# GRASS/ OSGeo-News

**Open Source GIS and Remote Sensing informations** Volume 1, June 2004


# Approche générique pour la gestion de standards de métadonnées


*par Barde Julien et Egdington Duane et Desconnets Jean-Christophe*


## Inroduction

La gestion des *Ressources Informationnelles*[1] (RI) est une composante essentielle de la gestion des ressources environnementales. En effet, l'amélioration du *traitement des données* et des *prises de décisions* est fortement dépendante de la capacité à *localiser les RI pertinentes*.

Néanmoins, la localisation exhaustive des RI pertinentes est un challenge pour les utilisateurs qui doivent, pour cela, faire face aux contraintes suivantes:

- les RI sont *hétérogènes* (*langage*, *sémantique*, *syntaxe/formats*, *métadonnées*, *contraintes d'accès* dues à leur *rareté* et à leur *coût*...),
- les RI sont *distribuées* dans des *Systèmes d'Information (SI) hétérogènes* dont l'interopérabilité nécessite la résolution de *correspondances sémantiques et syntaxiques*.

L'objectif prioritaire pour améliorer la localisation des RI pertinentes consiste à gérer efficacement les couples *{élément de métadonnée, valeur}* qui composent les fiches de métadonnées. En effet, une fois au courant des SI à interroger, la priorité pour localiser les RI pertinentes est la gestion des correspondances (syntaxiques) entre les éléments de métadonnée hétérogènes et de leurs valeurs hétérogènes (sémantiques). L'échelle globale du domaine environnemental ainsi que les contextes habituellement pluridisciplinaires des études associées augmentent fortement ces contraintes et le besoin de référentiels

[1]une *ressource informationnelle* (RI) recouvre l'ensemble des données, informations, connaissances que les acteurs produisent, recherchent et traitent. Indépendamment de leur format (numérique, papier...), selon les acteurs, ce terme peut par exemple représenter un rapport, une carte, une image, une photo, une vidéo, un jeu ou une série de données, une base de données, un modèle....

**Contents of this volume:**

sémantiques et spatiaux.

## Gestion des métadonnées

L'amélioration de la *gestion actuelle des métadonnées* est donc une priorité. D'une manière générale, on constate une utilisation croissante (bien qu'encore trop faible) de référentiels pour homogénéiser la terminologie des *éléments de métadonnées*, les *relations* qu'ils entretiennent et les *valeurs* qui leur sont associées. En conséquence, la qualité des outils de gestion de métadonnées doit être évaluée autant sur le respect d'une implémentation conforme aux spécifications des standards de référence (syntaxique: les éléments de métadonnées et leur structure) que sur la capacité à leur associer des valeurs issues de référentiels sémantiques pour éditer des instances.

L'*hétérogénéité des jeux d'éléments de métadonnées* perturbe l'*interopérabilité syntaxique des SI*:

- les standards utilisent souvent un noyau d'*éléments de métadonnées principaux* identique où les éléments font référence aux mêmes *concepts* en utilisant cependant des *termes* différents (ces *éléments principaux* répondent aux questions suivantes: Où? Quoi? Quand? Qui?...),
- l'utilisation de standards dont les *portées sont redondantes* génère des problèmes de mises en *correspondances* (étant donné qu'une même sorte de RI peut être potentiellement décrite avec différents standards),
- la mise en place récente de nouveaux *standards de métadonnées internationaux* s'appuyant sur les standards nationaux/locaux préexistants les rends obsolètes et génère des problèmes d'*archivages* (stockage des métadonnées en l'état ou mises en *correspondances* vers les standards de référence),
- l'usage systématique récent de *XML Schemas* pour standardiser l'implémentation des standards diminue la redondance de leurs portées et prévient des mauvaises interprétations de leur implémentation. En effet, les standards peuvent être utilisés à présent comme des bibliothèques de types W3C (2004) (c'est la cas, par exemple, des standards de l'*OGC* comme la récente mise à jour de l'*ISO 19115* par l'*ISO 19139*, *SensorML*...).

L'*hétérogénéité potentielle des valeurs* associées aux éléments de métadonnées interfère avec l'*interopérabilité sémantique des SI* si on ne peut pas les contrôler (quel que soit le standard de métadonnées choisi): leur contrôle par l'utilisation complémentaire de référentiels est un problème central pour améliorer la description des RI et donc leur localisation. En particulier, une gestion fine des valeurs associées aux *éléments de métadonnées principaux* est impérative puisque ces derniers sont utilisés prioritairement pour répondre aux requêtes les plus communes des utilisateurs. La gestion des descriptions suivantes est particulièrement cruciale tant elles sont complexes et ambigües:

- *description terminologique* avec des vocabulaires contrôlés/référentiels sémantiques (multilingues) pour valuer les *éléments de métadonnées de type "keyword"*. De plus, l'utilisation de tels référentiels aide à la mise en place de vocabulaires communs dans des contextes pluridisciplinaires,
- *description spatiale* avec des référentiels spatiaux/géographiques présentés avec des GUIs adaptés pour améliorer l'utilisation complexe de l'information géographique (*IG*) (en particulier lorsqu'il s'agit d'utiliser des formats comme *GML*, *WKT*...).

Le besoin d'outils de gestion de métadonnées multi-standards intégrant le contrôle de ces valeurs clés va croissant. Les demandes intra- ou inter-instituts/projets qui gèrent souvent plus d'une sorte de RI et ne peuvent se contenter d'un unique standard vont dans ce sens.

De plus, si on prend en compte les *tâches et besoins* respectifs des *utilisateurs* et des *développeurs* pour la gestion des métadonnées, il apparait que la plupart d'entre eux sont similaires et indépendants des standards considérés.

### Tâches et besoins des utilisateurs

Selon leurs rôles (administrateur...), quel que soit le standard, les *tâches des utilisateurs* associées à la gestion d'un standard de métadonnées consiste habituellement en :

- la mise en place de *profils* des standards de métadonnées pour couvrir les besoins de leurs utilisations specifiques,
- l'*édition* d'instance de ces standards pour décrire leur RI (en associant des *valeurs* aux *éléments de métadonnées* retenus dans les *profils*),
- la *localisation* (et l'éventuelle *acquisition* des RI selon les contraintes de droits d'accès associés) des RI pertinentes pour un travail donné en utilisant un *moteur de recherche multi-critères unique* qui autorise des *requêtes spatiales* évoluées,
- l'*import/export* d'instances de standards de métadonnées (habituellement au format *XML*),

*Les besoins des utilisateurs* se concentrent sur la convivialité (actuellement perfectible) des outils proposés pour exécuter ces tâches. La plupart du temps, les utilisateurs exprime le besoin d'un outil centralisé avec des GUIs dont l'ergonomie conviviale sera conservée d'un standard à l'autre puisqu'elle a pour objectif d'aider à réaliser les mêmes tâches. En effet, l'hétérogénéité actuelle des logiciels ou SI, qui implémentent pourtant les mêmes standards, augmente les efforts d'adaptation des utilisateurs alors que les objectifs visés sont paradoxalement les mêmes (étant entendu que l'utilisateur doit passer -voire perdre- du temps pour se familiariser avec l'ergonomie de chaque outil). Enfin, un maximum d'aide est attendu (*interface cartographique en ligne*, *vocabulaires contrôlés*, *calendrier*...) lorsque les valeurs à associer aux éléments de métadonnées sont trop complexes à fournir sous la forme demandée par les standards (*GML*...). Ces cas d'utilisation sont illustrés par un diagramme UML dans la présentation de la conférence (ici).

## Tâches et besoins des développeurs

De la même manière, les tâches et besoins des développeurs s'avèrent identiques d'un standard à l'autre. *Les tâches des développeurs* consistent à:

- conjuguer *la satisfaction des besoins des utilisateurs* tout en *se conformant aux spécifications des standards*,
- la gestion de *correspondances entre éléments de métadonnées principaux* de différents standards pour répondre aux requêtes de base aussi efficacement que possible,
- intégrer et gérer des *référentiels sémantiques et spatiaux* existants (comme AGROVOC, GEMET, ...) pour garantir la qualité des descriptions des RI et pouvoir gérer ultérieurement les *procédures d'expansions de requêtes*. En effet, la localisation efficace des RI implique la gestion d'autant de requêtes différentes que de sortes de SI à interroger. Les réponses à ces requêtes sont d'autant plus difficiles à obtenir que les termes et les objets géographiques utilisés dans les descriptions n'ont pas été contrôlés en amont,
- la mise en place parallèle d'une *infrastructure de données spatiales* riche permettant la gestion simultanée des RI décrites par les métadonnées et leurs éventuels traitements ultérieurs.

*Les besoins des développeurs* concernent la *diminution des efforts de développements* pour implémenter (automatiquement) un nouveau standard Bernstein et al. (2000) en répondant aux besoins identiques des utilisateurs de la même manière: en réutilisant un jeu de scripts unique et les mêmes composants pour les assister dans la réalisation de leurs tâches (WMS...). Ceci correspond finalement au besoin d'une *approche dite générique* car indépendante du contenu propre aux standards implémentés Barde (2005).

## Approche générique vs approche spécifique

Les implémentations spécifiques traditionnelles sont obtenues par une traduction directe des spécifications des standards en des *modèles physiques de données* qui constituent des systèmes de gestion des données hétérogènes et nécessitent donc des jeux de scripts hétérogènes pour les exploiter (voir figure 1).

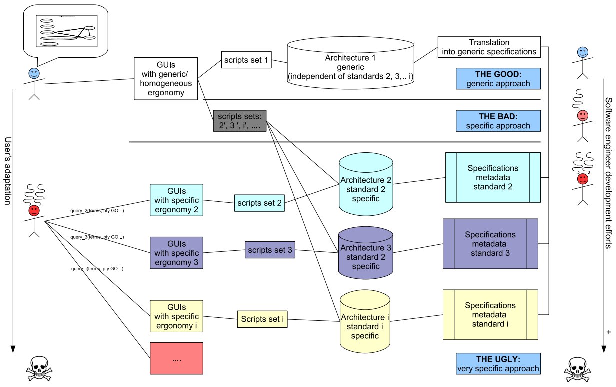


Figure 1: Bénéfice d'une approche générique pour la gestion multi-standards

Par exemple en utilisant une génération de code (automatique ou non) SQL basée sur les spécifications des standards en UML ou XML-S/XSD, les modèles physiques de données qui en résultent vont être fortement hétérogènes (avec des noms de tables et de colonnes qui correspondent aux termes employés pour désigner le nom des éléments de métadonnées dans chaque standard spécifique). Les jeux de scripts en charge de traiter le contenu de ces tables doivent alors être adaptés à ces jeux de termes spécifiques (alors qu'il s'agit de réaliser les mêmes tâches, et sont donc hétérogènes d'un standard à l'autre).

En accord avec les listes de tâches et de besoins dressées auparavant, la figure 1 illustre le bénéfice d'une *approche générique* pour les utilisateurs comme pour les développeurs.

Jusqu'ici, les outils existants ne couvrent pas l'ensemble de ces besoins d'une manière satisfaisante, principalement parce qu'ils sont basés sur des approches spécifiques.

# Des modèles génériques pour une gestion efficace des couples *{élément de métadonnée, valeur}*

Nous présentons dans cette section les modèles génériques actuellement préconisés et implémentés pour mettre en place un outil de gestion de métadonnées multi-standards doté de composants additionnels visant à assister les utilisateurs pour contrôler les valeurs des éléments de métadonnées principaux. En particulier, les modèles présentés seront focalisés sur les valeurs des éléments de métadonnées les plus cruciaux qui concernent les *descriptions thématiques et spatiales*.

## Un modèle générique pour gérer des standards de métadonnées hétérogènes

L'objectif est d'ajouter une étape intermédiaire en proposant un patron générique (modèle conceptuel présenté dans le diagramme de classes UML de la figure 2) qui fournit une couche d'abstraction permettant de décrire n'importe quel standard de métadonnées et donc de mettre en place un système de gestion de métadonnées générique. Ce modèle intègre également le nécessaire contrôle des valeurs, en particulier celles associées aux éléments de métadonnées principaux. Pour cela, le modèle proposé en figure 2 considère qu'un standard de métadonnées est un *inventaire d'éléments de métadonnées structurés* dont le *contenu* est rempli par des *valeurs* potentiellement contrôlées (selon des règles conformes aux spécifications des standards ou complétées par volonté des développeurs selon le contexte d'implémentation).

Cette approche peut finalement être comparée à l'objectif du DOM[2] qui manipule les mêmes concepts pour gérer la *structure* des *noeuds* et des *valeurs* qui leur sont associées dans tout type de document. L'objectif étant dans ce cas restreint au cas spécifique des standards de métadonnées.

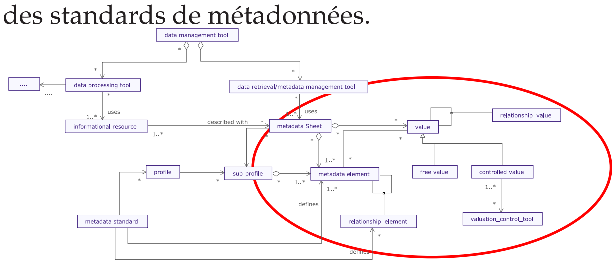


Figure 2: Expression générique des spécifications d'un standard de métadonnées

Toutefois, un standard n'a pas pour objectif de définir ni de contrôler toutes les valeurs potentielles associées aux éléments de métadonnées (principaux en particulier), souvent un standard se contente de suggérer leur contrôle. Le contrôle des valeurs (termes, objets géographiques. . . ) et de leurs relations fait l'objet d'autres standards spécifiques. L'intégration de ces outils complémentaires pour gérer efficacement des valeurs spécifiques est donc bien une tâche du développeur.

Nous nous focaliserons par la suite sur le cas spécifique de la gestion des valeurs associées aux descriptions spatiales et thématiques. Nous spécifions donc le modèle précédent pour permettre la gestion complémentaire de ces valeurs particulières et de leurs relations.

## Un modèle générique pour la gestion des valeurs (spatiales et thématiques) hétérogènes

Les différentes sortes de couples *{élément de métadonnée, valeur}* sont plus ou moins cruciales pour la localisation de données. En particulier, certaines *valeurs* sont spécialement complexes à contrôler. Parmi elles, les *descriptions thématiques et spatiales* sont toutes les deux et cruciales et complexes puisqu'elles correspondent aux éléments de métadonnées principaux impliqués dans l'essentiel des requêtes des utilisateurs (elles répondent aux questions *Où*? et *Quoi*?).

Nous allons donc nous concentrer sur ce cas particulier en proposant une formalisation des propriétés et des relations des concepts thématiques et spatiaux (illustrée en figure 3).

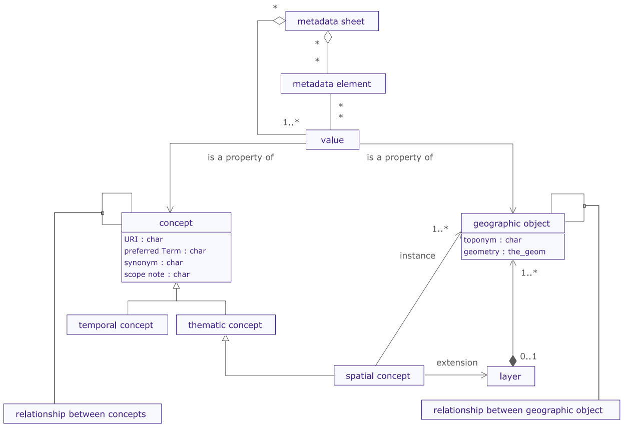


Figure 3: Propriétés et relations entre concepts thématiques et spatiaux

L'utilisation de termes pour valuer des éléments de métadonnées principaux peut prêter à confusion:

- les utilisateurs formalisent souvent leurs descriptions ou recherche de RI en utilisant des termes similaires à ceux de l'exemple suivant: "*espadon, température de surface, Madagascar,*

[2]Document Object Model

*hiver...*",

- la gestion des *relations sémantiques* est essentielle puisqu'elle va par exemple permettre d'explorer les relations sémantiques de ces termes pour étendre ce type de requêtes et améliorer la localisation exhaustive des RI pertinentes. Par exemple, en étendant la requête à la recherche de RI décrites avec le terme ("*Xiphias gladius...*") qui est un synonyme d'"*espadon...*"), étant entendu que ces deux *termes* désignent le même *concept* (qui pourrait tout aussi bien être désigné par une représentation graphique comme une image, une photo...),
- le cas d'un *toponyme* doit être considéré à part puisque plus complexe. En effet, un *toponyme* peut être considéré à la fois comme un terme et comme une description (géo)graphique pointant un *objet géographique*. Par exemple, l'objet géographique associé au terme/-toponyme "*Madagascar*" pourrait tout autant avoir été désigné dans une interface cartographique sans rien changer à l'intention de l'utilisateur.

Pour limiter ces confusions, nous suggérons une gestion des relations sémantiques et spatiales entre les concepts thématiques et spatiaux ainsi que les objets géographiques qui intègrent la complexité des exemples précédents en s'appuyant sur la formalisation suivante: "*un concept spatial est une sorte de concept thématique dont les instances sont des objets géographiques*" Barde et al. (2006), voir figure 3. Toutefois, il convient de préciser qu'un objet géographique n'est pas forcément désigné par un terme ou un toponyme du langage courant.

La figure 4 synthétise et précise le contenu des modèles précédents en un modèle unique permettant d'améliorer l'implémentation générique des standards de métadonnées en intégrant par défaut l'emploi de référentiels qui permettront de contrôler ces valeurs essentielles.

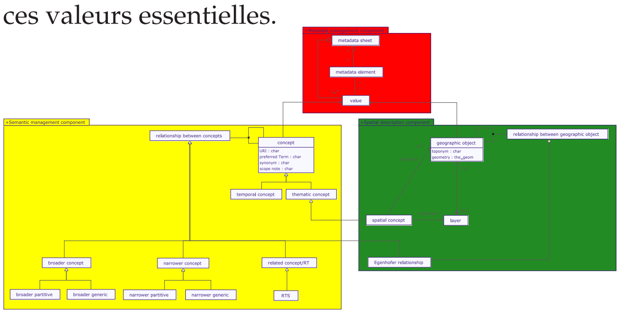


Figure 4: Synthèse de l'approche générique proposée

Ce modèle a été conçu pour pouvoir s'appuyer sur les initiatives de référence actuellement, aussi bien pour les *standards de métadonnées* (*XML Schemas*, *DTD...*), que pour la *sémantique* (travaux du W3C sur le (Web) Semantic avec les standards : *SKOS*[3]/*RDF*/*OWL*) ou l'*information spatiale* (prise en charge des principaux formats pour l'information géographique). En particulier, ce modèle permet, au sein d'une même architecture, de *lier physiquement les éléments de métadonnées avec des ontologies* qui permettent de contrôler les valeurs associées (incluant les descriptions spatiales) et d'étendre les requêtes efficacement.

De la même manière, il est possible d'ajouter des contrôles pour d'autres valeurs cruciales: en particulier les *descriptions temporelles et de contacts* associées aux RI. Ces descriptions sont aussi importantes (puisqu'elles permettent de répondre aux questions *Qui?* et *Quand?* mais moins complexes à contrôler. Les outils associés au contrôle de ces valeurs sont généralement des composants du type *calendrier* ou *annuaire* (en charge d'assister la désignation d'une date/période ou d'acteurs associés à la RI décrite).

La gestion de ces référentiels supplémentaires peut être faite indépendamment des standards de métadonnées implémentés. Cependant, les différents termes qui désignent les mêmes éléments de métadonnées principaux dans différents standards sont finalement valués en utilisant les différents objets gérés dans ces référentiels (traités pour calculer les valeurs sous la forme requise par les spécifications). Il est donc souhaitable de les gérer ensemble. En parallèle, en mémorisant dans une *table d'indexation générique* dédiée, quels objets (concepts, objets géographiques, dates/périodes, acteurs...) ont été associés aux RI pour les descriptions thématique, spatio-temporelle, humaine, il devient possible de répondre à l'essentiel des requêtes des utilisateurs indépendamment des standards qui ont été utilisés pour les décrire, en se basant sur des descriptions aussi riches que possibles (géometries 2D/3D, concept au lieu des termes...) qui ouvrent de nouvelles perspectives pour résoudre et étendre des requêtes complexes.

## Implémentation des modèles avec des logiciels open source

Nous présentons dans cette dernière section l'intérêt d'une implémentation basée sur des applications *open source*.

[3]sorte de RDF lié aux normes *ISO 2788* et *5964*

## Choix techniques sous-jacents

Les implémentations actuelles de l'architecture préconisée sont basées sur le logiciel open source *MDWeb* qui utilise lui-même d'autres applications *open source*. *MDWeb*:

- est un outil de catalogage *multi-standard* et *multi-lingue* qui implémente une approche générique (comme *M3Cat*, *MetaCat*...),
- utilise une architecture client-server *trois-tiers* avec:
  1. des *GUIs* (dans des navigateurs Web) et des composants additionnels (pop-ups) pour assister convivialement l'édition d'instances de standards ou la recherche de métadonnées, notamment:
     - *la description spatiale* avec des interfaces cartographiques en guise d'interfaces/GUIs conviviaux permettant en prime de visualiser les RI spatiales décrites: *Mapserver / Mapbuilder*,
     - *la description thématique* avec des GUIs de gestion de vocabulaires contrôlés pour mettre en place et parcourir des thésaurus / ontologies: "composant maison".
  2. *scripts de l'application* (PHP/Javascript/XML avec Apache comme serveur Http),
  3. *stockage de données*: un *SGBDR* pour la gestion des standards de métadonnées & IG & instances de métadonnées associées & vocabulaires contrôlés: *Postgres* avec *Postgis* pour l'information spatiale gérée localement (WMS pour l'IG distante...). Import de fichiers *SKOS* dans Postgres en utilisant l'API java *JENA*. Répertoires XML.

La présentation de la conférence illustre les principales caractéristiques de ce type d'architecture trois-tiers pour un outil de gestion de métadonnées multi-standards et détaille l'infrastructure physique de données.

### Exemples de jeu générique possible pour les GUIS

En utilisant *MDWeb* comme base pour implémenter cette approche, il est donc possible de couvrir à la fois les besoins des utilisateurs et des développeurs, en particulier en ayant un jeu unique de *GUIs* Radha et al. (2005) tel qu'illustre en figure 5 pour:

- *importer n'importe quel nouveau standard de métadonnées* en traduisant préalablement ses spécifications (formelles ou non) pour pouvoir les intégrer dans le MPD (pour le moment seules les spécifications au format XML Schemas sont importées automatiquement),
- mettre en place des *profils* des standards qui ont été importés,
- édition d'instances de métadonnées assistée et contrôlée par des composants complémentaires communs à tous les standards sous formes de GUIs conviviaux (en particulier pour les descriptions thématique et spatiale) (voir illustration en figure 5),
- un *moteur de recherche multi-critères générique*,
- l'*import/export* d'instances de métadonnées standardisées (généralement XML).

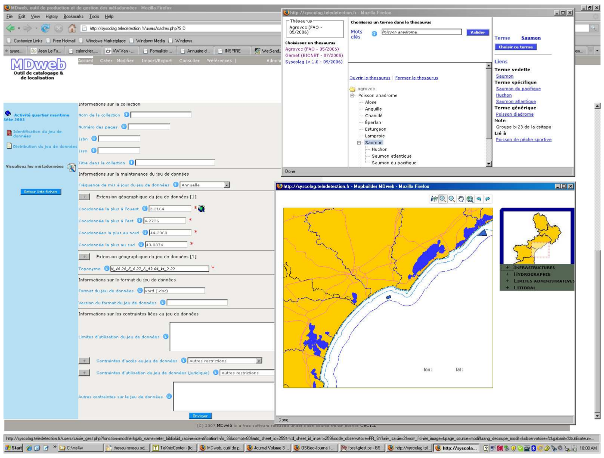

Figure 5: GUI pour l'édition d'instances de métadonnées

## Conclusion et perspectives

La localisation de RI peut être largement améliorée par une meilleure gestion des métadonnées grâce à l'implémentation d'approches génériques qui permettent la mise en place d'architectures uniques (GUIs, jeu de scripts, base de données) capables de gérer aussi bien:

- des *standards de métadonnées hétérogènes* (import, profils, édition...),
- des *valeurs hétérogènes*: en particulier des *termes* et des *descriptions spatiales* contrôlés pour décrire les *éléments de métadonnées principaux*,
- une *table d'indexation commune* qui duplique les *valeurs* associées aux *éléments de métadonnées principaux* sous une forme homogène visant à

améliorer l'efficacité du *moteur de recherche* (en évitant les wrappers), en particuliers par des mécanismes d'expansions de requêtes,

- les *RI spatiales décrites par les métadonnées* qui peuvent ainsi être traitées après les avoir localisées: soit localement soit à distance en utilisant des protocoles interopérables ou des clients riches (*WMS*, *Qgis*, *Udig*...).

Cette sorte d'architecture est pertinente pour satisfaire aussi bien les *besoins et les tâches des utilisateurs que ceux des développeurs* en *minimisant les efforts d'adaptation et de développement* tout en intégrant en même temps des outils complémentaires de contrôle des valeurs des éléments de métadonnées principaux.

La *localisation des RI* est ainsi améliorée. En particulier, en intégrant des outils de gestion des descriptions sémantiques et spatiales dans une même architecture, la localisation des données va pouvoir s'appuyer plus simplement sur des expansions de requêtes indispensables à la qualité des recherches. Il devient donc possible de se focaliser sur des cas d'utilisation type impliquant la gestion de relations spatiales et sémantiques comme: "trouver toutes les RI à moins de 1 mile de ce capteur (objet géographique) avec des mesures de tels paramètres physiques...", en s'appuyant en particulier sur des objets géographiques 2D ou 3D et des concepts, plutôt que des termes. De plus, en utilisant des relations standardisées (sémantiques par le *W3C* ou spatiales par l'*OGC*), les différentes requêtes deviennent réutilisables par d'autres systèmes/outils similaires.

Plus généralement, l'utilisation massive de standards de l'OGC et de logiciels qui les implémentent, accroit la capacité de l'application proposée à interopérer avec des systèmes extérieurs.

*Barde Julien*
*Monterey Bay Aquarium Research Institute (MBARI)*
`http://www.mbari.org/staff/julien/`
`julien@mbari.org`

# Bibliography

## Bibliography

L.M. Herborg, M.G. Bentley, A.S. Clare, S.P. Rushton (2003) The spread of the Chinese mitten crab (Eriocheir sinensis) in Europe; the predictive value of an historical data set. *Hydrobiologia* 503: 21-28.

World Wide Web Consortium (W3C) (2004) XML Schema Part 0: Primer Second Edition. *http://www.w3.org/TR/xmlschema-0/*.

World Wide Web Consortium (W3C) (2004) XML Schema Part 1: Structures Second Edition. *http://www.w3.org/TR/xmlschema-1/*.

World Wide Web Consortium (W3C) (2004) XML Schema Part 2: Datatypes Second Edition *http://www.w3.org/TR/xmlschema-2/*.

J. Barde (2005) Mutualisation de données et de connaissances pour la Gestion Intégrée des Zones Côtières. Application au projet SYSCOLAG. *universite Montpellier II* 285.

J. Barde, J. C. Desconnets, T. Libourel, P. Maurel (2006) Generic conceptual models for data and knowledge sharing. Application to environmental domain. *Hydroscience and Engineering, ICHE 2006* 16: 407-420.

Philip A. Bernstein and Laura M. Haas and Matthias Jarke and Erhard Rahm and Gio Wiederhold (2000) Panel: Is Generic Metadata Management Feasible? *The VLDB Journal* 660-662.

Chad Berkley, Matthew Jones, Jivka Bojilova, Daniel Higgins (2001) Metacat: A Schema-Independent XML Database System. *SSDBM '01: Proceedings of the Thirteenth International Conference on Scientific and Statistical Database Management* 171. IEEE Computer Society

Sergey Melnik, Erhard Rahm, Philip A. Bernstein (2003) Rondo: a programming platform for generic model management. *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data* 193-204. ACM Press

V. Radha, S. Ramakrishna, N. Pradeep Kumar (2005) Generic XML Schema Definition (XSD) to GUI Translator. *Distributed Computing and Internet Technology* 3816:290-296. IEEE Computer Society