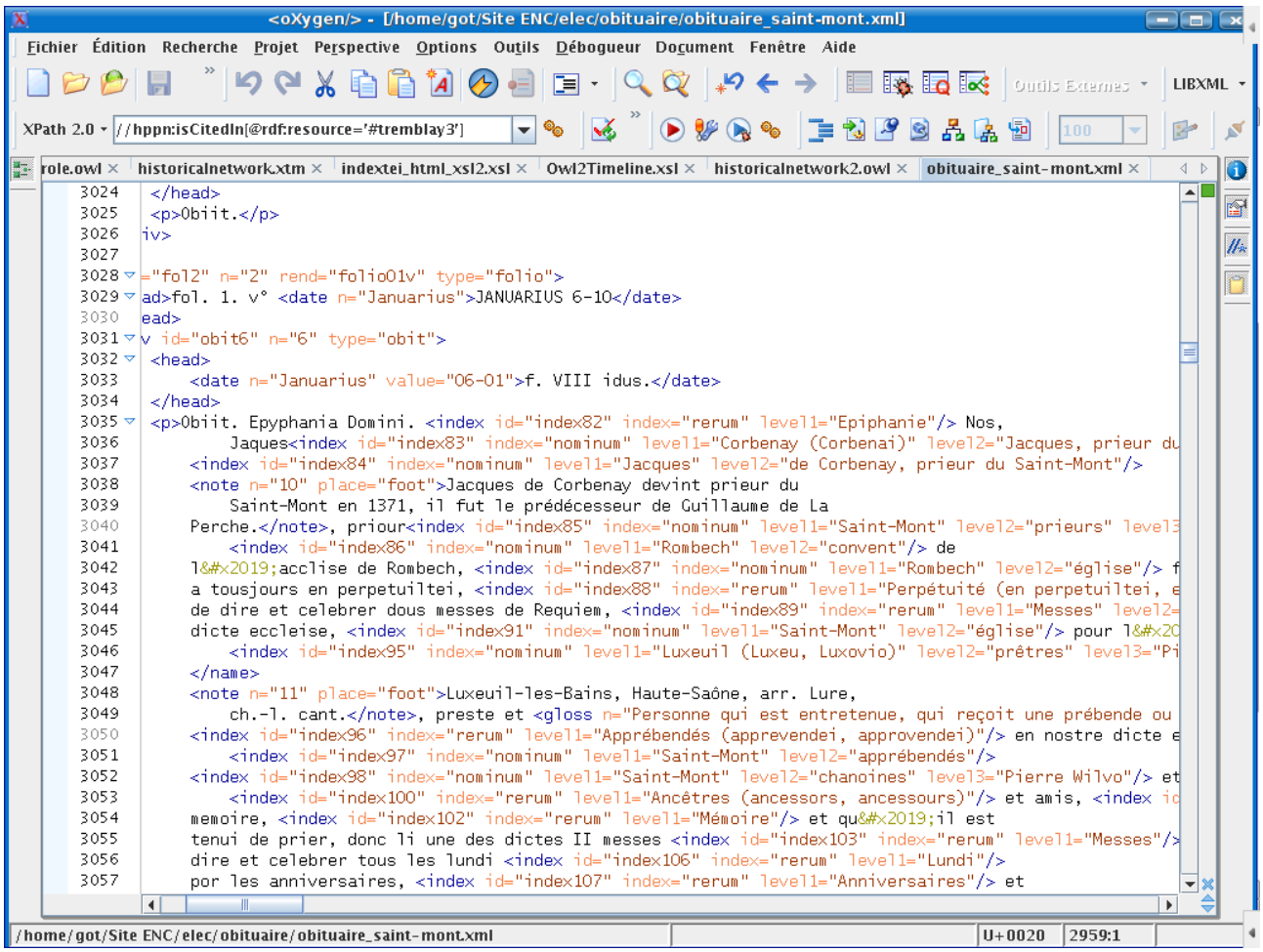
L'usage n'est pas la seule raison qui peut légitimer l'élaboration d'un index dans le contexte des éditions électroniques. Les toponymes et les noms de personnes constituent sans conteste les formes dont les graphies sont les plus disparates, surtout dans le cas de corpus de documents d'archives du Moyen Âge qui mélangent latin et langues vernaculaires. Dans ce cas, la lemmatisation automatique atteint ses limites et une indexation manuelle peut pallier en très grande partie à cette difficulté. Dans le cas d'absence de lemmatisation, l'indexation manuelle des noms de personnes et de lieux constituent une première approche qui peut se révéler payante.

L'élaboration d'un index se justifie donc complètement pour une édition électronique, d'autant plus que des éditions critiques proprement dites ont été publiées sur le Web ces dernières années, pour mémoire on citera le travail engagé au King's college ou à l'École nationale des chartes, et qu'il en fait partie intégrante ; dans ce cas, les habitudes de travail des chercheurs et les règles académiques poussent naturellement à l'élaboration de l'index, sans même se poser de questions sur la légitimité.

Pour autant, comme nous l'avons déjà dit pour le rôle de l'index, c'est son statut même et ses méthodes d'élaboration qui changent sur le support numérique ce que le guidelines de la TEI met bien en lumière. La liste d'entrées normalisées constituant l'index proprement dit n'est que le résultat d'un traitement rassemblant tous les points dans le texte que l'éditeur a référencé et associé à une entrée normalisée. Son élaboration est alors inversée, puisqu'on ne part pas de l'entrée pour aller vers la

référence, mais de la référence pour aller vers l'entrée. Ce système n'est d'ailleurs pas spécifique à la TEI, puisque les outils pour gérer les index dans des traitements de texte comme Word ou Open office fonctionnent de la même façon. Il présente en effet l'avantage de rendre l'index indépendant des changements de mise en page et donc de pagination pour le support papier et laisse libre le niveau de structure logique qui va servir à la référence pour l'édition électronique.

Dans le cadre de la TEI, l'élément <index> permet d'associer un point à une entrée normalisée avec les attributs level1, level2... en P4 ou avec l'élément <term> en P5. Mais plutôt que référencer un point, il peut être plus intéressant de référencer la portion d'informations à laquelle est associée l'entrée de l'index. Dans ce cas, il est possible d'utiliser les éléments <persName> et <placeName> en TEI associés à une entrée normalisée avec l'attribut key ou en référence avec une liste donnée dans le teiHeader. Ce système permet de dresser la liste de l'ensemble des formes d'un toponyme ou d'un nom de personnes et de constituer ainsi un lexique réutilisable dans d'autres corpus.



Malgré les services qu'il peut rendre, l'exploitation de l'index peut paraître encore trop minime par rapport au temps qu'a demandé son élaboration. Les outils de gestion des connaissances, en particulier ceux mis au point au W3C dans le cadre du Web sémantique, nous permettent alors d'exploiter facilement et rapidement l'index.

II- Exploiter l'index

Le Web sémantique est une série de travaux, de recherches et de spécifications mis au point au W3C dont le but principal est de donner du sens aux informations présentes sur le Web pour mieux les échanger. Cette activité du W3C est portée directement par Tim Berners-Lee, l'inventeur du Web qui en a dessiné les contours dès 1998 et qu'il a formalisée dans le diagramme appelé « W3C semantic stack ». Comme il est visible sur ce diagramme, le socle commun à toutes les technologies du Web sémantique est le modèle RDF qui offre un moyen d'écrire des prédicats sous forme de graphes et qui repose sur une syntaxe XML et la norme d'adressage URI. La formalisation des connaissances repose sur le concept d'ontologies composant le socle supérieur à RDF.

A l'origine, le terme « ontologie » désigne en philosophie l'étude des propriétés générales de ce qui existe. Les informaticiens spécialistes de gestion des connaissances l'ont repris à leur compte pour définir un « ensemble structuré de savoirs dans un domaine particulier de la connaissance » ou un ensemble de concepts organisés en graphe dont les relations peuvent être sémantiques ou de composition et d'héritage. Ce concept a donc naturellement trouvé sa place dans le Web sémantique et une norme basée sur le modèle RDF a vu le jour pour construire des ontologies : OWL (Web ontology Language).

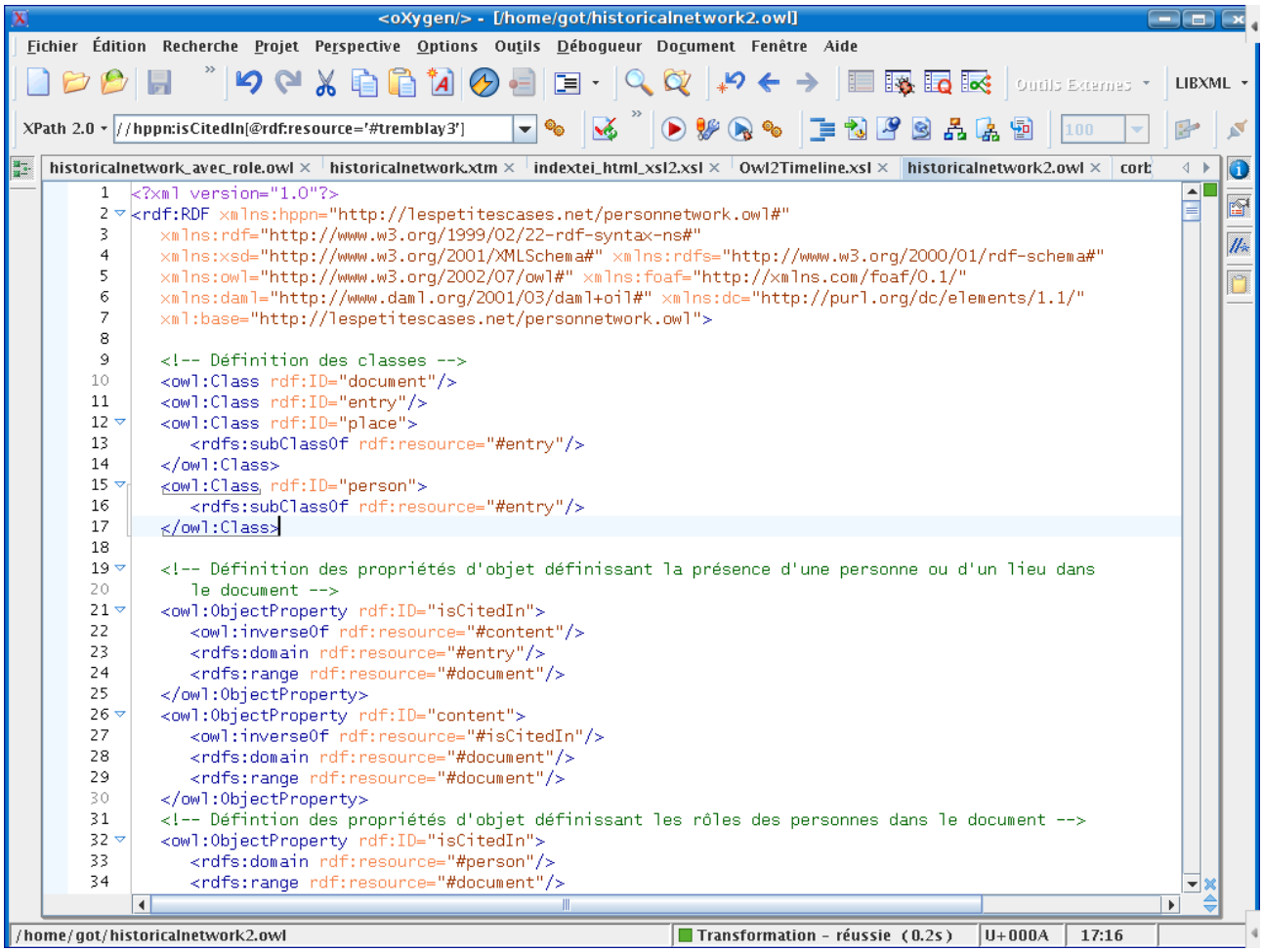
L'étude attentive de la structure des index fait immédiatement penser à la composition d'une ontologie. Malgré les disparités qu'il existe entre les différents index, nous pouvons faire apparaître une structure commune :

- l'entrée normalisé de l'index ;
- l'ensemble des formes dans le texte, éventuellement remis au nominatif, faisant référence à cette entrée ;
- des indications biographiques ou géographiques ;
- la liste des références dans le texte

Deux objets principaux ou classes se distinguent : les documents référencés et les entrées qui peuvent elles-même être séparées en deux objets : les personnes et les lieux. Une relation unit les documents et les entrées.

Pour clarifier mon propos, je vous propose un exemple utilisant un corpus d'actes médiévaux, en l'occurence un chapitre du Cartulaire blanc de l'abbaye de Saint-Denis dont l'édition est en cours d'élaboration et de mise en ligne sur le site Web de l'École des chartes. Dans le texte balisé en TEI, ici en P4, nous avons indexé les noms de personnes et de lieux avec l'élément <index/>. L'acte constitue la structure logique de base et nous pouvons le considérer comme un document en tant que tel. Cet encodage permet évidemment de dresser un index. Mais, suivant la structure de l'index décrite précédemment, il est assez simple avec une feuille de style XSL de transformer cet index en une ontologie composée de quatre classes : le document, l'entrée, la personne et le lieu, tous deux des

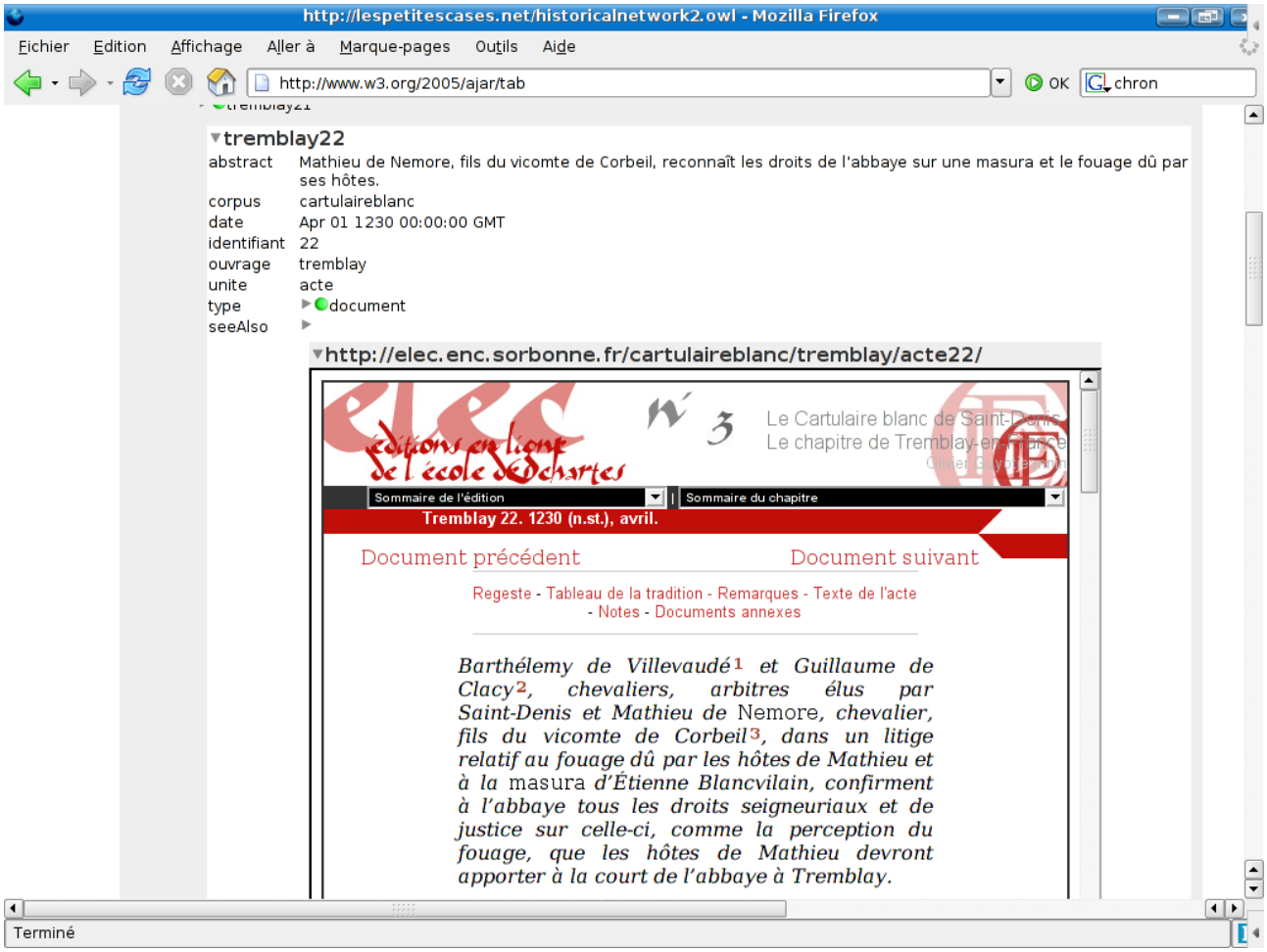
sous-classes d'entrée et de deux relations dites inverses : une qui unit le document à une entrée (qu'elle soit de type personne ou lieu) et une seconde qui unit l'entrée au document. En indiquant dans la structure de l'ontologie que ces deux propriétés sont inverses, le système informatique déduira automatiquement l'une de l'autre ce qui permet de faire des économies d'expressions.

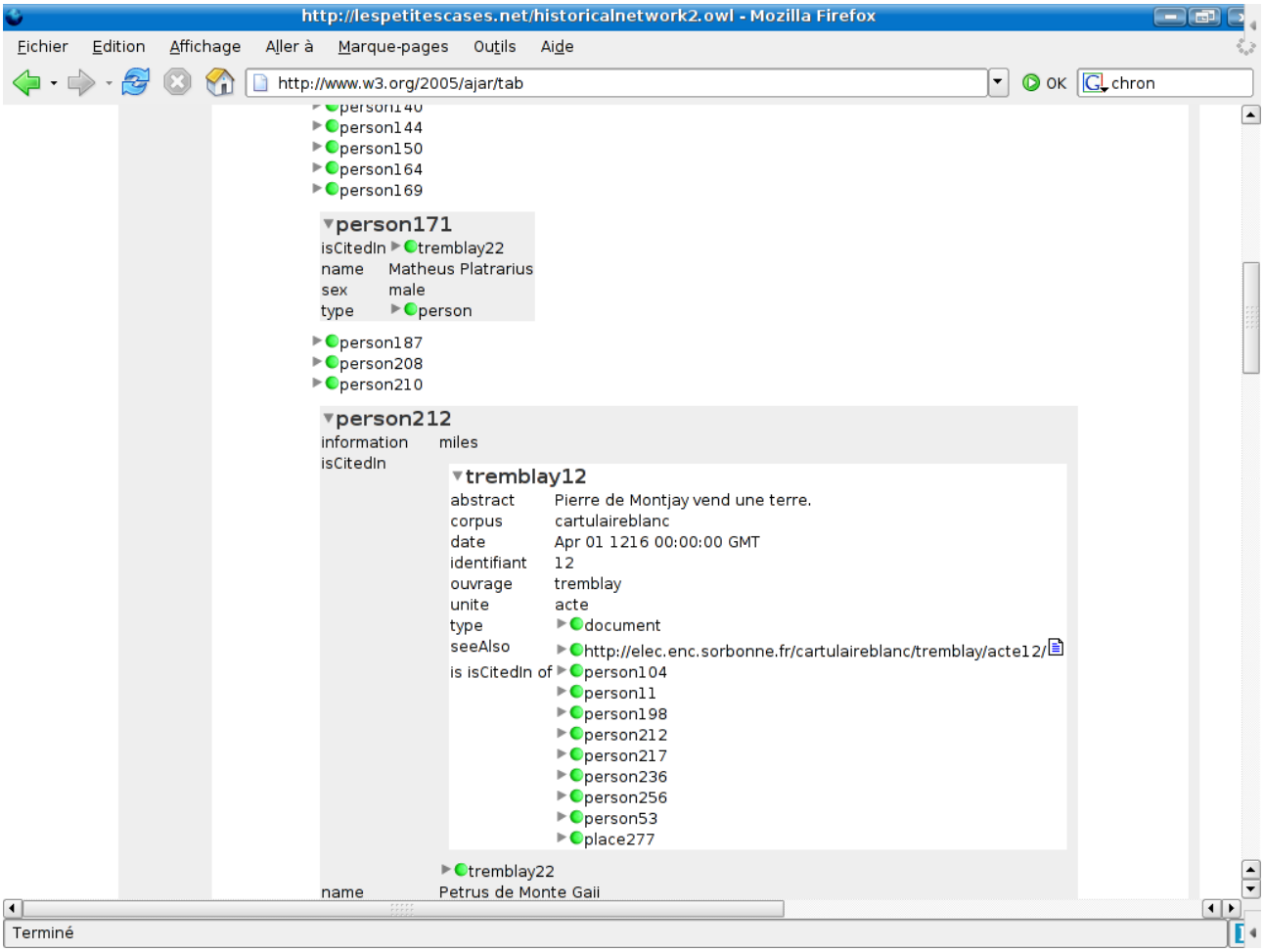


Ce fichier ne se suffit évidemment pas à lui même, il faut pouvoir le traiter. Dans le cadre de cette expérience, deux outils ont été utilisés. Le premier est tabulator, un navigateur de fichier RDF écrit en javascript et mis au point au W3C par Tim Berners-Lee et son équipe dont le but est de rendre accessible RDF et de montrer rapidement les possibilités du Web sémantique.

Dans notre cas, ce but est tout à fait atteint, car grâce à cet outil, il est très simple de naviguer dans notre ontologie et de faire ainsi apparaître de nouvelles informations directement dérivées de l'index. Il est ainsi aisé de visualiser pour une personne les actes dans lesquels elle est présente et de comparer pour chacun de ces actes les personnes présentes. La grande force de cet outil est de pouvoir afficher plusieurs prédicats et ainsi de les comparer facilement. Cet outil permet même de visualiser directement une page Web et dans notre cas le texte de l'acte. Avec notre ontologie et tabulator, on voit apparaître des réseaux en comparant les personnes présentes dans chaque acte, on peut alors par exemple déduire l'existence d'une relation entre deux personnes si elles prennent part à l'action juridique de deux actes différents. Évidemment, cette hypothèse demandera une confirmation que seul

l'historien peut apporter, mais dans ce cas l'outil informatique a fait apparaître de nouvelles connaissances ou du moins informations dont le traitement manuel peut s'avérer fastidieux surtout dans le cas d'un corpus très important.

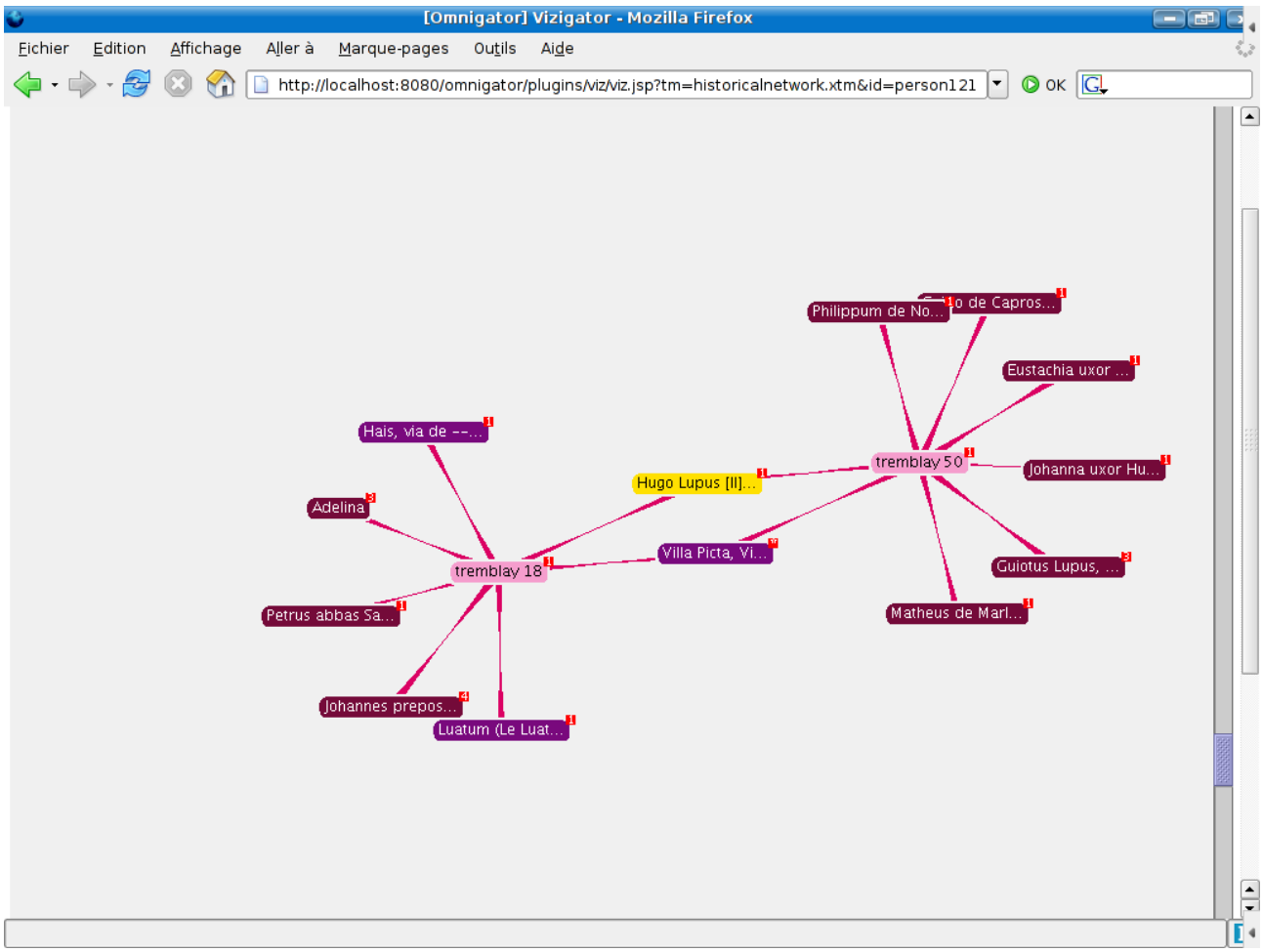




Le second outil permet de parvenir aux mêmes conclusions, mais propose des méthodes de visualisation un peu différentes. Il s'agit du logiciel Omnigator de la société Ontopia qui ne repose pas sur OWL, mais sur Topic Maps. Topic maps est une norme ISO qui ne fait pas partie des spécifications issues du Web sémantique, mais dont les principes généraux ressemblent à peu près au principe de RDF avec lequel des essais d'interopérabilité voient régulièrement le jour. Pour passer de OWL à Topic maps, une feuille de style XSL suffit, utilisant tous les deux une syntaxe XML. En lieu et place de la visualisation hiérarchique proposée par tabulator, Omnigator propose de visualiser les fichiers Topic maps de deux manières :

- une interface hypertextuelle, chaque page Web correspondant à la visualisation d'un topic et de l'ensemble des associations auxquels il participe
- une interface sous forme de graphiques, permettant de déplier ou replier les sujets intéressants.

C'est évidemment la deuxième interface qui peut nous rendre de grands services en faisant apparaître graphiquement les relations entre les personnes et les documents.



Cet exemple montre les possibilités de création d'une première couche de connaissances à partir des données brutes issues d'un index. Pour autant, les possibilités offertes par les ontologies permettent d'aller plus loin et de dépasser la notion même d'index.

III- Dépasser l'index traditionnel

Le fichier OWL obtenu à partir de l'index peut ensuite être réutilisé de diverses manières qui ne sont pas concurrentes, mais plutôt concomitantes. En premier lieu, il peut jouer le rôle d'un fichier d'autorité, devenant un méta-index pour différents corpus. L'École des chartes mène actuellement une campagne de numérisation d'éditions de chartriers d'établissements ecclésiastiques d'Ile-de-France tombées dans le domaine public. Il serait très utile de disposer d'un index général sur l'ensemble de ces éditions. Ce fichier pourrait non seulement servir de référentiel pour indexer les différents ouvrages au fur et à mesure. Mais, il constituerait surtout une matière première complètement inédite pour étudier les rapports entre la société et les établissements ecclésiastiques d'Ile-de-France, décuplant à l'échelle de plusieurs milliers d'actes les résultats de l'exemple précédent.

Encore plus intéressant, des informations peuvent être ajoutées à ce fichier : le rôle tenu par la personne dans chaque document dans lequel elle est présente, son sexe, des mentions biographiques et surtout la mention de relation entre les différentes personnes et entre les personnes et les lieux en typant si possible ces relations qu'elles soient de sang ou juridiques pour les personnes : filiation, seigneur/vassal, époux/épouse ou sur les rôles tenus par une personne : roi de, abbé de, seigneur de... Pour ce genre de relations, l'utilisation du langage OWL révèle alors toute sa puissance, car un de intérêts de ce langage est la possibilité de définir des caractéristiques sur les propriétés. A partir des caractéristiques et des propriétés, un logiciel appelé « raisonneur » peut effectuer des inférences sur ce fichier. Dans la description des réseaux sociaux, deux caractéristiques sont particulièrement intéressantes :

- SymmetricProperty : un propriété symétrique, par exemple, soit la propriété « frère/sœur de » symétrique et soit A frère/sœur de B, automatiquement un logiciel d'inférences peut déduire que B « frère/sœur de » A.
- inverseOf : Deux propriétés sont inverses, soit une propriété « enfant de » inverse de la propriété « parent de », soit A enfant de B, alors B est parent de A.

Ces mécanismes ont l'air triviaux pour un humain mais ils ne le sont pas pour une machine et ramené à plusieurs centaines de personnes, ces deux caractéristiques ouvrent des perspectives nouvelles pour les chercheurs du domaine. Combiné à la première couche de connaissances obtenue avec l'index, cette deuxième couche permet de mettre en lumière tous les réseaux sociaux présents dans les corpus référencés dans ce fichier qui forme alors une base de données prosopographique amendée au fur et à mesure de l'avancée des recherches et du dépouillement des sources.

Les outils utilisés précédemment offrent avec cette nouvelle couche d'annotations de nouveaux résultats. Mais, dans ce cas, un autre outil peut se révéler particulièrement utile : SPARQL (prononcé sparkle). Il s'agit d'un langage de requêtes spécifiques à RDF actuellement au statut de *candidate recommandation* au W3C dont la syntaxe est assez proche de SQL. Il permet d'interroger simplement

les prédicats d'un fichier RDF. Sa relative simplicité laisse envisager une utilisation assez aisée par les chercheurs. Enfin, il est tout à fait possible d'envisager la génération automatique d'arbres généalogiques en SVG par feuilles de style XSL à partir des données contenues dans le fichier OWL.