

# Proposition

*Autour du rôle des métadonnées*

T. Libourel

26 novembre 2002

## 1 Préambule

*Document de travail*

*Les réflexions portent sur l'introduction d'un niveau « méta » afin de faciliter les recherches et l'accès ultérieur aux données et traitements nécessaires à la résolution de nouvelles situations.*

## 2 Contexte

Les nouvelles technologies de l'information et de la communication (NTIC) ont augmenté le volume et les flux d'informations mis à disposition de tout public depuis le grand public jusqu'au public de la recherche la plus spécialisée. Les utilisateurs sont en mesure, compte tenu de ces avancées technologiques, d'attendre, voire d'exiger des environnements de travail perfectionnés. Dans les domaines du complexe (biologie, environnement), les projets d'acquisition de données ont généré des masses d'informations considérables qu'il s'agit maintenant d'utiliser, réutiliser à des fins diverses d'analyses et interprétations. La gestion et l'accès à ces diverses informations (expérimentales brutes et interprétées) se révèlent être un des facteurs essentiels pour aller vers une meilleure connaissance. Les systèmes d'information abondent mais constituent une mosaïque hétérogène d'architectures et de propositions.

En effet,

- ▷ les données peuvent être « à plat », semi-structurées, ou bien structurées de diverses manières et suivant différents modèles de données (relationnel, objet, etc.),
- ▷ les programmes relèvent de divers paradigmes (compilateurs, interpréteurs, bibliothèques, composants, etc.),
- ▷ les interfaces peuvent être plus ou moins sophistiquées.

Devant une telle hétérogénéité des systèmes, le problème qui se pose est double :

- ▷ comment gérer la « cohérence » au sein de ces systèmes,
- ▷ comment permettre une gestion et une recherche « intelligente ».

La première difficulté porte sur la cohérence. Bien que cela ne soit pas l'objectif direct de ce premier document, il ne faut pas perdre de vue cette notion primordiale tant au niveau de l'organisation structurelle de l'information qu'au niveau sémantique global du système.

En ce qui concerne la gestion et la recherche d'informations, la contribution de notre travail doit s'inscrire dans la définition d'un cadre méthodologique permettant de doter les concepteurs et utilisateurs d'outils adéquats.

Savoir comment structurer et surtout comment retrouver l'information d'intérêt sont autant de questions que de problèmes posés.

La gestion et la recherche d'informations peuvent se résumer à deux sous-problèmes, la saisie des informations et la restitution de celles-ci à la suite de requêtes ou traitements divers. Tout outil d'aide à la structuration de l'information saisie, devrait entraîner de meilleures performances au niveau des recherches. En effet, on peut admettre que plus la structuration des données est facilitée, plus les manipulations de ces données seront aisées. Et donc, la restitution des informations, lors d'une recherche, augmentera en pertinence.

Les travaux effectués dans le cadre de la modélisation des SI et BD, ont largement privilégié la structuration des données et des traitements. Cependant, les données appartiennent à un domaine « métier » précis qui est rarement décrit dans le système et ce sont les travaux en représentation des connaissances qui, le plus souvent abordent, de manière la plus satisfaisante, la description de ces domaines en termes de concepts et de relations entre ces concepts (niveau ontologique).

Les propositions faites autour des systèmes médiateurs permettent d'une part d'accéder à des environnements multi-bases en proposant leur intégration, et d'autre part d'interroger de manière uniforme l'interface utilisateur en reconstituant l'information recherchée à partir des diverses sources impliquées.

Plusieurs questions sous-jacentes peuvent être formulées :

- ▷ que doit-on intégrer ? de l'information structurée, semi-structurée, du texte, du multimedia, des programmes, etc.
- ▷ que peut-on chercher ? des ressources de base, de l'information construite après sélection parmi des ressources de base, l'information nécessaire à la mise en place de processus de résolution (enchaînant données et traitements).

Nous allons tenter de dégager le fait que les métadonnées peuvent jouer un rôle primordial dans la réalisation de véritables infrastructures de partage de connaissances.

Autrement dit les métadonnées peuvent sûrement aider dans la phase de construction de la médiation.

### 3 Les métadonnées

Le concept de métadonnées tend à être de plus en plus utilisé dans le monde informatique. Donnons une définition de cette notion, tout d'abord à partir de son acception primitive, puis au travers de certains domaines de prédilection dans lesquels les métadonnées sont assignées à des rôles spécifiques que nous détaillerons.

**Définitions et origines** Si on se réfère à la notion de métadonnée dans son acception première la définition est la suivante : les métadonnées sont des données sur les données ou des informations structurées qui décrivent des données.

Pour notre part, nous retenons que l'originalité du terme vient de sa structuration en deux parties : « méta » et « donnée ». La composante « méta » révèle une volonté d'abstraction à un niveau supérieur et introduit aussi la notion de réflexivité. Cela signifie

que les métadonnées doivent compléter l'information relative aux données mais à un niveau d'abstraction supérieur, tout en étant capables de se décrire elles-mêmes.

La deuxième composante « donnée » indique simplement que les métadonnées sont aussi des données, certes d'un niveau différent mais tout de même des données. Et en tant que telles, elles peuvent être manipulées, donc être structurées et interrogées.

Dans le contexte des bases de données, la notion de métadonnées a très rapidement été intégrée au sein même des outils SGBD. Dans de nombreux SGBD, les métadonnées constituent une base au-dessus des bases de données utilisateurs que l'on qualifie de méta-base. La méta-base, à l'instar des métadonnées, permet d'effectuer des descriptions relatives aux bases utilisateurs (considérées comme des données de niveau inférieur). Toute méta-base facilite la vision sur la structuration et la manipulation des données des bases utilisateurs.

L'apparition de la technologie du numérique et plus particulièrement l'explosion des données accessibles sur Internet a provoqué une prise de conscience de l'utilisation potentielle de données de différents niveaux.

**Rôles des métadonnées** Les métadonnées peuvent être utilisées à des fins diverses, depuis leur fonction naturelle d'aide à la structuration et à la recherche d'information, jusqu'aux fonctionnalités plus sophistiquées mises en œuvre dans le cadre d'applications interopérables :

▷ *Documentation.*

La métadonnée est informative, selon le contexte elle apporte une documentation sur les ressources de « base ». Le standard Dublin-Core <sup>1</sup> a bien été conçu dans cet esprit là, permettre la documentation sur des ressources du Web. De manière générale, dans tout domaine expérimental, les métadonnées vont « annoter » l'information en précisant leur(s) auteur(s), leur localisation, leur datation, les protocoles divers, etc.

▷ *Standard d'échange et aide à la recherche.*

Les métadonnées constituent un niveau d'information supérieur et à ce titre peuvent jouer le rôle d'accélérateur d'accès. Les moteurs de recherche du monde Web utilisent déjà les informations « méta » des fichiers *HTML*, l'étape suivante pourrait être la recherche dans des bases de métadonnées structurées (ou semi-structurées). L'information des métadonnées peut aussi servir de base à l'identification du format sous-jacent de l'information concernée et constituer ainsi une aide précieuse lors d'échanges.

▷ *Aide au contrôle et à la protection.*

La métadonnée peut être déconnectée de sa donnée de référence tout en autorisant un contrôle et par suite elle peut constituer un *niveau de sécurité*,

▷ *Aide à la « veille ».*

Les croisements d'information du niveau méta peuvent aider à comparer, à juger de la qualité ou à détecter incohérences, et erreurs.

▷ *Croisement de domaines / sémantiques.*

L'hétérogénéité des bases de données et de connaissances est un fait, elle constitue un écueil bien identifié mais l'uniformisation ou la standardisation universelle ne reste qu'un rêve. Les métadonnées peuvent jouer un rôle important en permettant

---

<sup>1</sup>Dublin Core RFC 2413 <http://purl.org/dc>

« le croisement » de sémantiques liées à différents secteurs (par exemple contexte mécanismes fonctionnels des molécules versus contexte clinique, etc.). Si la sémantique des domaines transparait au sein des métadonnées, il semble possible de guider l'utilisateur à affiner ses recherches en naviguant dans ce niveau « ontologique ».

**Les domaines cibles** Les principaux domaines d'applications des métadonnées, qui semblent apparaître de manière récurrente, sont les suivants :

▷ *Les bases de données*

Les SGBD ont intégré, la notion de méta-niveau, tout d'abord à des fins d'administration des données de toute base. L'administrateur de base de données dispose, grâce à la métabase, d'informations pertinentes et suffisantes pour gérer les bases utilisateur, sans avoir besoin de parcourir les données de ces bases elles-mêmes.

▷ *Les documents électroniques*

Depuis l'arrivée dans le monde professionnel du Web, divers types de documents électroniques sont de plus en plus utilisés. Ces documents intègrent toutes sortes d'informations (textes, sons, images etc.). Les utilisateurs potentiels, sont un petit peu *perdus* devant la quantité d'informations à parcourir, et c'est à ce niveau qu'entre en jeu la métadonnée qui constitue une sorte d'indexation facilitant les recherches. La recherche documentaire multimedia et le commerce électronique font partie des grands courants actuels d'utilisation des métadonnées. Le projet Dublin Core, initié en 1995, pose les jalons pour la mise en place de la gestion (qu'il souhaite optimale) des bibliothèques et archives numériques. La méthodologie objet *OODHM* fait école et s'enrichit de la prise en compte des métadonnées.

▷ *Les systèmes complexes en sciences expérimentales*

Les systèmes relatifs aux sciences expérimentales, notamment les environnements de résolution de problèmes en biologie moléculaire, en médecine, en chimie, en imagerie, cartographie devraient être (sont déjà) de grands utilisateurs de métadonnées. Les expériences pratiquées fournissent des masses d'informations non négligeables. Les données, issues de l'expérimentation, doivent être archivées, interprétées, traitées et interrogées. Les conditions dans lesquelles les expériences se sont déroulées, l'expertise des manipulateurs constituent autant de connaissance complémentaire dont la sauvegarde s'avère précieuse et pertinente, notamment dans le cadre de suivis expérimentaux temporels (épidémiologie, désertification de territoires par exemple).

Les biologistes ont amorcé via l'annotation des séquences nucléiques (c'est-à-dire notamment l'extraction et l'interprétation des régions fonctionnelles de l'ADN) un large pan de recherche dans lequel les informations produites relèvent des métadonnées.

En ce qui concerne l'information géographique, l'usage des métadonnées est quasiment institutionnalisé. Le Federal Geographic Data Committee (FGDC) fut le premier des organismes qui s'est attaché à standardiser les informations relatives aux relevés géographiques, dans le but de faciliter les échanges entre divers utilisateurs (les rôles distincts producteurs /utilisateurs étant bien identifiés). Les résultats des travaux de ce comité ont abouti à la production d'un premier standard pour les métadonnées. Bien d'autres organismes, comme le Comité Technique 287 du Comité Européen de Normalisation (CEN), l'Australia and New Zealand Land Information Council (ANZLIC), le Comité Technique de l'ISO, l'Open-GIS proposent aussi des

normes conçues autour des propositions du FGDC tout en développant des secteurs d'intérêt qui leur sont propres. Le même type de préoccupation existe dans le domaine de la biologie. Des comités essentiellement composés de biologistes et d'informaticiens tentent d'harmoniser la structuration des données issues du monde de la biologie (BIOML BIOpolymer Markup Language, PROML PROtein Markup Language, etc.).

De plus, bien d'autres standards ou formalismes (RDF, XML, MPEG, etc.) sont proposés à l'heure actuelle pour décrire le champ des métadonnées.

Toutes les normes et standards proposés présentent une structuration hiérarchique de l'information, c'est-à-dire une répartition en sections découpées elles-mêmes en sous-sections et ceci sur plusieurs niveaux. L'ensemble des propositions souhaite aussi répondre aux problèmes généraux :

- ▷ la sémantique des métadonnées doit être bien comprise et rapidement assimilée,
- ▷ l'internationalisation des ressources (qui devient de fait quasi une obligation),
- ▷ l'interopérabilité entre diverses collections de ressources.

Certes la variété et la richesse des informations proposées constituent un réservoir de connaissances potentiel, mais en contre-partie elles sont souvent soit mal perçues, soit pire considérées comme un handicap, car, au-delà des institutions, les chercheurs qui souhaitent utiliser les métadonnées hésitent devant l'ampleur du travail de saisie nécessaire.

Les chercheurs des sciences expérimentales sont de plus en plus concernés cependant par le souci de gérer leur patrimoine scientifique et par celui de diffuser leurs travaux (tout en conservant leur « paternité »).

A partir d'expériences diverses, on peut identifier trois grandes catégories de service de métadonnées :

- ▷ les services de type *inventaire de ressources*. En fait il s'agit de gérer au mieux des collections de ressources au sein d'une communauté de recherche,
- ▷ les services de type *observatoire*. Ces services gèrent des ressources, tant données que protocoles, pour fédérer les acquisitions et développements effectués sur des sites de référence dispersés et effectuer des suivis sur du long terme,
- ▷ les services de type *mémoire d'entreprise*. Ici, il s'agit de constituer des dossiers de référence remobilisables ultérieurement.

Dans tous ces services, les métadonnées servent de « socle » à la médiation.

La question immédiate devient : ce socle peut-il servir à la fédération, à l'intégration ?

## Vers une double stratification des métadonnées

Les efforts de standardisation travaillent, tous à définir en quelque sorte un « référentiel » des informations liées aux domaines concernés, mais passent sous silence la variété d'usages et d'utilisateurs.

Plusieurs travaux suggèrent de stratifier les métadonnées en fonction du niveau de spécialisation des utilisateurs. En effet, selon le degré de connaissances de l'utilisateur et selon sa perception du domaine, les recherches effectuées et les réponses obtenues (selon les métadonnées interrogées) risquent d'être plus ou moins pertinentes.

Les trois niveaux suggérés sont établis selon la capacité technologique des usagers.

- ▷ Le *niveau 1* s'adresse aux utilisateurs lambda qui cherchent des informations dans des métadonnées généralistes sans trop connaître, en détail, ni le domaine ni les outils utilisés,
- ▷ Le *niveau 2* est celui des décideurs. Les utilisateurs concernés connaissent bien le domaine et les produits,
- ▷ Le *niveau 3* est celui des experts. N'oublions pas que nous sommes dans les domaines expérimentaux, les experts maîtrisent les spécifications des ressources (données et outils).

Cette stratification (suggérée par toutes les normes) reste un peu réductrice, on peut admettre un filtre différent pour hiérarchiser les métadonnées. Il s'agit en fait de s'appuyer sur le niveau d'abstraction implicite.

- ▷ Le *niveau 1* constitue la connaissance du domaine, *ontologie ? thésaurus ?* décrivant concepts et relations pertinents du domaine,
- ▷ Le *niveau 2* concerne l'information générale relative aux schémas et aux protocoles partagés par des lots de ressources,
- ▷ Le *niveau 3* adresse des informations spécifiques relatives à des données typiques (ou plutôt atypiques car assez singulières pour mériter des annotations particulières).

Si les métadonnées permettent d'introduire un niveau supérieur, capable d'orienter les recherches dans un système médiateur, il semble qu'il faille dégager un modèle assez souple et « générique » pour constituer un niveau de ressources « méta » qui à la manière d'un index rendent l'accès à l'information de base plus rapide et plus pertinent (en prenant en compte la double stratification ?).

De nombreuses questions (en vrac) :

- ▷ on introduit un ensemble de bases supplémentaires (en supposant qu'on gère les métadonnées dans des bases). Faut-il considérer ces bases comme les autres ? ou jouer sur la différence niveau « méta », niveau « de base »,
- ▷ problème des méta-métadonnées ?
- ▷ langage d'interrogation, de navigation ?
- ▷ passerelles et règles d'appariement entre niveaux « ontologiques ».