Le projet SyMoGIH et la modélisation de l'information : une opération scientifique au service de l'histoire

Francesco Beretta, Pierre Vernus

Le projet SyMoGIH et la modélisation de l'information :
une opération scientifique au service de l'histoire

par
Francesco Beretta et Pierre Vernus
(Université de Lyon – CNRS UMR 5190 LARHRA)

Le nom du projet SyMoGIH, 'Système modulaire de gestion de l'information historique', exprime une idée centrale du projet : mettre en place un système de stockage collaboratif de l'information permettant aux historiens d'utiliser différents outils logiciels (statistique, analyse des réseaux, etc.) pour exploiter les données stockées dans des bases de données structurées selon une sémantique commune. Le but visé n'est donc pas celui de produire un nouveau logiciel mais de mettre au point une méthode au service de la recherche historique portant sur une utilisation conforme aux standards actuels de bases de données et de textes codés en xml. Dans cette démarche, la modélisation des données occupe une place centrale car, comme dans tout système informatique, elle permet de produire une sémantique grâce à laquelle on peut interpréter correctement et donc échanger les données stockées : l'acronyme SyMoGIH pourrait donc tout aussi signifier 'Système de *modélisation* et de gestion de l'information historique'.

En effet, dès que l'historien souhaite stocker l'information qu'il récolte dans une base de données, il va entreprendre, de façon consciente ou inconsciente, une opération de modélisation qui produira une sémantique structurant ses données. Preuve en soit le fait que même dans l'approche courante, qui consiste à installer un logiciel de gestion de bases de données sur son ordinateur personnel, par exemple FileMaker, puis à créer une table consacrée aux acteurs contenant différentes propriétés les concernant (le nom, la date de naissance, la ou les professions, etc.), on produit implicitement un modèle de données : la table reproduit un ensemble abstrait –la classe 'acteur'–, les lignes représentent les individus, les champs de la table représentent les propriétés considérées comme communes aux acteurs, etc. La modélisation est bien là, mais elle est inconsciente ou non réfléchie.

De plus, cette approche se limite souvent à organiser l'information en champs texte : certes, l'auteur de la base connait sa sémantique, c'est-à-dire la signification des tables et des champs qu'il a créés, et il peut ainsi interroger ses données. Mais les possibilités de recherche sont limitées à une interrogation dans les textes, sans pouvoir tirer profit de toute la puissance du modèle relationnel propre aux bases de données[1]. Aussi, la fait que la sémantique ne soit pas documentée explicitement rend parfois difficile à l'auteur lui-même la réutilisation de ses données quelques années plus tard.

Dès les débuts du projet SyMoGIH, en 2007, nous avons pris conscience de ce problème et, par conséquent, de l'importance d'une réflexion approfondie sur l'opération de modélisation. Cette démarche s'imposait car le projet résultait de la volonté de quelques collègues spécialistes de différents disciplines (histoire intellectuelle, économique, religieuse, politique, ...) de mutualiser la méthode de stockage des données et, si possible, les données elles-mêmes. Il a ainsi paru opportun de se mettre à l'école des informaticiens et de s'approprier une méthode de

1 Cf. http://fr.wikipedia.org/wiki/Modèle_relationnel (consulté le 31 janvier 2012).

modélisation pour l'appliquer ensuite à notre projet. Il fallait, en d'autres termes, 'inventer' une méthode de modélisation de l'information historique visant une certaine 'objectivité', c'est-à-dire créer une sémantique adaptée au stockage de tout type d'information historique, indépendante de la problématique particulière de chaque recherche et susceptible d'être utilisée par les spécialistes de différentes disciplines[2].

Au sein du projet SyMoGIH, cette méthode a été appliquée à trois niveaux : un niveau de construction de bases de données à usage individuel ; un niveau articulant entre elles bases individuelles et bases produites dans le cadre d'un projet de recherche, à l'aide de dictionnaires d'autorités communs permettant de mutualiser les identifiants d'objets et types d'informations ; un niveau de gestion collective et cumulative de l'information, celle-ci étant stockée dans une base commune que nous avons appelée la Base d'hébergement de projets (BHP). Le fait d'avoir saisi les données selon une structure commune, explicitée par une documentation accessible à tous les utilisateurs, permet leur interopérabilité et leur récupération pour de nouveaux projets. Le but de cet article n'est pas de rendre compte des différentes expériences concrètes faites ces cinq dernières années mais de présenter le processus scientifique qui a conduit à la mise en place de la sémantique propre à SyMoGIH.

La méthode Merise appliquée à l'histoire

A la base du processus de stockage numérique de l'information se situe le choix de la technologie qui hébergera les données, ainsi que celui de la méthode de modélisation. Concernant la technologie, on peut opter entre le choix plus 'classique' d'une base de données relationnelle, ou alors celui de textes codés en xml, par exemple selon la sémantique de la *Text encoding initiative* (TEI)[3] qui s'est imposée comme l'un des standards en sciences humaines, ou encore celui de la mise en place d'une ontologie utilisant le langage de représentation des connaissances *Web Ontology Language* (OWL)[4]. Pour des raisons liées à la fois à nos compétences personnelles, à la souplesse de l'outil et au volume important d'informations à stocker, nous avons opté pour une approche fondée sur l'utilisation des bases de données

2 Une démarche analogue à la notre a déjà été adoptée en France, entre autres, par le système FICHOZ, cf. Dedieu, Jean-Pierre, « Les grandes bases de données. Une nouvelle approche de l'histoire sociale. Le système Fichoz », *Revista da Faculdade de Letras HISTÔRIA* 3(2005), 99-112 (http://halshs.archives-ouvertes.fr/halshs-00004690/fr/, consulté le 30 janvier 2012), ainsi que par le projet Préfen, cf. Landry, Yves (éd.), *Registres paroissiaux, actes notariés et bases de données*, Caen, Centre de recherche d'histoire quantitative, 2005 (http://www.unicaen.fr/mrsh/prefen/index.php, consulté le 30 janvier 2012). Notre méthode est issue des mêmes interrogations mais nous avons souhaité y répondre en appliquant la méthode de modélisation Merise à la même problématique, c'est-à-dire celle du stockage collaboratif des informations historiques sous forme base de données. Notre projet découle d'une double volonté : celle d'expliciter la sémantique mise en place pour faciliter la communication entre historiens au niveau du modèle conceptuel ; celle de s'inscrire d'emblée dans les standards informatiques actuels et de pouvoir ainsi utiliser des systèmes de gestion des bases de données robustes, tel PostgreSQL, tout en confiant aux informaticiens la réalisation et la gestion de la base de données.

3 http://www.tei-c.org (consulté 30 janvier 2012).

4 Evans, Colin / Segaran, Toby / Taylor, Jamie, *Programming the Semantic Web*, Sebastopol (CA), O'Reilly, 2009 ; Dengel, Andreas, *Semantische Technologien. Grundlagen, Konzepte, Anwendungen*, Heidelberg : Spektrum Akademischer Verlag, 2012.

relationnelles, tout en utilisant de façon complémentaire le stockage des textes sous forme XML selon la sémantique de la TEI. Une fois ce choix opéré, se pose la question de la méthode de modélisation : faut-il adopter la méthode plus classique entité-association, intégrée à la méthode Merise dans le monde francophone, ou s'orienter vers la plus récente modélisation UML qui fournit un outil bien plus riche et puissant car il intègre le modèle des données et celui des traitements[5] ?

Etant donné qu'il s'agit pour nous d'exprimer l'articulation relativement statique entre objets historiques, bien qu'évoluant dans l'espace et dans le temps, et non pas de gérer des flux de marchandises en temps réel, ce qui aurait impliqué de modéliser en même temps les cas d'utilisation, les données et les traitements, le Modèle conceptuel des données (MCD) selon le formalisme de la méthode Merise a paru suffisant : il fallait éviter que le niveau d'abstraction de la modélisation soit d'emblée trop élevé pour être accessible aux historiens. Ceci d'autant plus que nous avons souhaité, en tant que spécialistes impliqués dans la production de nos données, nous approprier cette méthode pour l'intégrer à notre propre démarche scientifique.

En effet, la production des données à partir des connaissances que contiennent les sources représente le fondement de la méthode historique. Or, la fonction d'un MCD étant de « rendre compte correctement de la sémantique du domaine modélisé », en produisant « une description naturelle du monde réel » apte à répondre « aux requêtes potentielles des applications qui utiliseront la base de données »[6], il est indispensable que l'opération d'extraction des connaissances des sources et de leur stockage dans une base de données soit effectuée selon un modèle sémantique construit par l'historien en conformité avec les critères scientifiques de sa propre discipline. Notre démarche n'est pas inédite en histoire et elle a été déjà pratiquée avec profit tant au niveau de recherches individuelles[7] que collectives[8]. Dans le cadre du projet SyMoGIH, il s'agissait de dépasser ces approches liées au traitement d'objets et de problématiques historiques spécifiques et de tenter de mettre en place un système généraliste.

Ce projet soulève d'emblée de nombreuses questions : quel type de connaissances faut-il modéliser ? Celles que contiennent les sources comme telles ? Celles reconstituées grâce à l'analyse de plusieurs sources et à la synthèse effectuée par l'historien ? Comment construire des données réutilisables par d'autres projets, en évitant que la problématique particulière du chercheur influence excessivement l'extraction des connaissances ? Comment gérer le sourçage de l'information, fondement de la méthode historique ? Le stockage de connaissances issues de textes dans une base de données relationnelle ne comporte-il pas par définition un

5 Soutou, Christian, *UML 2 pour les bases de données*, Paris, Eyrolles, 2007.

6 Audibert, Laurent, *Bases de données de la modélisation au SQL*, Paris, Ellipses, 2009 p.27.

7 Cellier, Jacques / Cocaud, Martine, *Traiter des données historiques : méthodes statistiques, techniques informatiques*, Rennes, Presses universitaires de Rennes, 2001 ; Lewis, M. J. / Lloyd-Jones, Roger, *Using Computers In History. A Practical Guide*, London New York, Routledge, 1996, chapitres 8 et 9 (2e éd. London e.a., Routledge, 2009) ; Harvey, Charles / Press, Jon, *Databases in Historical Research*, Basingsoke, Palgrave, 1996, notamment le chapitre 5 ; Merry, Mark, Databases for historians (2011), chapitre E : http://training.historyspot.org.uk/mod/book/view.php?id=75&chapterid=144 (consulté 30 janvier 2012, inscription nécessaire).

8 Gast, Holger / Leugers, Antonia / Leugers-Scherzberg, August H., *Optimierung historischer Forschung durch Datenbanken. Die exemplarische Datenbank "Missionsschulen 1887-1940"*, Bad Heilbrunn, Verlag Julius Klinkhardt, 2010

appauvrissement de l'information qui la rend inutilisable ? Nous verrons dans la suite de l'article les réponses que la méthode SyMoGIH permet d'apporter à ces questions. Au préalable, nous présenterons les fondements de la modélisation selon la méthode Merise appliquée à l'histoire, dans une démarche qui, comme nous l'avons dit, n'a rien d'original puisqu'elle est déjà pratiquée par de nombreux collègues, tant au niveau individuel que collectif : l'originalité de notre projet réside donc dans la manière d'appliquer cette démarche.

La modélisation propre à la méthode Merise propose une distinction fondamentale entre données et traitements[9]. Nous nous limiterons ici à la modélisation des données, c'est-à-dire à la statique du système d'information. Elle s'articule en trois niveaux : le modèle conceptuel des données (MCD) qui décrit le monde réel en termes d'entités, propriétés et associations ; le modèle logique des données (MLD) qui transcrit le MCD sous forme de tables, selon une structure relationnelle qui peut être implémentée dans n'importe quel système de gestion de bases de données (SGBD) ; le modèle physique des données qui, après avoir choisi le logiciel de SGBD (par ex. FileMaker, Oracle, MySQL, PostgreSQL, etc.), réalise concrètement le stockage selon la structure envisagée par le MLD.

De cette distinction découle la mise en place de systèmes de bases de données robustes qui permettent de faire tourner les services que nous connaissons : des horaires des trains, aux sites commerciaux, aux cartes bancaires. Si on revient à l'exemple, évoqué ci-dessus, de l'historien qui utilise une base de donnée FileMaker, comme plusieurs d'entre nous l'ont fait pendant des années, force est de constater le caractère 'naïf' de l'utilisation d'un outil dont toute la puissance découle précisément de l'application des méthodes de modélisation. Combien d'entre nous ont dû relire systématiquement et recoder 'à la main' des données textuelles stockées dans des fiches individuelles, tout simplement parce qu'ils ignoraient l'existence des systèmes sémantiques qui font tourner depuis des décennies les bases de données du monde entier ? Pourquoi ne pas les adopter nous-mêmes pour rendre plus efficace notre travail et plus conforme aux standards actuels ?

Nous nous limiterons ici à présenter le niveau conceptuel de la modélisation des données qui permet, grâce à la réalisation d'un MCD, de construire une sémantique apte à transformer les connaissances extraites des sources en données structurées correctement. Ce premier niveau est essentiel car la structuration concrète de la base de données dans un MLD (représentant les tables, les champs, etc.) découle du MCD en appliquant à la sémantique qui a été définie au premier niveau les règles appropriées et systématiques qui font partie de la méthode Merise. L'attention de l'historien doit donc porter principalement sur le MCD car celui-ci explicite et documente les découpages sémantiques opérés pour rendre compte de la réalité historique, en permettant ainsi aux chercheurs de comprendre la structure des données et de communiquer entre eux sans ambiguïtés. C'est la condition indispensable pour qu'on puisse ensuite extraire les données de la base en les recomposant par des requêtes appropriées en vue de leur représentation visuelle et de la production de nouvelles connaissances.

9 Pour une synthèse des concepts essentiels de la méthode Merise, notamment pour ce qui concerne la sémantique d'un MCD, cf. Audibert, op. cit., chapitre 2, qui est une excellente introduction aux fondements de cette méthode.

Le 'modèle conceptuel des données' : un exemple

Le principe fondamental du MCD réside dans le découpage d'objets sous forme d'« entités » qu'on met en relation par des « associations ». Les entités sont des ensembles (ou classes) d'individus (ou instances) qui possèdent les mêmes propriétés[10]. Par exemple, on peut créer une entité 'acteur' comprenant tous les humains. Les individus appartenant à cette classe sont rassemblés car ils possèdent des propriétés communes, dites également attributs : une propriété 'nom', une propriété 'sexe', une date de naissance, un lieu de naissance, etc.. Pour représenter les entités, on utilise des rectangles : dans la partie supérieure, un substantif au singulier indique le nom de la classe d'individus, c'est-à-dire le nom de l'entité, tandis que dans la partie inférieure on liste les propriétés ou attributs qu'on souhaite retenir pour cette entité (cf. figure 1.1).

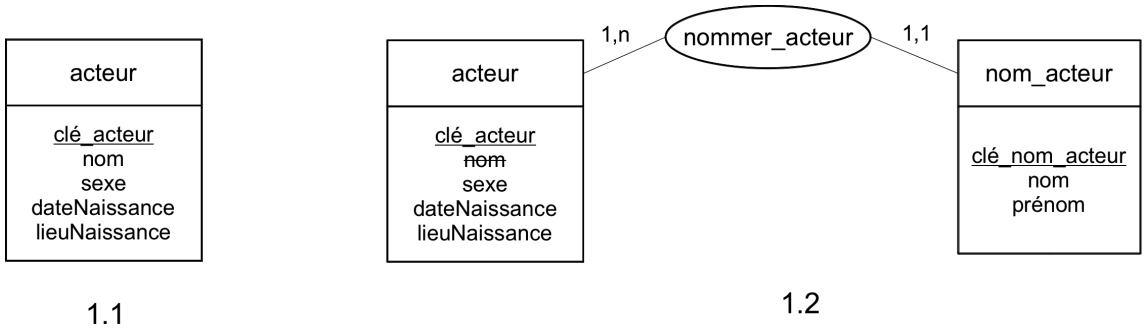

Figure 1.

L'attribut souligné représente la clé de l'entité 'acteur', c'est-à-dire une propriété qui a une valeur unique et distincte pour chaque individu et qui permet donc de l'identifier. On pourrait utiliser comme clé le nom de l'acteur, ou le numéro de sécurité sociale, mais ce faisant on ne pourrait ni traiter les homonymes, ni les acteurs d'autres époques ou pays. On a donc généralement recours à des entiers pour identifier les individus d'une entité.

Chaque entité possède, nous l'avons dit, une série d'attributs : nom, sexe, date de naissance, lieu de naissance, etc. Il importe de souligner que, dans le MCD, l'entité ne représente pas un individu mais une classe d'individus : sur un MCD, on ne représente jamais les individus ! Par principe, chaque propriété doit être possédée par chacun des individus qui composent la classe, du moins virtuellement, car elle peut ne pas être connue. Elle possède un domaine de valeurs, c'est-à-dire un ensemble de valeurs possibles qui peuvent être de différents types : chaînes de caractères, entiers, décimaux, énumérés, etc. La valeur d'un attribut doit être unique pour chaque individu. Tel est le cas du sexe – et encore –, mais pas celui du nom car une personne peut avoir plusieurs noms puisque, par exemple, le prénom d'usage n'est pas forcément celui de l'état-civil, ou selon les sources, l'orthographe du nom peut varier. Il est donc préférable de transformer le nom de l'acteur en une nouvelle entité, appelée 'nom_acteur', et de l'associer à l'entité acteur (fig.1.2). L'association est représentée par une ellipse et nommée par un verbe, alors que les

10 On pourrait plus proprement parler de « types-entités » pour les classes d'individus en réservant le terme d'entité aux individus. Nous suivons ici le langage courant qui utilise entité pour classe d'individus. De la même manière, on aurait pu parler de « types-associations », cf. Audibert, op. cit., 29-30.

entités sont nommées par des substantifs.

Les chiffres représentés sur les deux 'pattes' de l'association 'nommer_acteur' indiquent la cardinalité. Pour la connaître, on se place dans la perspective de l'une des entités reliées à l'association, par exemple l'acteur, et on se pose la question : combien de fois un individu représenté par cette entité pourra intervenir dans cette association ? La cardinalité minimale est représentée par le chiffre avant la virgule (dans ce cas : 1), la maximale par le deuxième chiffre (dans ce cas : n). Le MCD de la figure 1.2 exprime clairement notre choix : un acteur entrera en association avec au moins un nom, et pourra en avoir un nombre indéfini (cardinalité = n), tandis qu'un nom n'entrera en relation, ne nommera qu'un et un seul acteur (cardinalité minimale = maximale : 1,1). Cette sémantique sera décisive pour l'informaticien qui va construire la base de données : par exemple, il imposera dans l'interface qu'on mette au moins un nom à un acteur avant de pouvoir renseigner d'autres propriétés.

Si la valeur multiple d'une propriété pour un individu (le nom dans notre exemple) peut cacher une entité distincte, le fait qu'on rencontre la même valeur d'une propriété pour plusieurs individus peut également cacher, sous certaines conditions, une entité indépendante. Tel est le cas de la propriété 'lieu de naissance' : il est préférable de le traiter comme entité distincte, 'lieu', et de l'associer à l'acteur grâce l'association 'naître' (fig. 2).

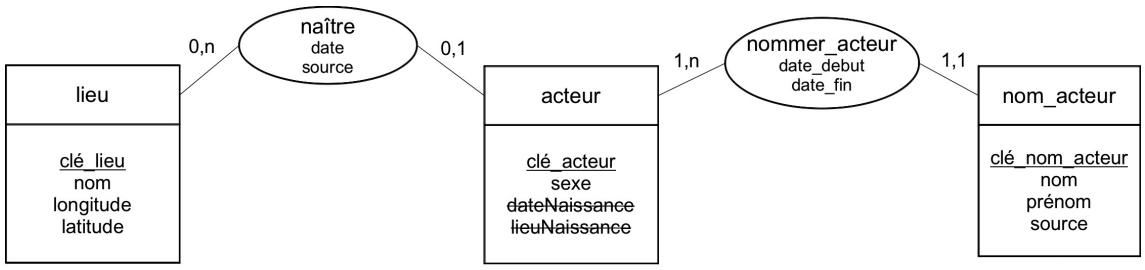

Figure 2.

La nouvelle entité 'lieu' pourra elle aussi disposer de propriétés, par exemple des coordonnées géographiques renseignant sa localisation et permettant de situer le lieu sur une carte. On déplacera également la propriété 'date' vers l'association 'naître' car elle est un attribut de la 'naissance'. La cardinalité 0,1 sur la patte droite de cette association signifie qu'un acteur ne peut être associé qu'à un lieu de naissance (on ne naît qu'un fois et dans un lieu unique !) ou à aucun (si son lieu de naissance est inconnu) ; la cardinalité 0,n sur la patte gauche exprime le fait que, toujours dans le contexte de l'association 'naître', un lieu peut n'être associé à aucun acteur (par exemple si aucun des acteurs du corpus étudié n'y a vu le jour) ou à plusieurs (lorsque plusieurs acteurs du corpus y sont nés). De plus, une propriété « source » sous forme de texte relatera le ou les documents permettant de connaître cet l'événement. Puisqu'il est possible d'ajouter des propriétés aux associations, on pourra retenir les dates de début et de fin de l'utilisation du nom d'un objet, comme le montre le cas de l'association « nommer acteur » : si un acteur a pris un nom en religion, puis a été élu pape, tel Sixte Quint, on pourra ainsi suivre l'évolution chronologique de ses trois noms, tout un ajoutant à l'entité nom-acteur une propriété

'source' qui permet de spécifier l'origine de la connaissance de chaque nom.

On pourrait aussi connaître différentes sources évoquant la naissance d'un même acteur à des dates ou des lieux différents. Dans ce cas, on pourra revoir la cardinalité de l'association 'naître' du côté de l'entité 'acteur', en admettant une cardinalité maximale 'n' : on pourrait ainsi renseigner plusieurs 'naissances' pour un acteur, comportant un 'sourçage' différent, et on indiquerait grâce à l'attribut booléen 'si_standard' (domaine de valeurs: vrai, faux) celle qui est à retenir (fig.3).

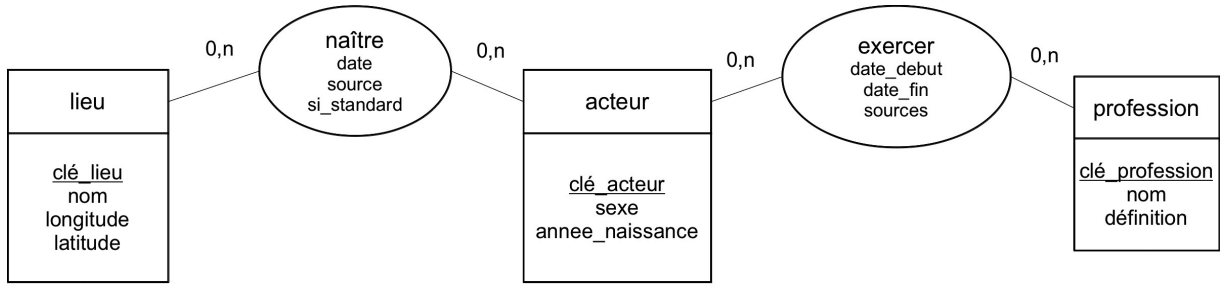

Figure 3.

De la même manière, on pourrait traiter les professions d'un acteur en les associant à ce dernier et en indiquant les dates de début ou de fin de la profession, ainsi que la ou les sources. Tout autre événement, segment de carrière, mais aussi relation entre acteurs, expression d'opinions politiques, religieuses, etc. pourra être traité de la même manière.

On reconnaît immédiatement la puissance de la sémantique qu'on peut mettre en place grâce à l'utilisation d'un MCD, pour l'étude d'une population composée de milliers d'acteurs : le fait d'avoir décomposé et saisi l'information sous cette forme permettra, grâce à des interrogations simples de la base de données, non seulement d'étudier les professions exercées en fonction du lieu de naissance ou de la génération, mais aussi de croiser ces données avec toute autre information saisie. Il permettra également d'en analyser les aspects quantitatifs, de réaliser des cartes avec les logiciels de systèmes d'information géographique (SIG), ou encore d'appliquer les outils d'analyse de réseaux, d'analyse des séquences ou d'*event history analysis*[11]. Tout historien connaissant la méthode Merise pourra comprendre la sémantique spécifique mise en place pour stocker ces données et les réutiliser à son tour.

Toutefois, si le potentiel de la méthode apparaît clairement, son application soulève de nombreuses questions auxquelles il appartient à l'historien, selon notre conviction, de répondre car elles relèvent du cœur même de sa méthode. Une première question est liée au 'découpage' des objets historiques : comment tracer les limites entre eux ? A quel niveau d'abstraction se situer en créant des classes d'individus ? Ensuite, se pose la question de leur identification : comment traiter les homonymies, les incertitudes que laissent les sources ? Et encore : quelle est la fonction des attributs des objets historiques ? Doivent-ils servir à la simple identification des individus ou au stockage d'une partie de l'information ? Car, comme le montrent les exemples des figures 2 et 3, il est beaucoup plus souple et efficace —donc préférable— de stocker les

11 Pour une introduction à ces techniques, cf. l'ouvrage fort utile de Cellier/Cocaud cité ci-dessus et surtout Lermercier, Claire / Zalc, Claire, Méthodes quantitatives pour l'historien, La Découverte, coll. Repères, Paris, 2008 à compléter par le site lié au livre http://www.quanti.ihmc.ens.fr/ (consulté le 31 janvier 2012).

informations historiques dans les associations entre objets que dans les propriétés des objets.

Et, enfin, que faut-il stocker dans la base des données : les contenus de multiples sources portant sur une même information, avec toutes leurs variantes et nuances, ou une sorte de 'synthèse' des connaissances disponibles qui opère un choix entre les variantes, en appliquant la critique historique ? C'est l'alternative qu'illustre la figure 3. En effet, un œil expert n'aura pas manqué de relever que le sens des deux associations dans le MCD n'est pas le même : à gauche il s'agit de rendre compte de plusieurs variantes de lieu ou de date pour le même événement « naissance » ; à droite, on saisira plusieurs segments de carrière de type « exercice d'une profession », déjà 'synthétisés' et donc distinct l'un de l'autre[12]. Si on voulait retenir pour chacun d'entre eux toutes les variantes qu'on retrouve dans de multiples sources, il faudrait complexifier le MCD. Nous nous attellerons à cette tâche par la suite mais revenons d'abord à la question des objets historiques.

Le 'découpage' et la fonction des objets historiques

Que faut-il entendre, dans la sémantique de SyMoGIH, par 'objet historique' ? Un objet historique est une entité qui regroupe une collection ou classe d'individus qui présentent les mêmes propriétés et partagent la même 'essence'. Spontanément, le premier exemple qui vient à l'esprit est celui des êtres humains : on aurait certes pu construire des entités en fonction d'un rôle et on aurait ainsi eu une entité 'marchand', une entité 'épouse', une entité 'député', etc. ; mais on aurait ainsi confondu l'essence de l'individu et l'une de ses propriétés, sa position sociale à un moment donné, en construisant un système sémantique voué à l'échec. Cet exemple quelque peu paradoxal illustre clairement le problème, bien plus délicat pour des cas moins évidents, et que nous avons déjà rencontré à propos du lieu de naissance : il faut distinguer entre l'objet lui-même, dans son essence, et les propriétés qui le caractérisent et qui ne sont, en réalité, que d'autres objets qui sont associés au premier. Mais qu'est-ce donc que l'essence d'un objet historique qui permet de le constituer en entité dans un MCD ?

La réponse à cette question ne peut pas être donnée à priori mais doit être fournie en adéquation avec le travail concret de l'historien. Suite aux discussions et aux expériences de ces dernières années, nous sommes arrivés à un découpage qui constitue onze objets : acteurs, acteurs collectifs, objets matériels, objets abstraits, caractères sociaux, bibliographie, unités documentaires, objets digitaux, ressources web, lieux, immeubles. Ces objets sont spécifiés par des types et des classes qui permettent de les répartir plus finement. Le type opère une séparation exclusive entre individus, en créant des sous-ensembles sans intersection, car un individu ne peut appartenir qu'à un seul type. En revanche, les classes permettent de créer des sous-ensembles avec intersections puisque un individu peut appartenir à plusieurs classes. Cette distinction s'exprime dans la différente cardinalité sur les pattes de ces deux associations du côté de l'entité « acteur » (Figure 4.).

12 Pour un exposé de cette problématique distinguant, avec une autre terminologie, entre contenu d'une source et reconstitution par l'historien de l'information, voir Merry, Mark, Databases for historians (2011), chapitre C : http://training.historyspot.org.uk/mod/book/view.php?id=75&chapterid=133 (consulté 30 janvier 2012, inscription nécessaire).

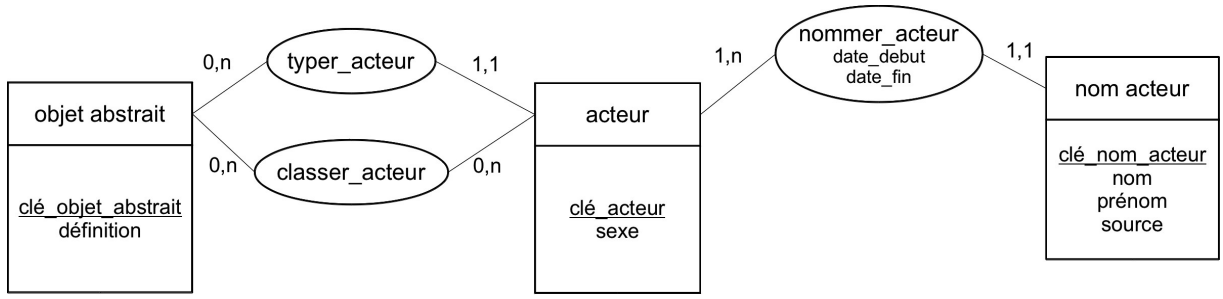

Figure 4.

Dans le système sémantique de SyMoGIH, l'entité « Acteur » comporte trois types : « être humain », « personnage fictif », « animal ». On aurait pu séparer les personnages fictifs figurant dans un récit en les distinguant des acteurs historiques ayant eu une existence réelle et en créant deux entités distinctes. Mais il a apparu plus opportun de ne pas multiplier le nombre d'objets historiques, et donc les entités, et de n'en créer qu'un. L'entité « Acteur » regroupe ainsi tous les individus dont l'essence consiste dans le fait d'agir, activement ou passivement, dans les récits et autres sources à notre disposition, sous forme d'individualités.

Ces acteurs individuels se distinguent d'une autre classe d'acteurs : ceux qui agissent collectivement et que nous avons réunis dans une entité appelée « Acteurs collectifs ». Cette entité regroupe les personnes morales, les institutions politiques, scientifiques, religieuses, les familles, etc., c'est-à-dire tout type d'acteur qui intervient dans les sources en termes de collectivité. Certes on aurait pu regrouper ces deux entités, acteurs individuels et collectifs, en une seule mais il a paru plus opportun de les distinguer, leur 'essence' étant visiblement différente. Un préfixe sémantique ajouté à l'identifiant numérique des individus permet d'identifier l'entité à laquelle ils appartiennent : par exemple, René Descartes sera identifié en tant que 'Actr134', tandis que l'Académie des sciences de Paris aura la clé 'CoAc6' dont le préfixe correspond à la dénomination anglaise de l'entité : « collective actor ».

S'il n'y a donc pas de principe absolu permettant de découper les objets, il est en revanche clair qu'une construction 'large', se limitant à un 'essence' très générique, permet de préserver leur caractère 'objectif', ou neutre, en renvoyant leur définition précise aux types et aux classes qui les caractérisent et, surtout, à la collection d'informations disponibles sur eux qu'on trouve dans la base de données. En effet, les objets historiques ont principalement une fonction analogue à celle des notices d'autorités des bibliothèques[13] : attribuer un identifiant à un individu –identifiant qui sera utilisé dans l'ensemble du système d'information– et fournir suffisamment d'éléments pour l'identifier grâce à une notice qui en décrit les caractéristiques essentielles, un type, des classes. En revanche, toute propriété de l'objet, ou information le concernant, doit être stockée séparément, selon les méthodes que nous verrons, mais ne doit pas être introduite en tant que propriété de l'objet lui-même. Il existera donc dans la base de données une collection d'informations associées à cet objet grâce à son identifiant qui représentent le stockage effectif

13 Cf. l'exemple de la notice 'Descartes, René ', pour les autorités de la BNF : http://catalogue.bnf.fr/ark:/12148/cb11899775j/PUBLIC (consulté 30 janvier 2012) ou pour les autorités du Sudoc : http://www.idref.fr/027287165 (consulté le 30 janvier 2012).

des connaissances le concernant : un objet historique n'existe donc réellement, pour ainsi dire, que par les informations qui s'y rapportent et grâce à ses associations aux autres objets.

Concernant les autres objets définis dans SyMoGIH, mentionnons tout d'abord l'entité « Objet abstrait » : cette entité englobe tous les concepts génériques ou individuels qui ne correspondent à aucun objet concret. S'il s'agit d'un concept générique, tel la notion de livre, d'hérésie ou de pouvoir, on aura affaire à un individu de type 'générique' ; si on traite en revanche une doctrine précise, telle la première loi de Kepler, ou la norme ISO 639 qui définit les codes internationaux des langues, il s'agira d'un objet abstrait de type 'individualisé'. On pourra ensuite récolter toute une série d'informations sur la genèse et la diffusion d'une doctrine, ou d'une loi, et les retrouver aisément, ainsi que leurs auteurs, grâce à leur association avec l'objet abstrait correspondant. Les objets abstraits ont une fonction clé dans le système puisqu'ils sont utilisés en tant que types et classes des autres objets (cf. Fig. 4) : ce choix confère une grande souplesse au système d'information car il permet d'expliciter le sens de chaque terme utilisé pour le classement, en l'associant à son tour à des objets et des information, et d'en créer à loisir de nouveaux.

De l'entité « Objet abstrait » ont été détachés les individus qui ont été regroupés dans l'entité « Caractère social »: il s'agit de l'ensemble des professions, fonctions administratives, politiques et religieuses, qualités personnelles, etc. qui, tout en étant en elles-mêmes des concepts ou réalités symboliques, méritaient un traitement spécifique en raison de leur importance pour le fonctionnement et la compréhension de la vie sociale. Une autre distinction entre objets similaires a été introduite entre les lieux en tant que réalités géographiques entendues au sens de localisation ou de surface, et les immeubles, comportant une dimension de verticalité due à la construction. Cette distinction s'avère fort utile pour le traitement des objets géographiques dont la nature peut être ainsi plus facilement distinguée, quitte à être articulée plus finement par des types : lieu habité, territoire, élément géographique naturel, adresse, pour l'entité « Lieu » ; bâtiment, infrastructure, ensemble de bâtiments, etc. pour l'entité « Immeuble », ainsi que par de nombreuses classes ajoutées ultérieurement.

Enfin, un soin particulier a été apporté au découpage des entités qui permettent le sourçage des informations et le stockage sous forme digitale : Objet digital, Bibliographie, Objet matériel. En particulier, la distinction entre ces deux dernières entités mérite attention car il a fallu trouver un critère simple et efficace pour démêler une matière relativement intriquée : étant donné que « Bibliographie » correspond à tout objet existant en plusieurs exemplaires identiques, chaque individu de cette entité correspond donc à une classe d'exemplaires ; en revanche l'entité « Objet matériel » regroupe des objets réels et individualisés, tel un volume d'archives, une monnaie conservée dans un musée, un tableau appartenant à une collection privée. Toutefois, cette construction des entités ne permet pas de traiter l'identification d'un chapitre dans un livre, ou d'une entrée précise dans un registre de baptêmes, identification qui seule aurait permis de retrouver aisément toutes les connaissance extraites d'un même passage. Une nouvelle entité a donc été introduite, l'« Unité documentaire », qui permet de documenter le découpage de l'unité textuelle, iconographique, etc. à partir d'une source, tout en indiquant la référence exacte et le sens de cette construction.

L'extraction des Contenus

C'est à partir de cet exemple, et sur la base des concepts introduits jusqu'ici, que nous aborderons maintenant les questions décisives qui se posent à l'historien lorsqu'il se propose d'imaginer un système généraliste de stockage des connaissances.

Prenons à titre d'exemple la transcription d'une entrée de registre de baptême : « Le dit 11 novembre 1721 j'ai baptizé Marguerite née hyer fille de Sieur Jean Claude Chirat marchand et de demoiselle Esparron son épouse. Parrain Sieur Jean Baptiste Esparron aussy marchand. Marraine demoiselle Marguerite Colaud fille »[14]. Une première connaissance qu'on tire de ce texte est logiquement l'événement 'baptême' : celui-ci met en relation un ensemble d'acteurs, chacun intervenant à titre différent. Comment représenter cette connaissance ? On peut en faire une entité « Baptême » et regrouper dans cette classe tous les 'individus' de type événement-baptême (Figure 5).

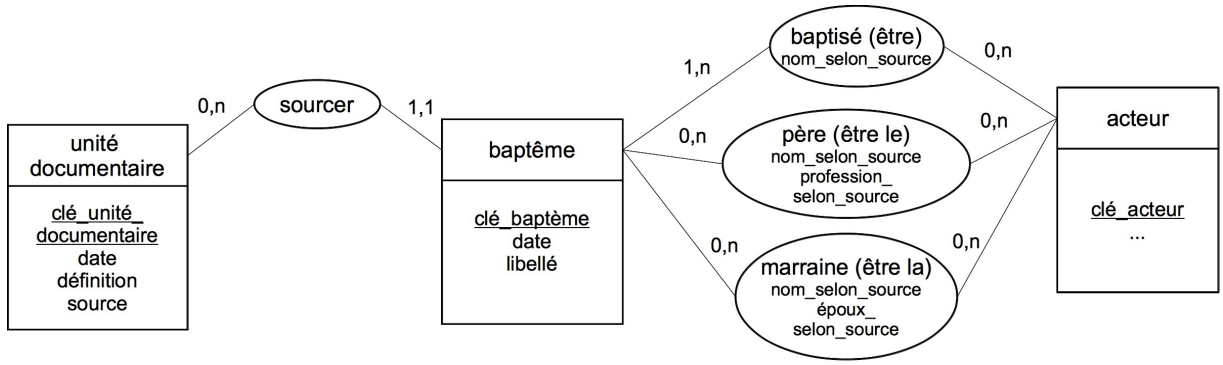

Figure 5.

Cette entité sera associée à tous les acteurs qui participent au baptême : ils sont représentés par l'entité « Acteur » tandis que la fonction, le *rôle* que chacun d'eux a dans l'événement est spécifié par le nom de l'association. Nous n'avons représenté, à titre d'exemple, que trois rôles : si on appliquait effectivement cette méthode, il faudrait représenter les rôles de tous les acteurs qui interviennent au baptême. Une propriété « nom selon la source » a été ajoutée qui permet de transcrire la dénomination exacte de l'acteur : comme l'association de l'objet se fait concrètement grâce au report de la clé de l'acteur (l'identifiant qui figure dans sa notice d'autorité) dans la table qui représente l'association, on pourra ainsi retenir l'orthographe précise du nom tel qu'il figure dans la source. Le sourçage de cette connaissance se fait grâce à l'association à l'Unité documentaire dont elle est tirée et qui représente, dans ce cas, l'entrée du registre. Comme le montre la cardinalité de l'association « sourcer » du côté de l'événement « baptême », cette association est obligatoire (cardinalité minimale = 1) et limitée à une seule unité documentaire (cardinalité maximale = 1) : celle-ci est donc l'unique source de cette connaissance.

Dans le MCD de la figure 5, nous avons également ajouté dans les associations des acteurs – à titre purement hypothétique – des propriétés permettant de stocker d'autres connaissances que peut fournir le document : la profession du père, le nom du mari de la marraine, etc. Toutefois,

14 Archives municipales de Lyon, 1GG155, Registre BMS, paroisse Saint-Nizier, f°100.

comme nous l'avons déjà indiqué, il est préférable de transformer ce type de propriétés en objets qu'on va ensuite associer. Par exemple, si on veut saisir la profession des participants aux baptême, il vaut mieux créer une entité « Exercice de la fonction » à laquelle on va associer l'acteur, la profession exercée sous forme d'un objet « Caractère social », éventuellement le lieu d'exercice de la profession.

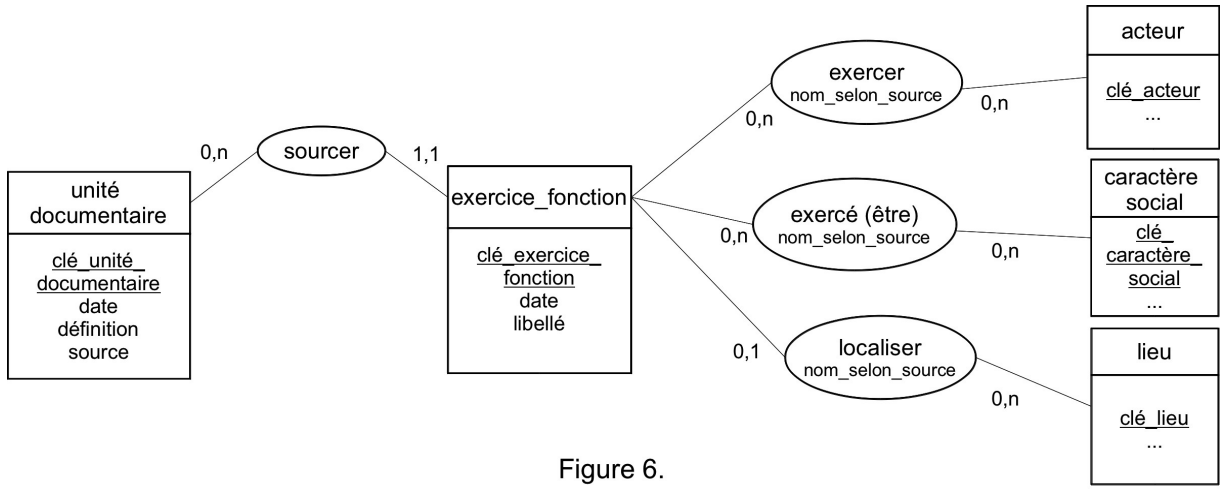

Figure 6.

On remarquera que la terminologie choisie pour définir ces entités est volontairement 'large' : le but est de mettre en relation ces connaissances et ces objets sans en définir à priori l'essence. Nous nous limitons donc à les rendre 'identifiables' grâce à leurs clés respectives, puis à les associer entre eux. Leur dénomination exacte selon la source sera retenue en propriété de l'association et/ou sous forme d'un libellé de l'entité « Exercice de la fonction » qui permet de décrire sous forme de texte le lien sémantique qui subsiste entre les objets associés. On relèvera aussi que nous avons décomposé les connaissances fournies par le document en unités atomisées autant que possible : dans la sémantique de SyMoGIH, les « Contenus » représentent ainsi des unités de connaissance atomisées issues d'une et une seule unité documentaire.

Toutefois, comme c'est le cas pour les objets historiques, il est impossible de multiplier indéfiniment les entités regroupant les différents types de contenu : baptême, exercice d'une fonction, etc. Par conséquent, il faut imaginer un modèle plus abstrait, une sorte de méta-modèle capable d'englober l'ensemble des unités de connaissance dans un modèle conceptuel synthétique. Si on reprend l'exemple de la Figure 6, il se présentera sous cette forme (Figure 7) :

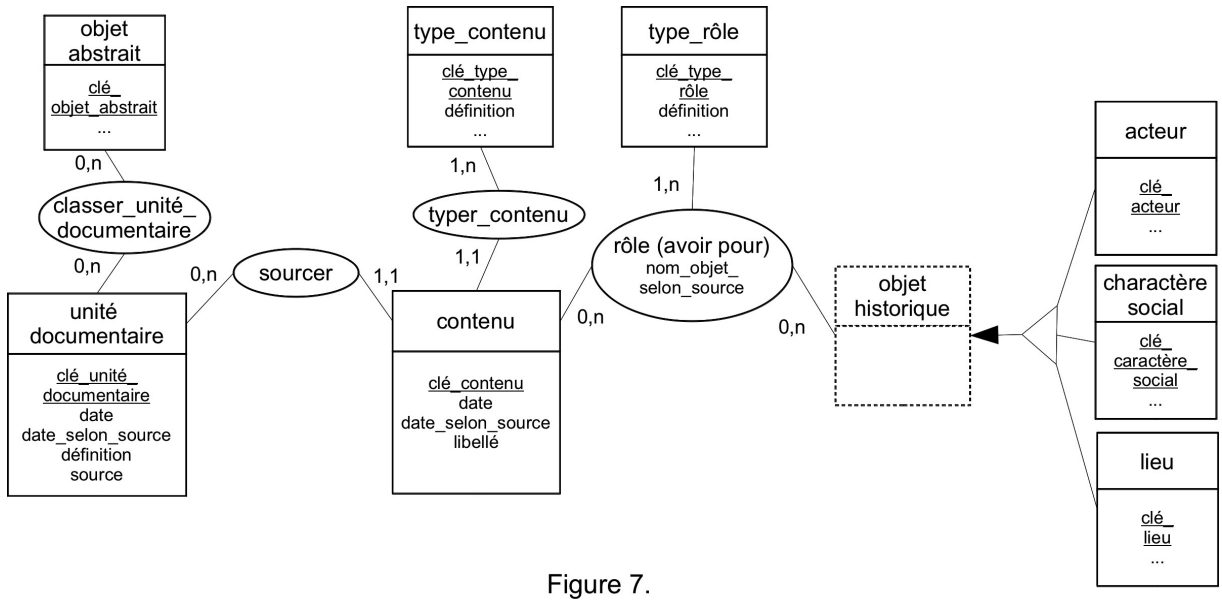

Figure 7.

Au centre du MCD, l'entité « Contenu » représente l'ensemble de toutes les unités de connaissance qu'on tire des documents. Une entité « Type de contenu » permet de spécifier pour chaque individu ou instance de l'entité « Contenu » de quel type de connaissance il s'agit : baptême, exercice d'une fonction, etc. L'association « rôle » – en fait « avoir pour rôle » pour respecter la méthode Merise qui décrit les associations en utilisant des verbes – rattache les objets au Contenu en question tout en spécifiant, grâce à l'association à l'entité « Type de rôle », à quel titre un objet intervient dans la connaissance.

L'entité abstraite « Objet historique » ne possède pas d'individus propres mais représente les autres objets. En d'autres termes, chaque objet peut se trouver associé au « Contenu » via un rôle : la reconnaissance de l'essence de l'objet se fera grâce à la partie sémantique de l'identifiant, « Actr » pour les acteurs, « SoCh » pour les caractères sociaux (nom anglais de l'entité : *Social Characteristic*), « NaPl » pour les lieux (*Named Place*), etc. Tout objet défini dans le système d'information pourra être associé de la même manière à un Contenu. Quant aux Unités documentaires, elles pourront être classées en utilisant les Objets abstraits : on pourra de cette manière retrouver toutes les entrées de registre de baptême, les chapitres d'ouvrage, etc. Concernant la date, elle sera retenue dans une forme standardisée, tandis qu'un champ « Date selon la source » permettra de saisir, si on le souhaite, le libellé exact de la source, exprimé éventuellement dans un style ou dans une ère différente. Les Contenus pourrons hériter la date de l'Unité documentaire dont ils sont tirés ou posséder une date propre.

Cette manière de construire la sémantique permet de stocker les unités de connaissance de type « Contenu » conformément aux requis de la méthode historique. L'atomisation et la simplicité d'un modèle conceptuel qui se limite à exprimer des associations entre objets, tout en explicitant le sens de leur relation, permet de séparer le niveau de la problématique de la recherche de celui du stockage des connaissances : les réponses au questionnement de l'historien se feront grâce à des requêtes qui vont recomposer, en fonction de la problématique, des données atomisées, munies du minimum possible de connotation sémantique et préservant, sous forme de

texte, la formulation originale de l'objet. La documentation précise de tout Type de contenu et Type de rôle permet à l'ensemble des utilisateurs du système de comprendre la sémantique ayant produit les données, même s'ils n'en sont pas les producteurs, et d'en tirer profit, ou alors de créer de nouvelles définitions et de faire évoluer le système. En effet, la récolte des données se fera toujours, inévitablement, en fonction d'une problématique de recherche : par conséquent, il est illusoire de vouloir extraire de façon 'neutre' l'ensemble des connaissances que contient un texte. La sémantique de SyMoGIH permet d'expliciter l'opération accomplie lors de l'extraction des Contenus et d'en reconnaître éventuellement les limites pour une autre problématique de recherche. Dans ce cas, de nouveaux types d'unités de connaissances pourront être produits afin d'extraire de nouveaux Contenus des mêmes Unités documentaires.

Cependant, en stockant les connaissances sous forme d'associations de clés reportées à partir de dictionnaires d'autorités ne risque-t-on pas de perdre la 'substance' du texte de la source, dont la richesse excède les capacités de tout codage ? Le stockage du nom précis des objets dans une propriété de l'association elle-même, c'est-à-dire dans le « Rôle », permet de réviser le codage si on décèle des cas d'homonymie, en insérant la clé de l'acteur correct. Le cas des professions semblerait être plus délicat, étant donné la richesse des dénominations qu'on rencontre dans les sources. Là aussi, la sémantique retenue ne fait que reproduire le travail quotidien de l'historien : d'une part, la transcription du libellé précis de l'objet permet de préserver un rapport fidèle à la source ; d'autre part, la création d'un objet « Caractère social » et son association au Contenu via un Rôle permet de stocker l'interprétation, le classement que le spécialiste produit lorsqu'il analyse le texte. Il pourra ensuite revenir sur son choix par une simple modification de la clé de l'objet dans le Rôle, en créant éventuellement un nouveau Caractère social plus pertinent.

Cette manière de procéder soulève toutefois un problème quant à la technologie adoptée pour stocker les Contenus, c'est-à-dire la base de données. En effet, la transcription des noms de tous les objets mentionnés dans le texte revient à reproduire virtuellement le texte de la source tout en perdant sa structure originale. Si on se limite à extraire quelques unités de connaissance d'un texte long, on peut faire l'économie d'une transcription complète. Mais si on opte pour une analyse détaillée de la source, il vaut mieux transcrire le document en entier et y ajouter un balisage XML. Pour ce faire, on utilisera avec profit la sémantique de la *Text encoding initiative* (TEI) comme l'illustre l'exemple d'un balisage possible de l'extrait d'un registre de baptêmes reproduit ci-dessus (Figure 8).

```
<div ana="#baptême">
  <p> Le dit <date when="1721-11-11">11 novembre 1721</date> j'ai baptizé
    <persName  ana="#baptisé" key="Actr1"><forename>Marguerite</forename></persName> née hyer fille de
      <persName key="Actr2" ana="#père">Sieur <forename>Jean Claude</forename>
      <surname>Chirat</surname><roleName key="SoCh1" ana="#fonction_exercée">marchand</roleName></persName>
    et de <persName key="Actr3" ana="#mère">demoiselle <surname>Esparron</surname> son épouse</persName>.
    Parrain <persName key="Actr4"  ana="#parrain">Sieur <forename>Jean Baptiste</forename>
       <surname>Esparron</surname> aussy <roleName key="SoCh1" ana="#fonction_exercée">marchand</roleName></persName>.
    Marraine <persName key="Actr5"  ana="#marraine">demoiselle <forename>Marguerite</forename>
      <surname>Colaud</surname><roleName key="SoCh2" ana="#état">fille</roleName></persName>.</p>
</div>
```

Figure 8.

On reconnaît dans ce fragment des balises qui, selon la sémantique de la TEI, indiquent la

structure de texte, pour division, <p> pour paragraphe, ainsi que des balises sémantiques qui spécifient le sens des objets nommés qu'on rencontre dans le texte, <persName> pour l'ensemble de la dénomination d'une personne, <roleName> pour indiquer une fonction exercée dans la société, etc. On peut ensuite ajouter au balisage les attributs opportuns : @key pour introduire l'identifiant issu de la notice d'autorité de la base de données ; @ana pour spécifier un type d'analyse de l'objet en question – dans notre cas il s'agit respectivement du baptisé, du père, et ainsi de suite– tandis que l'ensemble de la portion de texte relate un baptême. On obtient ainsi un codage des unités de connaissances qu'on souhaite extraire de ce texte qui correspond à la même sémantique que celui produit par la base de données en utilisant les Contenus mais qui, en même temps, en préserve la structure textuelle. De plus, on pourra faire 'disparaître' les balises dans n'importe quel logiciel approprié –par exemple LibreOffice– pour lire le texte comme tel, ou on pourra le publier sur un site web en l'indexant avec les identifiants des objets mentionnés.

Le stockage des Informations

Le traitement des Contenus avec les deux technologies complémentaires que nous venons de présenter ne répond pas à tous les besoins des historiens. En effet, les connaissances ainsi stockées sont pour ainsi dire 'brutes' et fréquemment répétitives ou redondantes. A partir d'elles, le spécialiste opère spontanément une synthèse et une abstraction qui produisent un type de connaissances ayant une signification différente des Contenus et qui méritent d'être stockées comme telles : nous avons appelées « Informations » ce type différent d'unités de connaissance. Par exemple, à partir d'une série de mentions de l'exercice d'une même profession par un acteur, à des dates et par des sources différentes, on obtiendra une seule information qui indique que tel acteur a exercé telle profession de telle date à telle date. Cette nouvelle unité de connaissance, qui correspond au volet de droite de la Figure 3, doit être unique, par principe, dans le système d'information, même si elle est issue de plusieurs sources, puisqu'elle a une autre signification : dans ce cas, on passe d'une connaissance ponctuelle à un segment de carrière.

Ce n'est toutefois pas la durée qui fait la différence entre un Contenu et une Information mais le degré d'abstraction, le changement de perspective : le Contenu exprime une unité de connaissance telle qu'elle est fournie par un et un seul document ; l'Information opère une synthèse de plusieurs unités de connaissance, tranche parmi les variations possibles des sources – de date ou de contenu– et produit une connaissance nouvelle qui va permettre, tel le fragment d'une mosaïque, la reconstitution d'un 'monde historique'. Elle a une fonction et une signification analogue aux connaissances fournies par la notice d'un dictionnaire biographique : les connaissances sont soumises à l'opération de la critique historique, elles sont pour ainsi dire épurées, on n'en retient que l'essentiel, elles ne correspondent plus nécessairement à la structure ou au contenu des sources dont elles sont tirées. En même temps, le principe d'atomisation et la structure de la sémantique de SyMoGIH doivent être maintenues car elles garantissent la séparation entre le stockage des connaissances et leur exploitation pour répondre à un questionnement précis (Figure 9).

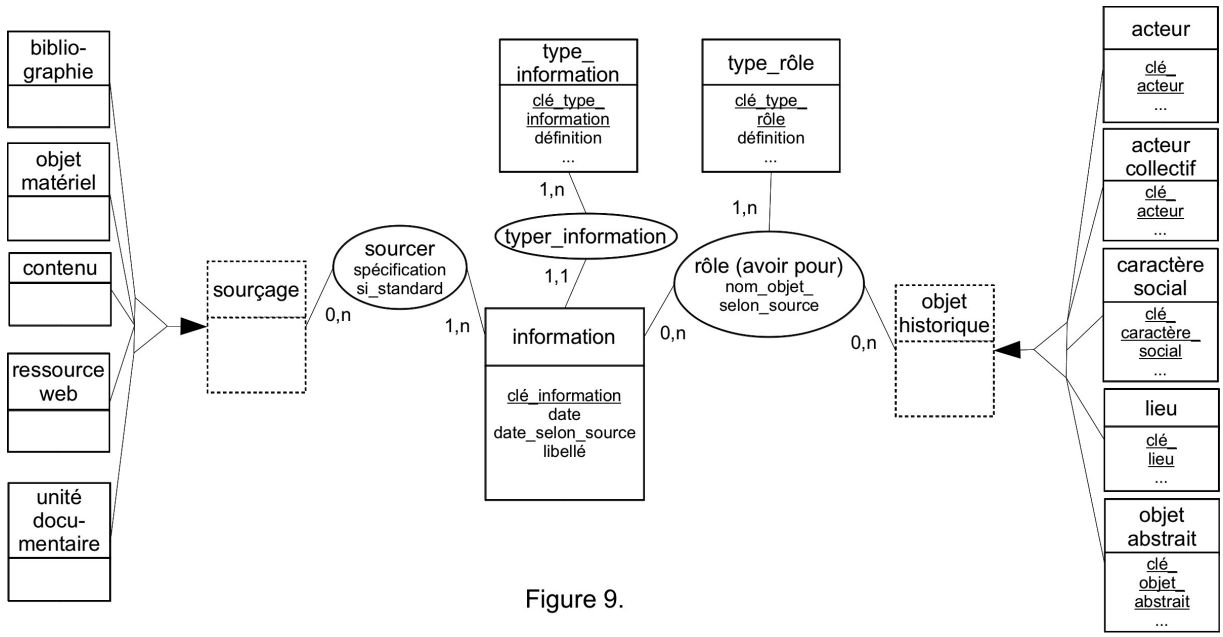

Figure 9.

On relèvera que la structure de la partie droite de la Figure 9 est identique à celle de la Figure 7. Mais le sens de la connaissance stockée est différent : il s'agit d'une synthèse critique obtenue à partir de connaissances qui sont virtuellement issues de plusieurs sources, comme le montre la partie gauche du MCD. L'entité abstraite « Sourçage » représente l'ensemble des entités qui peuvent intervenir pour sourcer une « Information ». La propriété « spécification » de l'association « sourcer » contient la référence précise d'où a été tirée l'information (page, feuillet, etc.). La cardinalité de la patte de cette association du côté de l'Information –1,n– montre qu'un sourçage au moins est obligatoire pour toute Information, ce qui correspond à l'un des requis de la méthode historique : indiquer l'origine de toute connaissance fournie. Insistons à nouveau sur le fait que cette cardinalité diffère de celle définie pour la patte droite de l'association 'sourcer' dans la figure 7 (1,1) : le Contenu doit et ne peut avoir qu'une source. Cette différence de cardinalité est un critère discriminant fondamental pour distinguer . Enfin, la propriété booléenne « si standard » permet d'indiquer, parmi les différentes sources possibles de l'Information, celle qui est à considérer en priorité.

Cette modélisation ne permet toutefois pas de rendre pleinement compte du processus de discernement critique qui a conduit au choix parmi les différentes sources pour fixer une connaissance de type Information. Aussi, un « libellé » ajouté comme propriété permet de reproduire l'Information sous forme de texte – dans sa dimension de synthèse, bien entendu, et non de reproduction du contenu la source – en indiquant sous forme littéraire le lien qui subsiste entre les objets associés à l'Information via les Rôles. Toutefois, ce MCD ne prévoit pas de stocker dans une forme exploitable systématiquement les composantes de l'Information, par exemple les données quantitatives qu'elle comporterait éventuellement. En ajoutant quelques éléments au MCD, on peut obtenir le résultat souhaité (Figure 10).

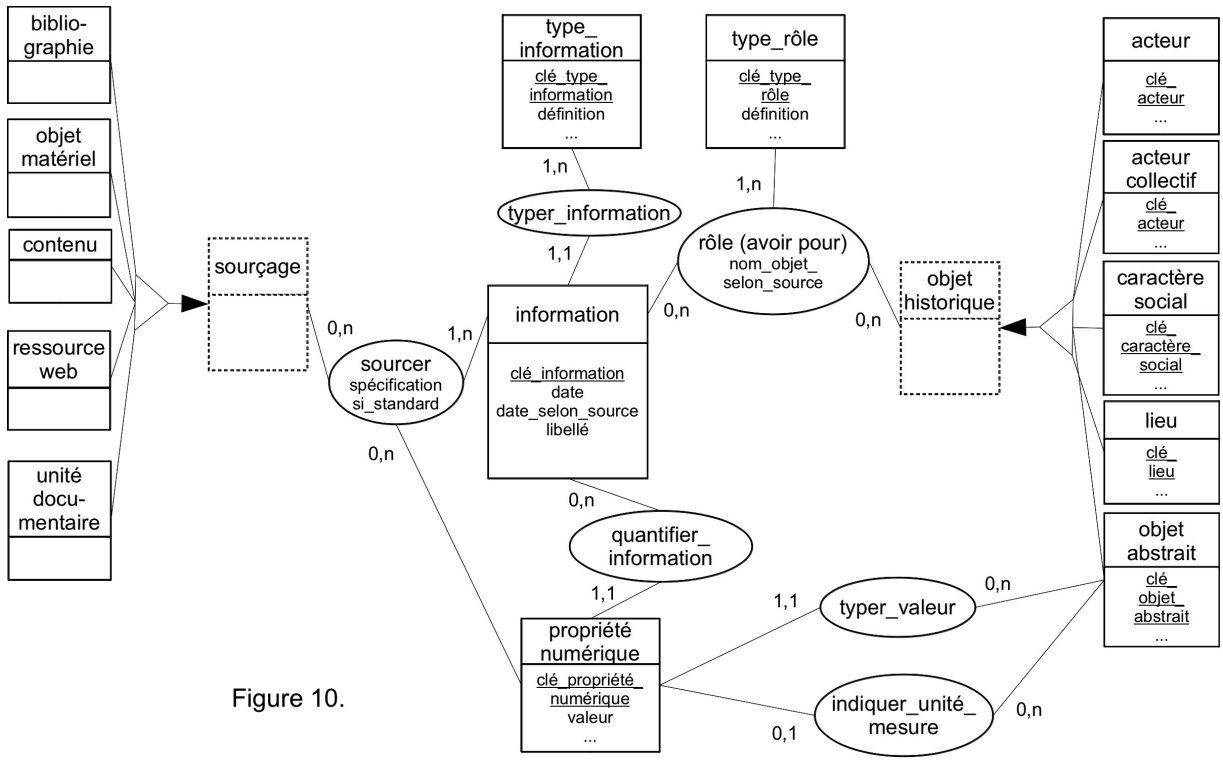

Figure 10.

L'entité « Propriété numérique » permet de stocker une valeur quantitative, par exemple le montant d'une vente, ou la longueur d'un pont. Des objets abstraits appropriés permettent de typer la valeur et d'en indiquer, le cas échéant, l'unité de mesure. Cette modélisation permet de saisir plusieurs occurrences de la même propriété – si les sources, par exemple, font état de montants différents pour la même vente – et, grâce à la propriété « si standard » de l'association « sourcer », de choisir, à l'issue du discernement critique du spécialiste, la valeur qui doit être retenu comme 'correcte'. A relever que des Contenus portant sur la même unité de connaissance peuvent être utilisés pour sourcer cette Information. Le même procédé peut également s'appliquer à des 'propriétés texte' qui pourront être ajoutées sous forme d'entités distinctes. Et enfin, les « Rôles » peuvent eux aussi être spécifiés par des propriétés texte ou numériques, permettant ainsi de retenir avec précision les propriétés relatives aux différents objets. Tout en restant dans le cadre d'un MCD relativement simple, il est ainsi possible de rendre compte des différentes opérations qu'accomplit l'historien depuis le dépouillement des sources jusqu'à la production de connaissances construites à partir d'un travail critique, en les stockant comme telles.

Concrètement, si on revient à l'exemple de l'extrait du registre des baptêmes, deux unités de connaissance pourraient être directement produites sous forme d'Informations : l'union des parents et la naissance de l'enfant. Il s'agit en effet de connaissances qui pourront sans doute être sourcées par d'autres documents et qui font partie du tissu essentiel indispensable à la reconstitution d'un monde historique – dont l'un des fondements réside dans la reconstitution des familles et des généalogies.

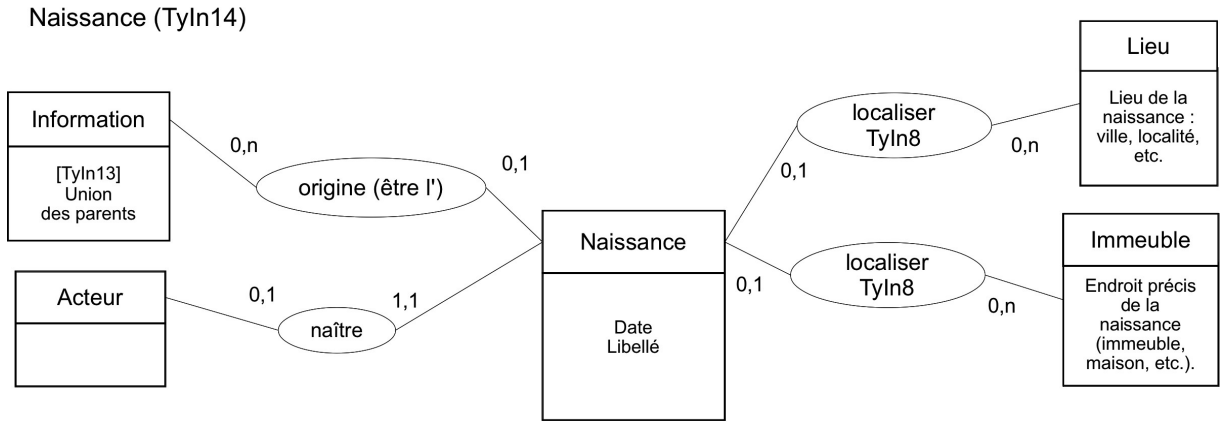

Figure 11.

La Figure 11 illustre le MCD propre à la « Naissance » qui représente un individu de l'entité « Type d'information ». Cette modélisation, assortie également d'un texte d'explication, indique à tous ceux qui participent au stockage collaboratif des connaissances de quelle manière coder une naissance, en lui associant un et un seul acteur qui naît, un lieu de naissance, éventuellement un immeuble ou une partie d'immeuble qui localisent l'événement. On peut également associer, en tant qu'origine de la naissance, l'information « Union des parents » : une information devient ainsi à son tour un objet qui peut être associé à d'autres objets. Cette structure sémantique correspond, dans un autre formalisme, au modèle GEDCOM utilisé par les généalogistes[15]. On relèvera aussi qu'on a préféré parler d'« Union » que de mariage pour permettre à ce Type d'information de rassembler tous les types d'unions, quitte à spécifier ensuite grâce à l'association à un Objet abstrait qu'il s'agit d'un mariage ou d'un autre type d'union.

Tous les Types de contenu et les Types d'information sont produits selon une même procédure : la modélisation est explicitée par un MCD assorti d'explications, puis elle est soumise à une réflexion critique collective qui s'efforce, lors de séances régulières de discussion, de valider et de perfectionner le système sémantique mis en place. La méthode de SyMoGIH réalise ainsi pour les données historiques le programme qui est formulé dans la définition d'« information » courante en informatique : « Élément de connaissance susceptible d'être représenté à l'aide de conventions pour être conservé, traité ou communiqué »[16]. En même temps, du point de vue de la méthode historique, elle confère à la modélisation selon la méthode Merise une dimension scientifique car elle l'intègre et l'adapte à ce domaine central de la recherche qui est celui de la production des données, susceptibles d'être ensuite analysées et interrogées pour produire les reconstitutions et les interprétations des historiens.

Le stockage des unités de connaissance historique : et après ?

La méthode Merise prévoit une série de règles et de procédés qui permettent, à partir des

15 http://fr.wikipedia.org/wiki/Norme_GEDCOM (consulté le 31 janvier 2012).

16 http://www.cnrtl.fr/definition/information (consulté le 31 janvier 2012).

MCD que nous avons construits, de créer les tables de la base de données qui reproduisent sous forme numérique le modèle conceptuel retenu et qui permettent le stockage des Contenus et des Informations. Une fois ce stockage opéré, comment exploiter ces données ? Une première étape indispensable consiste à explorer les données afin de repérer les éventuelles erreurs de saisie ou les incohérences de codage. A cette phase de nettoyage succède celle au cours de laquelle intervient pleinement la problématique de recherche. Après avoir été précisément formulées en langage naturel, à partir de la sémantique documentée par les MCD propres aux Types de contenus et Types d'informations retenus, les interrogations sont traduites dans les langages SQL, pour les bases de données relationnelles, et Xquery, pour les textes stockées en XML, langages dont la puissance d'interrogation est élevée[17].

Ensuite, moyennant éventuellement des recodages à la volée, des tris ou une restructuration des données en fonction de la problématique retenue, on pourra extraire les données, les exporter sous format de classeur ou de csv (texte séparé par virgule) et les introduire dans des logiciels de statistique, analyse de réseaux, généalogie, cartographie, etc. pour les visualiser ou les soumettre à de nouveaux traitements. Ce qui permettra de produire de nouvelles connaissances, impossibles à 'percevoir' en parcourant simplement les données stockées.

Des exemples concrets d'exploitation des données dépasseraient le cadre qui nous est imparti ici. Dans les faits, le projet SyMoGIH a été initié en 2007 par les deux auteurs de l'article, en collaboration avec Alexandre Giandou, François Robert et Loïc Bonneval, dans une petite équipe de collègues qui souhaitaient travailler à la mise en place d'une méthode collective de stockage de l'information historique. Elle a ensuite évolué au gré de la motivation des initiateurs et de l'intégration de nouveaux participants. Bernard Hours s'est activement impliqué et a su initier ses étudiants à cette aventure du stockage modélisé et collectif de l'information. Charlotte Butez, géomaticienne, a contribué de façon décisive à intégrer la dimension géographique dans le système grâce à la mise en place d'un gazetteer, c'est-à-dire d'un répertoire de lieux géolocalisés qui est devenu l'un des pivots du projet. Actuellement, une nouvelle application web est entrée en production grâce à l'engagement de deux informaticiens, Djamel Ferhod et Sylvain Boschetto.

Dans la communauté des utilisateurs qui fait vivre ce projet de mutualisation et de partage des données relevons la dizaine de travaux de master aboutis, les trois doctorats en cours, la vingtaine d'utilisateurs individuels de la base commune. La méthode SyMoGIH a permis de stocker avec succès les quelques 16 000 informations concernant plus de 2 800 acteurs de la population du projet ANR Sippaf : les données ainsi structurées peuvent être interrogées et exploitées par les spécialistes ou publiées sous forme de notices biographiques sur le web[18]. Les données d'autres projets collectifs sont également hébergées dans la base commune (par exemple une partie des données de l'ANR Mosare portée par l'UMR 5206 Triangle) ou y seront versées prochainement. Ceci n'est qu'un début : le projet SyMoGIH dispose en effet du potentiel nécessaire pour mettre en place une plateforme généraliste d'hébergement des données à usage à

17 Il existe des logiciels pour les bases de données natives XML, tels eXist ou baseX, qui permettent de croiser, dans une même requête, des connaissances stockées dans une base de données avec celles présentés dans un texte balisé.

18 http://www.patronsdefrance.fr/Database/Acteur_fr.php (consulté 31 janvier 2012).

la fois individuel et collectif. La mutualisation des unités de connaissance ouvre à la recherche historique une nouvelle dimension dans la reconstitution des dynamiques des sociétés du passé en offrant la possibilité d'un travail collectif depuis la modélisation jusqu'au partage et à l'exploitation d'un volume sans cesse croissant de données qu'aucun chercheur, ni même groupe de recherche, ne pourrait, à lui seul, rassembler.