Proposition de structuration des métadonnées en géosciences : Spécificité de la communauté scientifique

Jean-Christophe Desconnets, Thérèse Libourel Rouge, Pierre Maurel, André Miralles, Michel Passouant

Proposition de structuration des métadonnées en géosciences[1] :

Spécificité de la communauté scientifique

Jean-Christophe Desconnets[*], Thérèse Libourel[**], Pierre Maurel[***], André Miralles[***], Michel Passouant [****]

[*] *IRD[2] – Projet ROSELT Maison de la télédétection, 500, avenue J.F. Breton 34093 Montpellier cedex 05. jc.desconnets@teledetection.fr*

[**] *LIRMM[3],161 rue Ada 34392 Montpellier Cedex 5. libourel@lirmm.fr*

[***] *Cemagref[4] UMR 3S Maison de la télédétection, 500, avenue J.F. Breton 34093 Montpellier cedex 05. pierre.maurel@teledetection.fr, andre.miralles@teledetection.fr*

[****] *CIRAD[5] - TERA, CIRAD, "TA 60/15", 73 avenue Jean François Breton, 34938 Montpellier Cedex 5. michel.passouant@cirad.fr*

RÉSUMÉ :

Cette communication est le résultat d'une réflexion conjointe menée entre chercheurs informaticiens et thématiciens confrontés à la gestion des données et métadonnées dans le cadre de projets propres à leur domaine. Après avoir précisé les spécificités du domaine visé ainsi que la base de travaux déjà existants, nous présentons un modèle général de structuration des métadonnées établi à partir d'une méthodologie objet.

Ce modèle est ensuite validé au travers d'exemples extraits de projets menés par les auteurs.

Finalement, nous suggérons d'intégrer les métadonnées comme une dimension à part entière dans nos systèmes d'information. En perspective, cette méthodologie peut être étendue aux systèmes d'information environnementaux à l'usage des gestionnaires et décideurs.

ABSTRACT:

This communication is the result of the joint thought between computer researchers and scientists dealing with the data and the metadata management within the scope of theirs owns

[1] Géosciences : les auteurs recouvrent par cette terminologie les domaines d'applications rattachés aux sciences de la terre et de l'environnement (agronomie, écologie, agriculture, hydrologie, etc…)

[2] IRD : Institut de Recherche pour le Développement

[3] LIRMM : Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

[4] Cemagref : Institut de recherche pour l'ingénierie de l'agriculture et de l'environnement

[5] CIRAD : Centre de Coopération Internationale en Recherche Agronomique pour le Développement

research projects. After describing the specificities of the concerned field as well as the base of existing works, we present a general model of metadata structuring established from a methodology object.

This model is afterwards validated for concrete examples extracted from projects carried out by the authors. Finally, we suggest to integrate the metadata as full dimension of their information systems. In prospect, this methodology can be extended to the environmental information systems used by the managers and the decision-makers.

MOTS-CLES :

Information Géographique, Métadonnées (structuration et service), Patrimoine scientifique, Méthodologie objet, UML

KEYWORDS:

Geographic information, Metadata (structuring and service), Scientific heritage, Object methodology, UML

---

1. Introduction

Les nouvelles technologies de l'information et de la communication (NTIC) ont augmenté le volume et le flux d'information mis à la disposition des utilisateurs. Cependant l'utilisation des données stockées nécessite une connaissance experte du domaine d'étude. Les utilisateurs potentiels de tels systèmes sont en mesure d'attendre, voire d'exiger des environnements de travail perfectionnés.

Force est de constater que, dans ce contexte, la notion de **métadonnées** bénéficie maintenant d'une large audience. Dans son acception première, le terme signifie " données sur les données ou données qui renseignent sur certaines données et qui permettent ainsi leur utilisation pertinente " (Vocabulaire de la géomatique publié par l'office de la langue française de 1993). Notre approche porte effectivement sur les deux composantes évoquées par le terme métadonnées : **méta,** étymologiquement état d'abstraction à un niveau supérieur et **donnée** suggérant qu'il s'agit effectivement de données de niveau supérieur structurables et manipulables.

Les domaines d'application des métadonnées sont relativement variés. Pour notre part, nous nous intéressons plus particulièrement aux systèmes d'information relatifs aux géosciences dont la particularité est de disposer de masses d'informations hétérogènes et localisées géographiquement. Elles sont, généralement, issues d'expérimentations variées et soumises à des traitements et à des interprétations diverses, mais que la propriété de localisation géographique autorise à utiliser concomitamment. On parle alors d'information géographique ou même d'information spatio-temporelle pour intégrer la dimension liée au temps. Le contenu des métadonnées peut varier d'une description simple de données (type, couverture géographique et fournisseur) jusqu'à une description détaillée (spécifications de la donnée, protocole d'acquisition, de traitements par exemple).

Notre objectif est de proposer, au-delà des usages institutionnalisés des métadonnées dans les catalogues des grands fournisseurs de données géographiques, une structuration générique utilisable et instrumentable par le chercheur dans le cadre de ses activités. De plus, nous pensons que les services construits sur cette structuration devraient contribuer à informer les utilisateurs (chercheurs, décideurs,

administration, grand public, etc.) de l'existence de ce type de données, de leur domaine de validité et d'en faciliter l'accès.

Notre proposition émane d'une réflexion conjointe menée entre chercheurs informaticiens et thématiciens faisant appel à la géomatique dans leurs domaines d'applications relevant des géosciences.

Dans la section 2, nous présentons plus en détail la place des métadonnées pour les géosciences afin de mettre en évidence la particularité de la production/utilisation des données dans les projets scientifiques. Dans ce contexte, un tour d'horizon des principales normes relatives aux métadonnées du domaine est présenté.

La section 3 positionne notre proposition au regard des enjeux et des contraintes d'un service de métadonnées dans la communauté scientifique et en propose une typologie.

La section 4 décrira le modèle de structuration, cœur de notre proposition. Il est articulé autour de **la donnée en tant que référence** (donnée " référence ") et a été réalisé à partir d'une méthodologie objet. Nous le déclinerons dans le formalisme UML. La validation du modèle proposé sera effectuée au travers d'exemples instrumentés.

Nous conclurons l'article par quelques remarques et perspectives sur l'utilisation de la proposition.

2. Métadonnées et géosciences

2.1. *Cycle de vie des informations*

L'information géographique nécessite des phases complexes d'instrumentation pour être disponible. Des données géographiques dites " de référence " sont produites par de grands organismes nationaux (exemple de l'IGN[6] en France pour les cartes et les bases de données topographiques), voire internationaux. Outre, ces grands producteurs de données, un certain nombre d'organismes de recherche ou d'acteurs de la société civile sont amenés, d'une part, à acquérir des données (ils sont alors utilisateurs) et, d'autre part, à produire à leur tour de nouvelles données, fruit d'acquisitions et/ou de traitements spécifiques.

La double spécificité, nature de l'information (thématique et géoréférencement) et cycle production/consommation, doit transparaître lors de la structuration des métadonnées.

2.2. *Rôle des métadonnées*

Il semble judicieux de catégoriser les rôles assignés aux métadonnées depuis leur fonction naturelle d'aide à la structuration et à la recherche d'information, jusqu'aux fonctionnalités plus sophistiquées dans le cadre d'applications interopérables.

Les métadonnées ont clairement un rôle d'identification et de spécification de la donnée " référence ". Elles peuvent ainsi servir de base aux moteurs de recherche. Elles peuvent de plus faciliter ou améliorer la diffusion de données de base en

[6] IGN : Institut Géographique National

établissant une sorte de " standard " d'échange [Coulondre 98]. Elles peuvent enfin assurer une aide au croisement de sémantiques liées à différents domaines d'expertises. Si les métadonnées traduisent une partie de la connaissance " ontologique " du domaine, le rapprochement de métadonnées relatives à plusieurs apports émanant de groupes d'expertise différente sur des données communes doit permettre des correspondances entre concepts.

2.3. *Les normes existantes*

Les métadonnées du domaine des géosciences font actuellement l'objet de recherche dans le cadre de la standardisation et ceci dans une trentaine de pays. Sans prétendre être exhaustifs, nous reprenons les normes qui sont le plus souvent citées dans le domaine de l'information géographique.

La plupart d'entre elles se fondent sur les propositions faites par le **FGDC**[7] dès 1992 [FGDC 98]. Cette norme suggère une nomenclature bâtie sur sept sections principales :

- **l'identification** : comprend le nom et le type de la donnée ;
- **la qualité** : procédé ou processus de création (généalogie), précision ;
- **les caractéristiques spatiales** : mode utilisé (raster ou vecteur), type géométrique ;
- **le système de référence spatiale** : référentiel géographique, projection, système de coordonnées ;
- **les entités et leurs attributs** : description du schéma de la base de données par exemple ;
- **la distribution** : le format, le logiciel, la version logiciel ;
- **la référence** : nom du créateur, date de création ;

et les assujettissant à un certain nombre de formats normalisés pour certains paramètres (latitude, longitude, heure, etc.).

Plusieurs autres organismes ont travaillé et travaillent dans le même sens : le Comité Technique 287 du CEN[8] : **CEN TC 287** [CEN 96], le Comité Technique 211 de l'ISO[9] : **ISO TC 211** [ISO 98], l'Australia and New Zealand Land Information Council : **ANZLIC** [ANZLIC 96], l'OpenGIS Consortium : **OPenGIS** [OpenGIS 98]. Ils proposent des normes conçues autour des propositions du FGDC tout en développant des secteurs d'intérêt qui leur sont propres.

3. Spécificité d'un service de métadonnées dans une communauté scientifique

3.1. *Contexte de la recherche*

Jusqu'à ces dernières années, les missions du chercheur et des organismes de recherche n'intégraient que rarement la gestion du patrimoine scientifique, en tout cas pas son optimisation.

---

[7] FGDC : Federal Geographic Data Commitee

[8] CEN : Comité Européen de Normalisation

[9] ISO : International Organization for Standardization

L'intérêt pour les métadonnées auquel on assiste maintenant peut s'expliquer par :

- le souci de gérer le patrimoine scientifique au niveau d'un projet ou de l'ensemble d'un établissement de recherche ;
- la prise de conscience du rôle de structuration et de représentation des connaissances scientifiques que joue la démarche de construction des métadonnées, qu'elle soit conduite *ex post* ou encore mieux *ex ante*;
- l'introduction progressive d'une **démarche " qualité "** au sein des organismes de recherche ;
- le développement des NTIC, en particulier les solutions internet et intranet, qui facilitent la diffusion des données et nécessitent la mise en place de véritables services de métadonnées.

3.2. *Les enjeux*

Contrairement aux données dites de « référence » qui sont élaborées à partir de protocoles bien établis, les données utilisées dans le cadre de projet de recherche sont le plus souvent le résultat de traitements non prédéfinis et évolutifs en fonction de l'avancement des travaux. De même, une donnée pourra subir des traitements successifs mis en œuvre par différents intervenants.

Au delà de la qualification des données proposée par les normes, un service de métadonnées propre à une communauté scientifique doit pouvoir retracer la généalogie d'une donnée de manière précise et mettre ces informations à disposition des thématiciens.

3.3. *La typologie*

A partir des expériences relatées dans notre communauté scientifique, nous avons identifié trois grandes catégories de services de métadonnées :

- les services de type " **inventaire des ressources en géomatique** " d'un établissement de recherche, avec en particulier les données disponibles, produites en interne ou acquises auprès d'autres organismes. Nous pouvons citer l'exemple démarré au Cemagref pour recenser ses ressources humaines, matérielles et informationnelles de nature géomatique (projets, données, documents) à des fins de mutualisation interne de ces ressources [Carlet 00], [Gabaston 00].
- les services de type " **observatoire** ". Dans ce cas, le service est caractérisé par l'acquisition répétitive de données sur des sites de référence, selon des protocoles précis et en général pour une longue période avec, bien sûr, un accès direct ou filtré aux données (droits d'accès, profils personnalisés,…). C'est le cas par exemple du projet ROSELT[10] [Roselt 95] mené dans la zone Circum saharienne. Il a, notamment, pour objectif la circulation d'informations autour du thème de la lutte contre la désertification dans une communauté d'utilisateurs variés et géographiquement

[10] ROSELT : Réseau d'Observatoires de Surveillance Ecologique à Long terme, sous la maîtrise d'ouvrage de l'OSS (Observatoire du Sahara et du Sahel) et la maîtrise d'œuvre de l'IRD (Institut de Recherche pour le Développement)

dispersés (décideurs, scientifiques et grand public du Nord et du Sud). Cela pourrait aussi être le cas de bassins versants de référence suivis en France par le Cemagref.

- les services de type " **mémoire d'expertise** ". Ici, il s'agit de constituer l'inventaire des données brutes ou élaborées selon des méthodologies précises, produites à l'occasion d'un projet scientifique particulier. L'objectif principal est de pouvoir remobiliser ultérieurement les données, y compris par des équipes de recherche autres que celle d'origine. Pour cela il faut mettre à disposition tout le **corpus de connaissances** qui a été **utilisé pour constituer le lot de données**. Tel a été le cas de deux applications menées par le CIRAD [Meftah 99] [Faye 00].

3.4. *Les contraintes*

Les caractéristiques du service de métadonnées mis en place (architecture informatique, nature des métadonnées) doivent ensuite être modulées en fonction de plusieurs paramètres :

- le couplage " temporel " entre le projet de recherche et la saisie des métadonnées. Il sera en effet toujours plus difficile de saisir a posteriori les métadonnées (oubli progressif du projet, faible motivation pour faire des recherches dans les archives du projet, départ du personnel temporaire, structuration initiale des données sûrement moins bonne à cause justement de l'absence d'un système de caractérisation des données et des traitements au début du projet) ;
- le niveau de précision des métadonnées, qui dépend à la fois des usages visés et des ressources humaines disponibles pour saisir les métadonnées ;
- le degré d'ouverture du service de métadonnées, qui pourra concerner, du plus restreint au plus élargi, le chercheur isolé, l'équipe ou les équipes d'un même projet de recherche, le laboratoire gérant plusieurs projets, l'ensemble d'un établissement scientifique, la communauté scientifique au sens large.

4. Proposition d'un service de métadonnées

4.1. *Proposition d'un modèle générique*

Le modèle proposé structure les métadonnées autour de la donnée " référence " selon le point de vue du chercheur, c'est à dire en décrivant sous forme de concepts, les informations relatives :

- aux protocoles d'acquisition et/ou de traitement associés à leurs opérateurs,
- au type de donnée (type simple s'il s'agit d'une donnée élémentaire ou d'un type complexe dans le cas par exemple d'une base de données ou d'une collection),
- à la localisation qui est essentielle car elle situe la donnée dans son référentiel,
- au support, informations utiles à la diffusion de la donnée (format, volume de stockage de la donnée, etc.),
- au domaine d'étude, qui permet de rattacher la donnée à sa perception en termes de connaissance par les acteurs du domaine.

Le modèle présenté avec le formalisme UML, volontairement simplifié (figure 1), ne présente que des classes réifiant les concepts précités, le détail du contenu de chaque classe étant omis. Nous distinguons la classe apportant des

informations de référence sur la donnée (fond blanc) de celles apportant une information sur la métadonnée (fond grisé). De plus certaines classes ou associations sont annotées d'un icône T, C ou O à des fins de lisibilité. Les icônes symbolisent respectivement les concepts d'entités temporelles (figure 2), d'agrégation (figure 3) [Gamma 94] et de terminologie afférente à un domaine qui pourrait s'assimiler aux travaux relatifs aux ontologies (figure 4). Les figures respectives précisent les notions mises en jeu, leurs propriétés et la structure de ces notions.

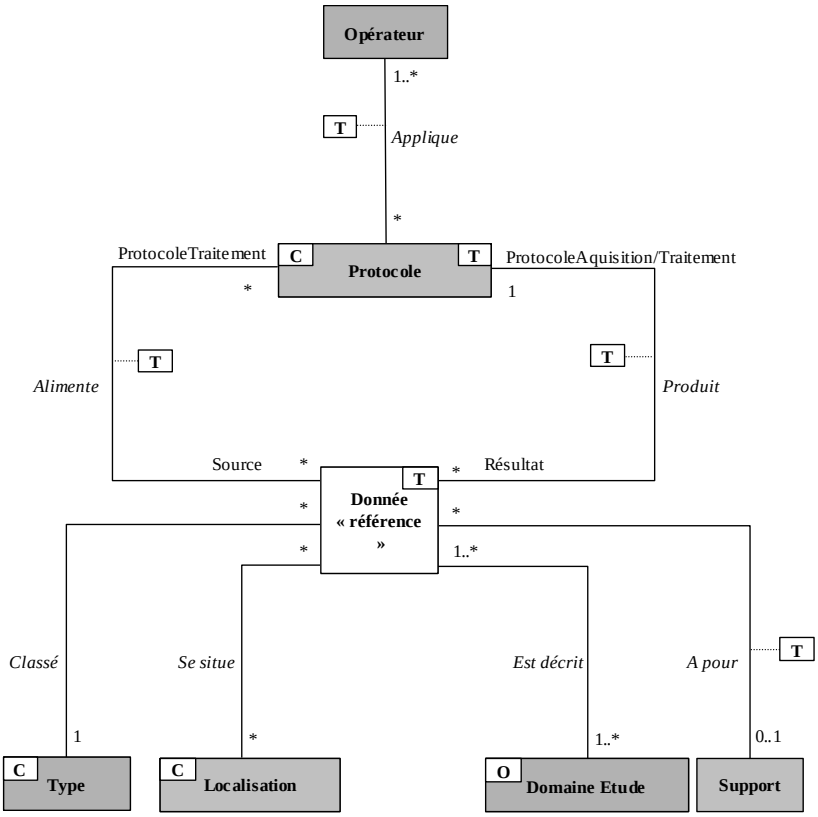

Figure 1. Modèle de structuration des métadonnées

Ce modèle est le résultat d'une réflexion qui est partie des expériences déjà menées et des modèles existants que nous avions établi à ces occasions. Elle utilise les principes de l'approche objet : abstraire pour dégager les concepts essentiels et identifier des motifs structurels réutilisables, cas des motifs symbolisés par les icônes T, C et O.

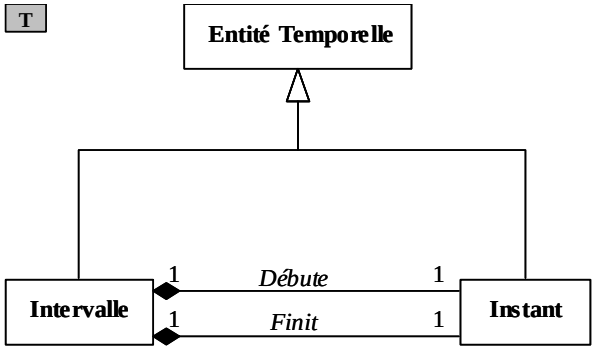

Figure 2. “ Motif ” de la notion de temporalité (icône T)

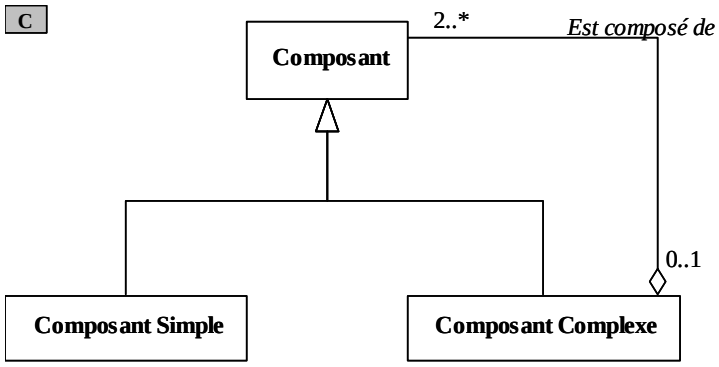

Figure 3. “ Motif ” de la notion d’agrégation (icône C)

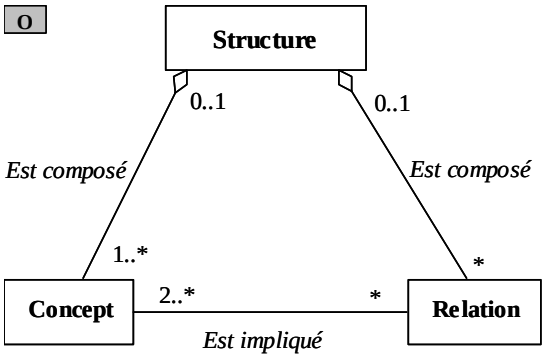

Figure 4. “ Motif ” de la notion de connaissances de type ontologique (icône O)

Ce modèle (Figure 1) est l'ossature à partir de laquelle pourra être développé le modèle de métadonnées spécifique à chaque application. Le concepteur a la liberté d'affiner les concepts primitifs du modèle initial selon le degré de détail auquel il veut aboutir. Nous donnons quelques exemples d'affinage de concepts pour la catégorie « observatoire » :

- le type de données (Figure 5)
- le type localisation (Figure 6)

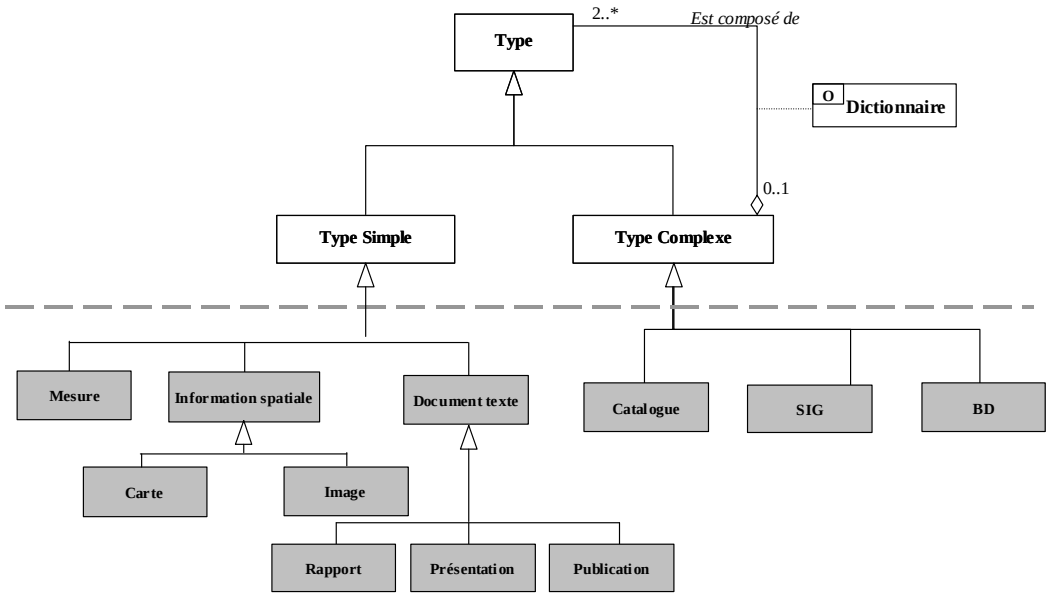

Figure 5. Spécialisation relative au type de la donnée dans le cas " observatoire ".

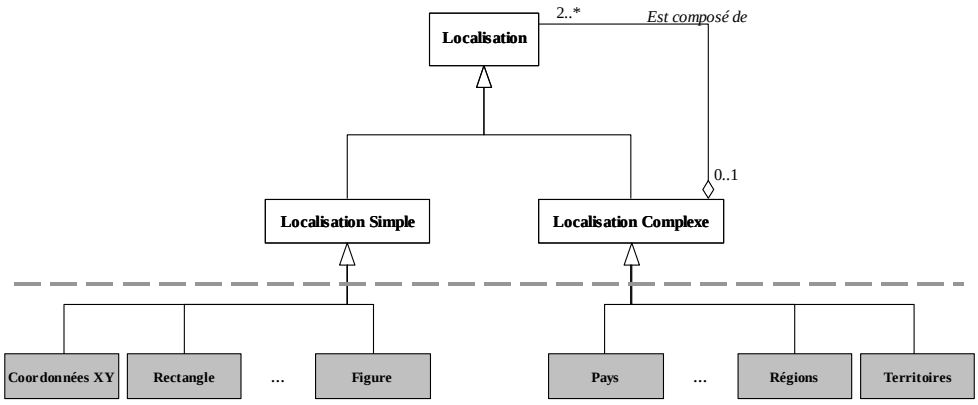

Figure 6. Spécialisation relative à la localisation de la donnée dans le cas " observatoire "

Notre proposition répond de manière pragmatique aux deux points de vue, producteur et utilisateur de données tout en s'inscrivant dans une démarche qualité. Une base de métadonnées construite sur ce modèle donnera :

- au producteur un moyen de gérer, de décrire et de promouvoir les jeux de données produits,
- à l'utilisateur les moyens d'évaluer la pertinence, le sens et la qualité des jeux de données accédées.

L'appariement avec les sections principales des nomenclatures normatives (section 3) reste compatible. Seul le thème qualité n'est pas identifié en tant que tel dans le modèle mais il est présent dans la description des autres classes (protocole, type, support, localisation). Notamment la notion de généalogie (c'est à dire le suivi des données dans leur cycle de vie) [Spery 99] est entièrement couverte au travers du concept protocole et de ses associations avec la donnée " référence ".

4.2. *Validation du modèle générique sur des exemples*

La proposition du modèle est née de l'intégration des modèles de métadonnées pré-existants spécifiques aux applications mentionnées en section 4.1. Elle nous a servi à construire le modèle par abstractions successives. Ceci a constitué la première forme de validation. De nouvelles applications sont en cours. Elles ont été conçues et implémentées sur la base de notre proposition et constituent la deuxième forme de validation.

A titre d'exemple, le modèle général a pu être directement utilisé dans le cadre du projet ROSELT [Roselt 01]. Un extrait du script d'implémentation de la base est donné en annexe.

Il semble se dégager un niveau élémentaire correspondant aux **métadonnées principales** qui permet à l'utilisateur de connaître l'existence de données produites en interrogeant la métabase par des requêtes du type : existe-t-il des données localisées sur tel territoire et concernant telle thématique ? Un deuxième niveau de **métadonnées élargies** peut être exploré. Il apporte une information utile à la diffusion de la donnée et peut détailler la capacité des données à satisfaire les besoins d'une application déterminée, à obtenir les renseignements nécessaires à une utilisation adéquate de la donnée.

4.3. *Exemple d'une architecture d'un service de métadonnées*

L'architecture illustrée en figure 7 préfigure le service de métadonnées adapté à une application de ce type, et plus largement aux différents types définis en section 3. Il s'agira de mettre à disposition, dans un environnement Internet ou Intranet, un service d'interrogation multicritères de la base de métadonnées s'appuyant sur les informations relatives au domaine d'étude et à la localisation de la donnée par exemple.

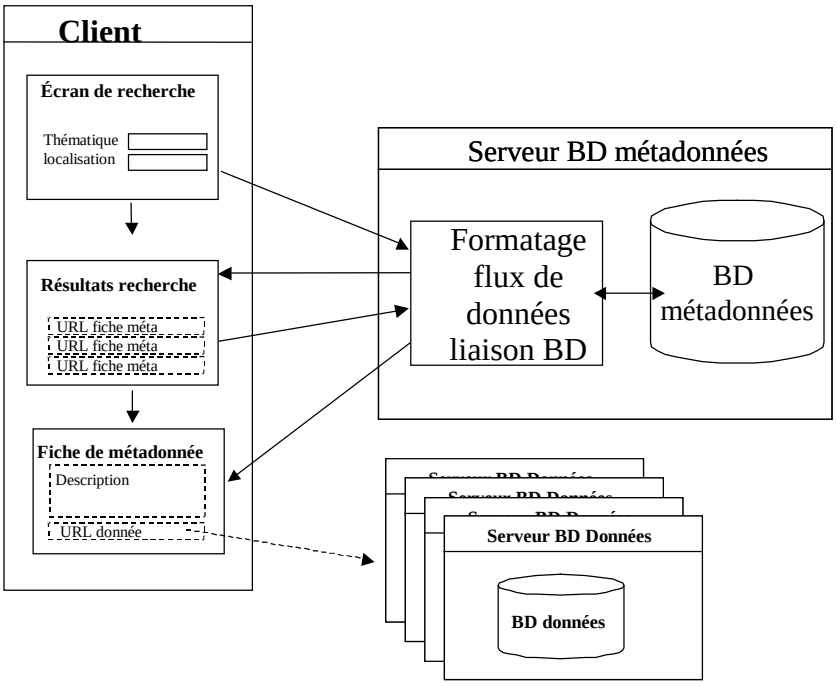

Figure 7. Préfiguration du service de métadonnées de ROSELT

Le résultat des requêtes pourra être, dans un premier temps, une liste de liens donnant accès, dans un deuxième temps, à une fiche descriptive de la donnée, fiche formatable selon le profil de l'utilisateur. La dernière étape doit pouvoir, au regard de la pertinence de la donnée proposée, permettre l'accès à la donnée à proprement parler à travers un protocole de connexion vers la base de donnée distante.

5. Perspectives

Nous venons de présenter un modèle générique de structuration des métadonnées adapté à la communauté scientifique des géosciences. Ce modèle est suffisamment ouvert pour permettre des implémentations propres aux différents types d'applications scientifiques. Il permet également de réutiliser les qualifications des données et des métadonnées que l'on retrouve dans les différentes normes. Enfin, il tient compte des spécificités du cycle de vie d'une donnée dans le cadre de projets scientifiques.

L'objectif est maintenant d'implémenter et de tester ce modèle en construisant des services de métadonnées pour différentes applications opérationnelles s'appuyant sur des bases de données environnementales hétérogènes et réparties. Ces services devraient pouvoir offrir plusieurs fonctionnalités comme : 1) l'automatisation de l'inventaire et de la saisie des métadonnées (en local ou en déporté), 2) la définition de profils utilisateurs pour les recherches personnalisées, 3)

la définition des zones d'intérêt complexes (pour les recherches et les traitements), 4) l'accès direct ou filtré aux données.

Ces développements devraient contribuer de manière significative à une meilleure gestion du patrimoine de données dans le contexte plus global d'une démarche « qualité » au sein d'établissements comme les centres de recherche, les services de l'état, les organisations internationales, etc.

Bibliographie

[Anzlic 96] ANZLIC ; 1996, ANZLIC Guidelines : Core metadata elements Version 1, 147 p

[Carlet 00] Carlet Y., Foucher S., Mas S., Reynaud C., 2000. Conception et modélisation d'une base de données recensant les données, le matériel et les compétences des agents du Cemagref dans le domaine de la géomatique. Stage d'analyse du DESS IAO, Université de Montpellier II – CNAM, mai 2000, 37 p.

[CEN 96] CEN / TC 287. Geographic Information – Data Description - References Model – Metadata.

[CNIG 99] CNIG, 1999. Le catalogage et les métadonnées. Fiche technique du CNIG n°23. 4p.

[Coulondre 98] Coulondre S., Libourel T., Spéry L., 1998. Metadata and GIS : a classification of Metadata for GIS. GIS Planet'98, International Conférence and Exhibition on Geographic Information. Lisbonne, Portugal.

[Gabaston 00] Gabaston S., 2000. Développement d'une base de données couplée à un intranet, pour l'inventaire des ressources en géomatique au Cemagref. DESS IAO, Université de Montpellier II – CNAM, 18 septembre 2000, 163 p.

[FGDC 98] FGDC Content standard for digital geospatial metadata. FGDC-STD-001-998

[Gamma 94] Gamma E., Helm R., Jonhson R., Vlissides J., 1994. Design Patterns Addison Wesley.

[Meftah 99] Meftah, N. (1999). Capitalisation des données d'une action de recherche. Réalisation d'une base de données avec intégration des méta-données, Conservatoire National des Arts et Métiers, Université Montpellier II: 127 p.

[Scholl 96] Scholl M., Voisard A., Peloux J-P., Raynal L., Rigaux P., 1996. SGBD Géographiques, Spécificités. International Thomson Publishing.

[ISO 98] ISO /TC 211. Geographic Information – Part 15 : Metadata. Numero 15046-15.

[OpenGIS 98] OpenGIS Consortium, 1998. The OpenGIS specification Model – Topic 11 : Metadata. Document Number : 98 – 111r2.

[Roselt 95] Roselt, 1995. Conception, organisation et mise en œuvre de Roselt. Document de référence du réseau, août 1995, IARE, Montpellier. 68 pp + annexes.

[Roselt 01] Roselt, 2001. Rapport d'analyse de la base de métadonnées Roselt : Propositions pour le développement d'une base de référence autour des données ROSELT sur Internet. Montpellier, IRD, Février 2001.

[Rouet 93] Rouet P., 1993. Les Données dans les systèmes d'information géographique. Paris, Hermès.

[Spery 99] Spery L., Claramunt C., Libourel T., 1999. A lineage metadata model for the temporal management of a cadastre application. Proceedings of the International Worshop on Spatio-temporal Models and Language, SDTDML'99, IEEE.