

UN MODÈLE DE MÉTADONNÉES POUR GÉRER LES ÉVOLUTIONS DE DONNÉES GÉOGRAPHIQUES HÉTÉROGÈNES ET DISTRIBUÉES

Par Christelle PIERKOT
christelle.pierkot@ign.fr

EADS
6 voie l'Occitane - BP 87671
31676 LABEGE Cedex - France

Université Paul Sabatier - IRIT
118 route de Narbonne
31062 TOULOUSE Cedex 4 - France

IGN - Laboratoire Cogit
2/4 Avenue Pasteur
94165 St Mandé Cedex - France

Cet article est la traduction d'une communication acceptée à l'ACI 2005. Il correspond à l'avancée d'un travail de doctorat réalisé à EADS par Christelle Pierkot et coencadré par Messieurs Abdelkader Hameurlain de l'IRIT et Sébastien Mustière du Laboratoire COGIT de l'IGN.

Résumé

La mise à jour des données géographiques distribuées pose encore aujourd'hui de nombreux problèmes dus essentiellement aux caractéristiques spécifiques de ces données (composante spatiale, topologie...). Nous proposons ici un modèle de métadonnées dont le but est d'aider un système à gérer divers acteurs répartis sur plusieurs sites, manipulant des données hétérogènes mises à jour régulièrement et parfois indépendamment les unes des autres. Ce modèle est basé sur la norme ISO 19115 qui est la norme de métadonnées pour l'information géographique.

Introduction

La baisse du coût de production des données géographiques et l'évolution des techniques informatiques telle que la mise en réseau des systèmes ont considérablement facilité la collecte et la distribution des données nécessaires aux besoins des utilisateurs. Néanmoins, la maintenance et la mise à jour de ces données restent encore à ce jour un problème majeur.

Notre travail s'intéresse de manière générale à la gestion de la mise à jour de données géographiques distribuées. Un domaine applicatif particulier mais bien représentatif de notre contexte est le cas de la gestion des évolutions lors de missions militaires. Dans ce contexte, les acteurs sont répartis sur différents sites (en métropole, sur différentes parties du terrain d'action...) et chaque unité collecte, met à jour, enrichit, transforme et distribue ses propres données en fonction de ses besoins et de son système d'information. Dans un tel environnement distribué, les bases de données évoluent parallèlement. Or, pour prendre les bonnes décisions et mener à bien une mission, il faut que chaque unité ait une même vision de la zone d'intervention et possède toutes les informations disponibles. Pour y parvenir, il faut donc synchroniser régulièrement les données de chaque acteur, tout en minimisant les problèmes de cohérence.

Mais comment faire pour maintenir au mieux la cohérence d'un système constitué de multiples acteurs répartis sur des sites différents, manipulant des jeux de données géographiques hétérogènes qui peuvent évoluer parallèlement et différemment? Pour répondre à cette question, nous propo-

sons un modèle de métadonnées basé sur la norme ISO 19115 (ISO 19115, 2003) qui spécifie les métadonnées nécessaires à la description de l'information géographique. Notre modèle doit assurer la traçabilité des données hétérogènes distribuées et transformées, et aider à la gestion de mises à jour par différents acteurs. Ce modèle définit les relations entre les acteurs, les jeux de données et les évolutions afin d'en connaître la provenance et de déterminer certains critères essentiels à une intégration stable, comme la fiabilité (notion de confiance) ou la qualité (notion de précision et d'exactitude) des données et mises à jour.

Nous allons dans la prochaine partie soulever les principaux problèmes dus à la mise à jour des données géographiques, puis nous passerons en revue certaines solutions proposées. Nous présenterons ensuite dans la partie 3 la norme ISO 19115. Puis, nous présenterons plus précisément notre modèle dans la partie 4, en distinguant l'existant de l'ISO 19115 de ce que nous avons ajouté. Nous concluons enfin sur les perspectives de notre travail.

Données géographiques et mises à jour

La mise à jour des données géographiques reste à ce jour un problème majeur. En effet, à cause du caractère spécifique des données géographiques que nous évoquerons dans la partie suivante, une mise à jour nécessite encore souvent la livraison de la base entière pour pouvoir être prise en compte dans le système utilisateur. Quelques travaux de recherches ont été proposés pour combler ce manque, nous les évoquerons dans la seconde partie de ce paragraphe.

Hétérogénéité des données

Une des spécificités de l'information géographique est l'hétérogénéité des données qui sont utilisées :

- Une première particularité concerne la modélisation des données géométriques dans les systèmes d'informations. Il existe deux grands modes de représentation : les bases de données vectorielles et les bases de données rasters. Dans les bases de données vectorielles, les objets sont représentés par des points, lignes ou polygones, alors que dans les bases de données raster, l'espace est découpé en cellules élémentaires et les objets sont retrouvés grâce à l'ensemble des pixels le composant.
- Une autre particularité concerne les différents niveaux de détails qui sont utilisés pour représenter une même réalité géographique. En effet, il est souvent utile de posséder plusieurs vues d'un même espace, chacune apportant plus ou moins de précision en fonction du besoin (par exemple, pour planifier un vol aérien long courrier, l'utilisateur aura besoin d'une carte mondiale et d'une carte détaillée de la zone d'atterrissage).
- Il y a aussi le découpage géographique de l'espace. En effet, sur une même zone, plusieurs bases de données peuvent coexister et pour obtenir les informations souhaitées, il faut superposer ces bases. Un recalage est alors nécessaire pour mettre en correspondance les différentes parties.

Toutes ces différences de représentation du monde réel contribuent à l'hétérogénéité des données manipulées en information géographique et posent donc de nombreux problèmes lors de la mise à jour si on veut garder la cohérence entre toutes ces données.

Problématiques spécifiques à la mise à jour

La mise à jour des données géographiques s'avère particulièrement difficile à cause de certains aspects qui doivent être vérifiés tout au long du processus.

Ainsi, certaines contraintes sont à prendre en compte lors de la mise à jour afin que l'intégrité spatiale des données ne soit pas remise en cause. Par exemple, une route ne pourra traverser une rivière sans qu'un pont ne soit créé entre les deux.

La cohérence inter-thèmes est également un autre facteur à considérer. Les données sont souvent réparties en thèmes (hydrographie, routier...). Il est nécessaire, lorsqu'une mise à jour doit être intégrée, de connaître le thème sur lequel elle s'applique mais aussi de savoir sur quels autres thèmes il faudra la propager. Par exemple, la création d'une route peut engendrer une incohérence sur le thème bâtiment si la route traverse une maison. Il est donc nécessaire de vérifier et de corriger l'effet que produit l'intégration d'une évolution sur tous les thèmes concernés.

La différence de mode de représentation évoquée dans le paragraphe précédent peut engendrer tout autant certaines difficultés dans le processus de mise à jour. En effet, une mise à jour effectuée sur une image peut être propagée dans une base de données vecteur (par exemple, dans le cas où l'utilisateur manipule uniquement ce type de données). Il faut alors interpréter et transformer l'évolution pour pouvoir l'intégrer dans la base utilisateur.

La coexistence de différents niveaux de détails peut aussi être source d'incohérences lors de la mise à jour. On souhaite régulièrement propager des mises à jour effectuées à une certaine échelle sur une autre échelle, mais il est difficile de retrouver les données représentant le même phénomène à des échelles différentes. En effet, les données peuvent ne pas exister d'une échelle à l'autre (par exemple un hameau ne figurera pas sur une carte nationale mais apparaîtra sur une carte départementale), ou encore avoir été simplifiées de telle sorte que la représentation n'est plus la même (par exemple, sur une carte routière, un rond point peut être vu comme un simple point à une échelle départementale, et comme un carrefour complexe à une échelle urbaine).

Le découpage géographique appliqué sur le jeu de données source pose également quelques problèmes lors de la mise à jour des données. Prenons l'exemple d'une image découpée en plusieurs morceaux, représentant un espace entier. Une mise à jour peut avoir lieu sur plusieurs morceaux de cette image, et il faut donc pouvoir recoller les parties concernées entre elles. Le raccord est important pour la visibilité et la cohérence de l'image finale.

Le caractère distribué des données sur plusieurs sites pose également le problème de la mise à jour. En effet, les données évoluent parallèlement en fonction des propres caractéristiques de chaque système et lorsqu'une mise à jour doit être intégrée, des problèmes de cohérence (cohérence inter-thèmes, intégrité spatiale...) se posent entre les deux systèmes. La solution utilisée à ce jour pour contourner ces difficultés est souvent que le système le plus à jour renvoie à l'autre système l'intégralité de ses données sous la forme d'une base de données complète. Cependant, les évolutions effectuées par l'utilisateur sur son propre système sont alors perdues et doivent être « réinjectées » manuellement.

Quelques travaux en gestion des mises à jour géographiques

Pour tenter de venir à bout des divers problèmes évoqués dans les paragraphes précédents, plusieurs travaux ont été abordés selon différents points de vue. De nombreuses recherches sur les versions dans le domaine de l'information géographique ont été entreprises par Jomier et son équipe (Cellary et Jomier, 1990) (Jomier et al., 2001). Une thèse a également été effectuée sur le sujet par Peerbocus (Peerbocus, 2001) dans laquelle l'auteur utilise les bases de données multi-versions pour détecter automatiquement les conflits dus aux mises à jour. Il s'appuie sur le modèle formel de base de données multi-versions défini par (Gancarski, 1994) qui permet la gestion des différentes versions des entités en fonction du contexte. Une application de ces techniques dans les bases de données géographiques a aussi été proposée par Ding qui a défini un système de mise à jour incrémental pour la ville de New York qui s'appuie sur les modèles orientés objet et sur les mécanismes de version de bases de données (Ding et al., 2004).

Badard (Badard, 2000) propose d'établir des liens entre les différents thèmes des jeux de données et entre les données à différentes échelles afin d'avoir une correspondance préalable et de pouvoir ainsi l'utiliser lors de la mise à jour. Ces liens lui permettent de propager une évolution à l'en-

semble des données concernées directement par celle-ci. En s'appuyant sur les travaux de Badard et de Devogèle (Devogèle, 1997), le projet SGME (Raynal et al., 2001) propose de propager une évolution entre les jeux de données militaires ayant des échelles différentes, en s'appuyant sur les liens de correspondance et les liens inter-thèmes.

D'autres proposent l'utilisation de bases de données multi-représentations. Ces bases de données permettent de représenter plusieurs visions du même espace, soit à différentes échelles, soit selon différents points de vue, dans une seule et même base de données (Vangenot et al., 2002). Le but étant de passer d'une représentation à l'autre le plus facilement possible. Ainsi, Kilpeläinen propose d'établir des liens bidirectionnels entre les différents objets représentés dans de telles bases de données. L'idée étant qu'en conservant tous ces liens, il est plus aisé de retrouver un objet correspondant à une évolution quel que soit le niveau de représentation (Kilpeläinen, 2000)

Une autre solution est l'utilisation des bases de données fédérées. Une base de données fédérée est une vue commune de plusieurs bases de données qui permet une coopération entre ces bases. Il faut donc créer un schéma commun à toutes les bases de données, ainsi que les règles de passage permettant à un schéma particulier d'accéder à celui de la base fédérée. Christensen propose en particulier de créer une base de données géographiques fédératrice permettant de regrouper les caractéristiques communes à certains objets provenant de différentes collectes (Christensen, 2001)

Dans notre cas, on se situe dans le contexte où les données sont très hétérogènes et gérées de manières indépendantes. Il semble donc difficile d'utiliser une base de données fédérée. En effet, chaque système possède son propre modèle en fonction de ses propres besoins, et chaque base peut évoluer en parallèle, il est donc impossible d'avoir un schéma commun à toutes les bases. Il est aussi impossible d'utiliser une seule base de données multi-représentations à l'intérieur de l'infrastructure globale, en revanche, chaque sous-système peut posséder sa propre base multi-représentations. Nous devons donc utiliser des multi-bases (plusieurs bases de données hétérogènes ou non, capables d'interopérer sans vue commune). En revanche, nous pourrions nous inspirer des travaux de (Badard, 2000) pour établir des liens entre les différents jeux de données ainsi que ceux de (Gançarski et Jomier, 1994) pour les mécanismes de version appliqués aux bases de données géographiques.

Données géographiques et métadonnées

Une solution pour gérer au mieux ces bases multiples, hétérogènes et distribuées est d'utiliser les métadonnées qui permettent de fournir des renseignements précis sur les données et les acteurs les manipulant.

Depuis 2003, ISO 19115 (ISO19115, 2003) est la norme des métadonnées spécifique à l'information géographique. Elle a été établie par le comité technique 21 1 de l'Organisation Internationale de Normalisation (International Organization for Standardization, en anglais). Cette norme définit les éléments de métadonnées, fournit un schéma et établit une terminologie, des définitions et des procédures communes à l'ensemble des métadonnées nécessaires à

l'information géographique. Elle est divisée en packages, tous dépendants du package «ensemble d'information des entités de métadonnées» (Metadata entity set information, en anglais) qui est obligatoire. Ce package est représenté par la classe MD_Metadata dans le schéma UML associé à la norme. Cette classe décrit les métadonnées générales sur les ressources (domaine d'application, date de création, nom et version du standard de métadonnées utilisé...). A cette classe, se rattachent toutes les informations obligatoires ou optionnelles qui sont présentes dans cette norme (cf. figure1). On y trouve entre autres :

Des informations sur l'identification des données (MD_Identification, obligatoire). Par exemple la description des données, un aperçu, le mode de représentation spatial utilisé ...

Des informations sur la qualité des données et des jeux de données (DQ_Quality, optionnel). Cette classe est divisée en deux pour fournir d'une part des informations de généalogie sur le producteur ainsi que le processus utilisé pour créer les données (LI_Lineage spécialisé en LI_Source et LI_ProcessStep) et d'autre part des informations quantitatives sur la qualité telles que la précision ou encore la cohérence des données (DQ_Element).

L'étendue et la fréquence des mises à jour (MD_MaintenanceInformation, optionnel). On y trouve des informations sur la fréquence, l'étendue et la date de la prochaine mise à jour. Cette classe permet aussi à l'utilisateur de choisir la période envisagée pour les futures mises à jour.

Des informations sur le distributeur des données et les options possibles pour obtenir les ressources (MD_Distribution, optionnel). Cette classe permet de connaître le média de stockage des données, utile pour savoir si les ressources sont disponibles sur le réseau ou non. Elle renseigne également sur le distributeur, le coût et la disponibilité d'un jeu de données.

Mais également, des informations sur l'étendue spatiale, temporelle et verticale du jeu de données (EX_Extent) ou encore sur les contraintes associées aux données (MD_Constraint, optionnel) où l'on trouve les restrictions d'usage sur les jeux de données (copyright, licence,...) ainsi que sur le niveau de confidentialité des données (confidentiel, top secret, ...). Et enfin, une description des informations de références (CI_Citation), qui fournit des informations sur le nom, la date de référence, la date d'édition ou encore la version de la ressource.

Enfin, un package permettant d'étendre la norme en fonction du besoin spécifique de l'utilisateur (MD_Metadata ExtensionInformation, optionnel), en particulier le nom, la définition et les conditions d'utilisation des nouveaux éléments de métadonnées, a également été prévu.

D'autres informations sont également disponibles mais ne sont pas détaillées ici car elles ne concernent pas directement notre problématique.

Il est important de noter que cette norme permet à l'utilisateur de connaître la provenance et la qualité des données en fonction du processus de production utilisé pour leur création. Des informations sur le rythme des mises à jour peuvent être disponibles, mais elles sont essentiellement utili-

sées pour définir la fréquence à laquelle les mises à jour doivent être appliquées et non pour gérer des évolutions qui arrivent en flux continus.

Un cas particulier de la norme ISO 191 15 est le format de fichier de métadonnées METAFOR. C'est une implémenta-

tion en XML d'un profil de la norme ISO 191 15 et des normes associées pour les besoins en métadonnées de la Défense française. Selon ces besoins, le profil défini étend et restreint la norme. Il ne concerne pas encore les données vectorielles. Aucune information sur les données d'évolutions n'y est à ce jour prise en compte.

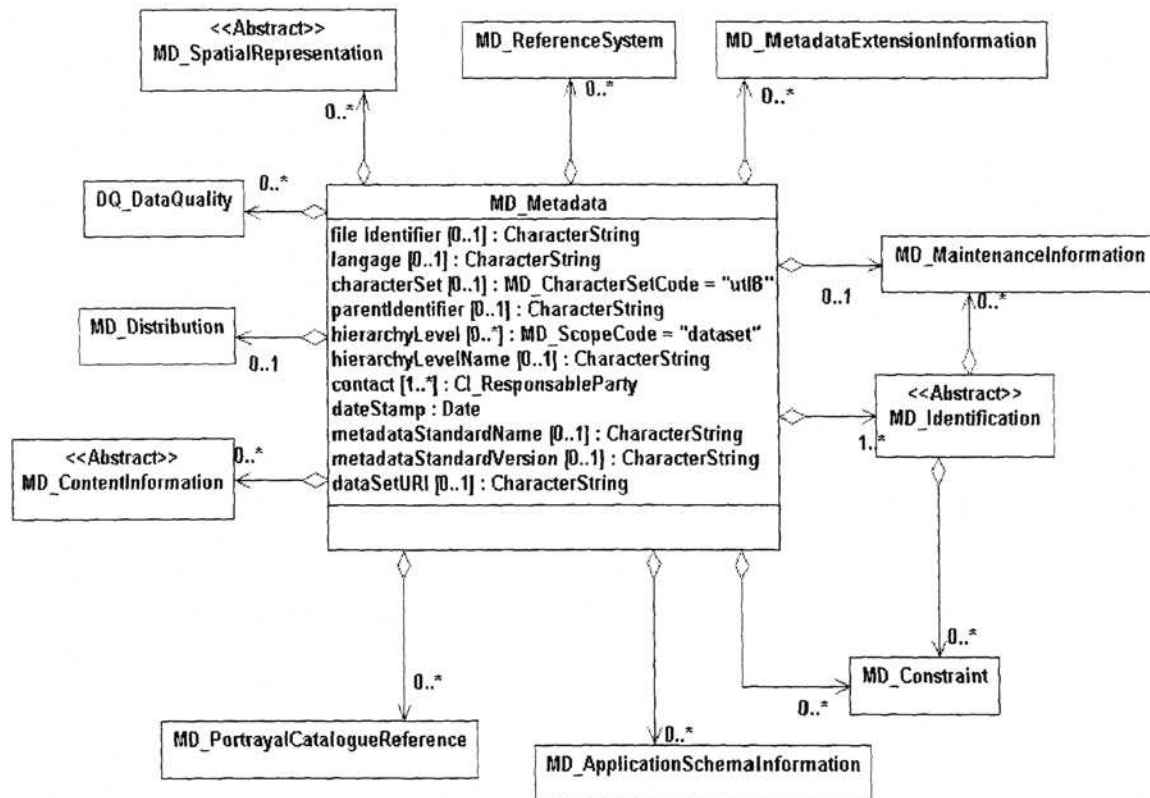


Figure 1 : Ensemble des entités de métadonnées dans ISO 19115

Modèle «Données, Acteurs, Evolutions»

De nombreux acteurs manipulent des jeux de données de types et de qualités variés, qui possèdent des modèles, des spécifications, des niveaux de détails et des formats différents et qui sont stockés sur des sites distribués. Ces acteurs doivent coopérer les uns avec les autres. En particulier, ils doivent échanger et synchroniser leurs données sans mettre en péril leur propre système et le système global (perte d'informations, incohérence...).

Les données peuvent être collectées, mises à jour, enrichies, transformées et redistribuées en fonction des besoins des utilisateurs. Chaque acteur possède son propre jeu de données (qui peut éventuellement dériver d'un autre jeu de données), qui évolue en même temps que les jeux de données des autres acteurs. Pour obtenir un résultat cohérent et utilisable, il est souvent nécessaire de valider interactivement les mises à jour.

Pour maintenir au mieux la cohérence d'un système contenant des données variées distribuées qui évoluent en parallèle, nous devons retrouver facilement certaines informations sur les jeux de données et d'autres sur les mises à jour. Il est nécessaire de connaître par exemple la source des jeux de données et des évolutions (qui les a envoyées), leur qualité (comment elles ont été acquises), sur quelle zone elles s'appliquent, la date à laquelle elles ont été créées ou modifiées.

Nous suggérons donc de définir un modèle de métadonnées pour gérer non seulement les jeux de données mais aussi les mises à jour et les acteurs. Ce modèle doit permettre d'obtenir facilement des informations sur les jeux de données (version, date actualité...), les acteurs (rôle, fiabilité...), les évolutions (format de livraison, qualité...) et sur les relations entre les jeux de données, acteurs et évolutions (collecteur des évolutions, jeu de données sur lequel s'appliquent les mises à jour...) afin d'aider à l'intégration des évolutions dans l'infrastructure entière mise en place.

Ce modèle étant un modèle de métadonnées, nous proposons qu'il s'appuie sur la norme ISO 19115. Certaines informations présentes dans l'ISO ont été reprises dans notre modèle, d'autres (notamment en ce qui concerne les évolutions) ont été ajoutées afin de répondre aux questions posées par la problématique. Nous le présentons ci-dessous en quatre parties : une vue générale des relations entre les acteurs, les jeux de données et les évolutions, puis le détail de chacune de ces trois entités. Dans les schémas ci-dessous, nous distinguons l'existant de ce que nous avons ajouté grâce à des textures différentes (fond blanc pour les classes ayant été ajoutées ou modifiées par rapport à l'ISO, fond de couleur pour celles qui n'ont pas changées).

Structure générale

Les jeux de données, acteurs et évolutions sont étroitement liés (cf. figure 2). Un jeu de données appartient au moins à un acteur et peut dériver d'un autre jeu de données

(restriction de la zone, changement de schéma, changement d'échelle...). Un acteur possède au moins un jeu de données qu'il peut acquérir de différentes façons (fournisseurs extérieurs, numérisation directe ou traitements particuliers sur des jeux de données existants chez d'autres acteurs). Il peut aussi produire ou recevoir des évolutions. On entend par «produire» le fait que c'est l'acteur lui-même qui saisit les évolutions (par exemple grâce à un levé de terrain), par «recevoir», le fait que les évolutions proviennent d'un autre acteur (collecte de nouvelles données par exemple). Les acteurs peuvent donc échanger de l'information entre eux, selon certains critères que nous définirons plus précisément dans la modélisation des acteurs. Dans ce modèle général, nous considérons les évolutions comme un ensemble contenant plusieurs évolutions élémentaires ayant été saisies ou collectées pour un jeu de données particulier. Les évolutions s'appliquent donc au moins sur un jeu de données. L'ensemble constitué des évolutions peut contenir des nouvelles données et des mises à jour, aucune distinction n'est faite à ce niveau. Les évolutions appartiennent au moins à un acteur, celui qui les a produit.

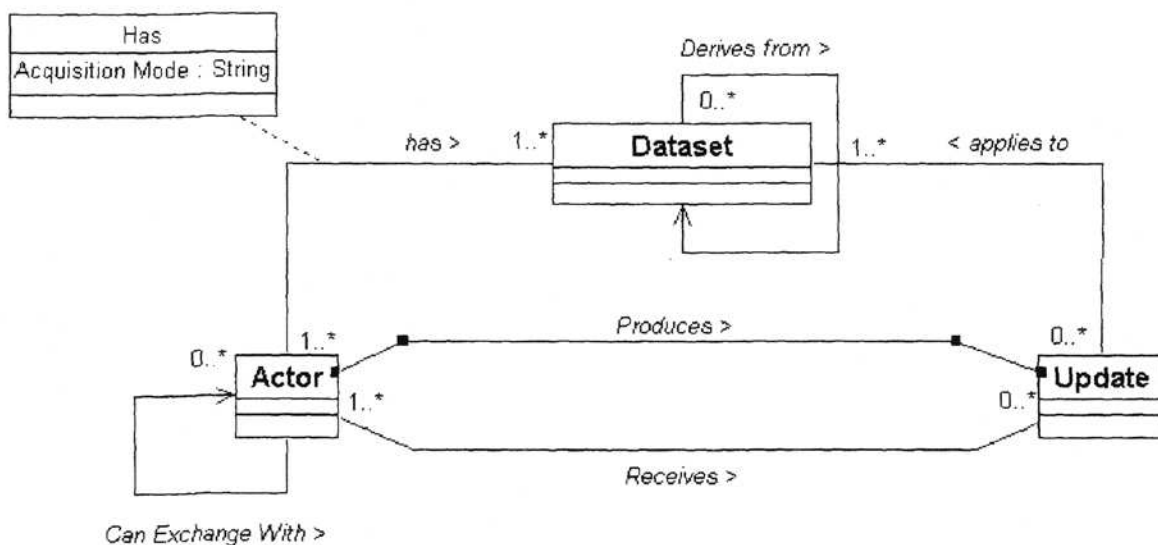


Figure 2 : Modèle général

Modélisation des acteurs

Les acteurs ne sont pas modélisés dans l'ISO 19115. Nous proposons donc un modèle permettant la gestion des échanges qui interviennent dans un contexte distribué, mais également qui définit les interactions possibles entre plusieurs acteurs répartis sur des sites différents.(cf. figure 3)

(Delavar et al., 2003) définit une infrastructure comme étant une «sorte d'organisation» qui est la structure de base à d'autres organisations. Dans notre modèle, nous considérons qu'une infrastructure globale est une organisation qui produit, transforme, utilise, met à jour et diffuse des données géographiques. Cette infrastructure

globale est divisée en infrastructures locales réparties sur différents sites.

Nous introduisons également deux types d'acteurs : les acteurs internes et les acteurs externes. Les acteurs internes font partie d'une infrastructure locale et peuvent échanger leurs données et évolutions avec les autres acteurs, en fonction de leur rôle dans l'infrastructure globale. Les acteurs externes quant à eux ne peuvent que fournir des nouvelles données ou mises à jour et ne peuvent aucunement recevoir des informations des autres acteurs. Dans notre contexte militaire, nous pouvons considérer que l'infrastructure globale est l'organisation toute entière mise en place pour exécuter la mission. Les infrastructures locales sont les unités déployées sur le terrain ou en métropole.

Les acteurs internes sont les militaires français qui utilisent les données et les acteurs externes sont les alliés déjà présents sur la zone d'intervention, qui possèdent des informations supplémentaires qu'ils veulent partager avec l'armée française.

En fonction de leur rôle, les acteurs internes peuvent échanger leur données et mises à jour plus ou moins facilement. Le mode de transmission doit également être fourni afin que les acteurs sachent comment se procurer les données. Ces indications nous permettent de savoir qui peut envoyer des évolutions et qui peut les recevoir. Par exemple, un acteur ayant pour rôle celui de simple utilisateur peut uniquement recevoir des données et des mises à jour mais

n'a pas le droit d'en envoyer aux autres acteurs. Si en plus, cet acteur a une liaison bas débit, il lui faut connaître le volume de données qu'il doit transférer afin de connaître le temps de téléchargement et de s'assurer que celui-ci ne sera pas trop important.

Pour que les échanges soient fiables, nous devons ajouter des informations sur la confiance que l'on peut accorder aux acteurs qui fournissent les données et mises à jour. En effet, la fiabilité est un critère important car elle permet à l'utilisateur de définir s'il a un intérêt ou non à acquérir les données qu'on lui propose. La confiance peut également servir à aider lors de l'intégration des mises à jour qui sera plus ou moins assistée en fonction du degré accordé.

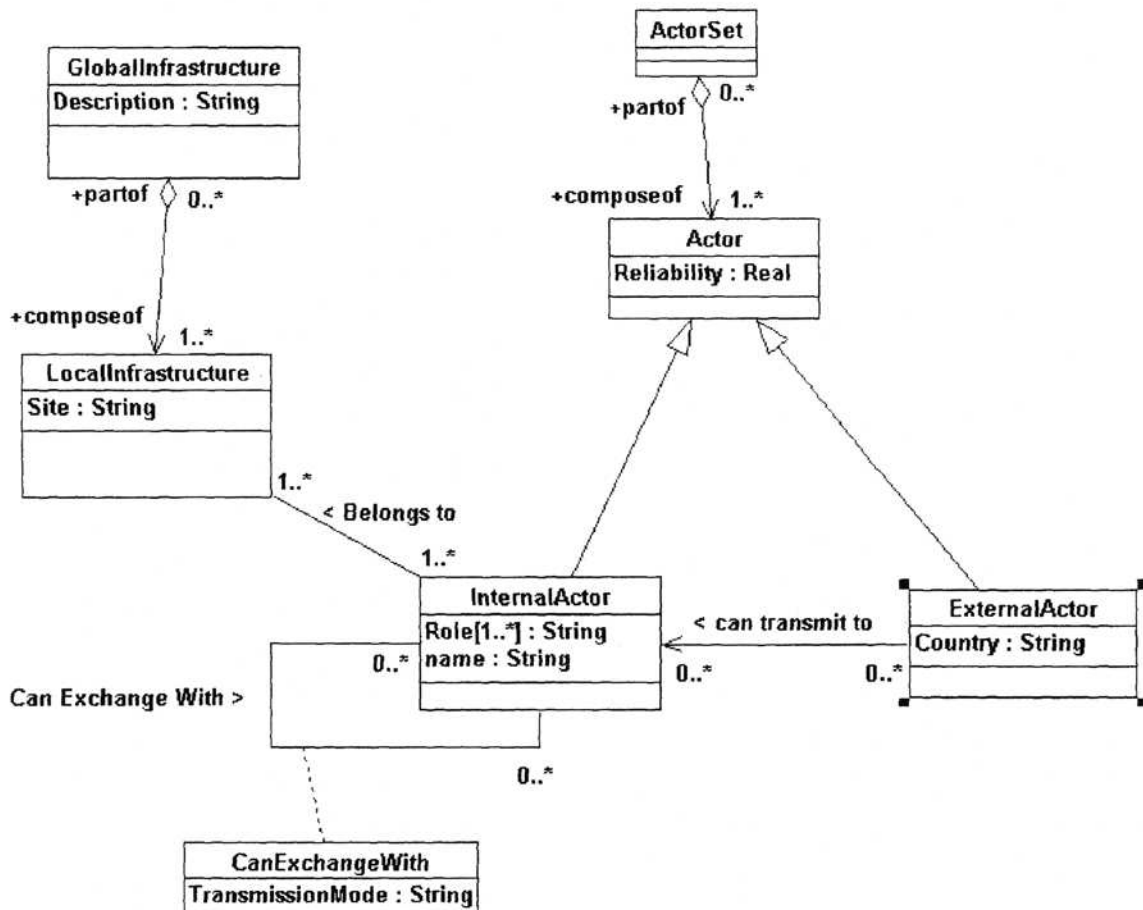


Figure 3 : Modélisation des acteurs

Modélisation des jeux de données

Les jeux de données sont définis dans l'ISO 191 15 et dans METAFOR à travers la notion de DS_Dataset. Néanmoins, dans notre contexte particulier, certaines informations nécessaires ne sont pas prévues, et nous les avons introduites.

Les jeux de données manipulés en information géographique sont de deux types :

les bases de données vectorielles et rasters. Ils sont représentés par les classes MD_GridSpatialRepresentation et MD_VectorSpatialRepresentation dans la norme ISO 191 15 et sont donc repris dans notre schéma (cf. figure 4)

La qualité des jeux de données est une information importante, elle est étroitement liée à la période d'acquisition des produits (période de crise, mission de reconnaissance), ainsi qu'au mode d'acquisition (à partir de levé terrain, d'images satellites...). Ces informations sont disponibles par le biais de la classe DQ_Quality existante dans l'ISO 191 15.

Des informations de généalogie sont également indispensables pour connaître la source des jeux de données et leur historique depuis leur création. Supposons que l'on possède une évolution provenant d'un jeu de données plus ancien que celui dans lequel on veut l'intégrer. Il est alors nécessaire de retrouver le jeu de données correspondant à l'évo-

lution ainsi que les modifications effectuées sur celui-ci afin d'intégrer correctement cette évolution. De telles métadonnées existent dans ISO 191 15 et sont disponibles dans LI_Lineage, qui est une classe agrégée à la classe DQ_Quality.

Un des manques de la norme lorsqu'on veut manipuler des jeux de données mis à jour régulièrement par des acteurs différents et distribués sur le réseau est la notion de version. Il existe cependant un attribut dans la classe CI_Citation permettant de renseigner sur la version d'une ressource, mais aucun lien n'est établi entre les différentes versions des jeux de données. Nous devons donc définir un mécanisme de version reliant les jeux de données entre eux.

L'attribut « actualité » sert quant à lui à savoir si notre jeu de données est à jour ou non, et si des mises à jour plus récentes sont disponibles chez d'autres acteurs. Deux attributs « date » existent dans la classe CI_Citation de l'ISO 19115 : le premier donne la date de référence de la ressource et le second la date d'édition (date de la version). Ces deux attributs peuvent être réutilisés pour nous donner l'actualité du jeu de données.

L'attribut « état » permet de savoir rapidement si un jeu de données a été transformé depuis son acquisition. Cet attribut n'existe pas et doit être rajouté à la norme.

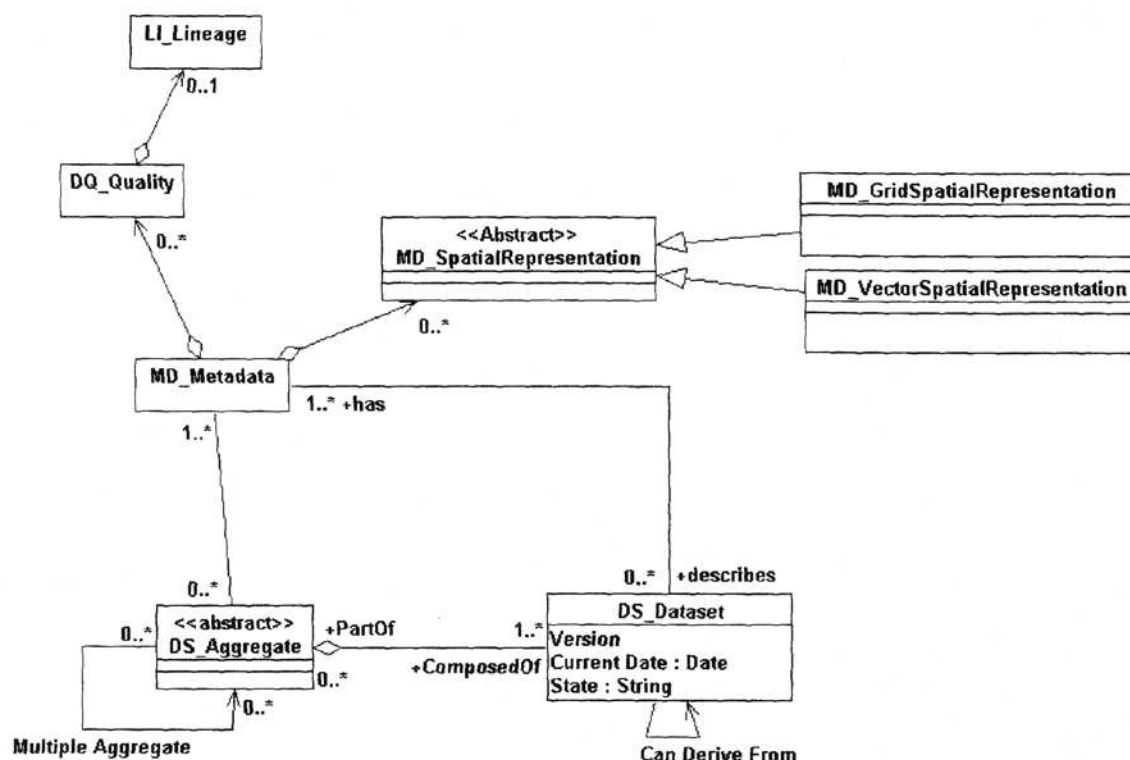


Figure 4 : Modélisation des jeux de données

Modélisation des évolutions

Les évolutions ne sont pas modélisées en tant que telles dans l'ISO 19115. Nous avons donc défini un modèle permettant de gérer, utiliser et diffuser les mises à jour (cf. figure 5).

Un ensemble d'évolutions est défini par une zone géographique, une unité d'échange et par une liste d'évolutions élémentaires effectuées sur un jeu de données particulier. Pour faciliter l'intégration des évolutions, nous devons ajouter de l'information sur les ensembles d'évolutions mais également sur chaque évolution. Cela est fait grâce aux métadonnées. Ce modèle utilise certaines classes existantes dans l'ISO, notamment pour définir la qualité des données.

T.Badard et D.Richard (Badard, 2000; Badard et Richard, 2001) ont défini les lots différentiels et les lots d'évolutions pour modéliser les ensembles des évolutions :

Les lots différentiels fournissent les données de mise à jour en identifiant les objets qui ont évolué et en donnant éventuellement des informations sur la façon dont chaque objet a évolué. Les lots différentiels ne gèrent que les objets anciens et nouveaux et la mise à jour de la géométrie ne se fait que par une succession de création et de destruction. Ils sont liés à la façon dont les objets sont gérés dans la base, en particulier à la gestion d'identifiants. Si les évolutions sont modélisées en lots différentiels, un attribut « DataSource » doit alors être renseigné pour chaque évolution élémentaire, permettant de connaître l'objet dans son état ancien et l'objet dans son état nouveau. Le type de l'évolution élémentaire ne sera dans ce cas qu'une création ou une destruction.

Dans les lots d'évolution, les objets sont une interpréta-

tion directe des mises à jour effectuées, c'est-à-dire une identification de la nature de l'évolution et dépendent donc de la façon dont les mises à jour ont été effectuées. La modélisation en lot d'évolutions implique de renseigner le type de l'évolution élémentaire : cela peut être une création, une destruction, une modification géométrique ou une modification sémantique dans le cas des données vectorielles ou encore la substitution d'une partie d'une image par une autre plus récente dans le cas des données raster.

Ces deux modes de représentation permettent de définir des formats de livraison adaptés au transfert des évolutions. L'unité d'échange de ces modes de livraison est étroitement liée aux données manipulées, ce peut être en XML ou GML (Geographic Markup Language) (OpenGIS, 1999) pour les données vectorielles ou encore un format spécial pour les données raster (JPEG 2000 ou GeoTiff).

Les évolutions peuvent être transformées pour être intégrées dans un jeu de données ayant des caractéristiques différentes de celui ayant servi d'appui à la collecte de la mise à jour (différent schéma, différente échelle...). Pour assurer un suivi des mises à jour, nous proposons d'utiliser des mécanismes de versions sur les ensembles d'évolutions, comme cela a déjà été proposé pour les jeux de données. Cet ajout nous permet de connaître précisément l'historique des mises à jour, de leur création jusqu'à leur intégration dans les différents systèmes (avec ou sans transformation). On a donc ainsi une trace des transformations que les évolutions ont subies depuis leur création. La qualité doit être fournie pour une évolution particulière mais également sur l'ensemble des évolutions, notamment pour savoir dans quelles conditions les mises à jour ont eu lieu. Cette qualité doit pouvoir nous renseigner sur le type d'acquisition des évolutions (par exemple, une saisie ou une livraison).

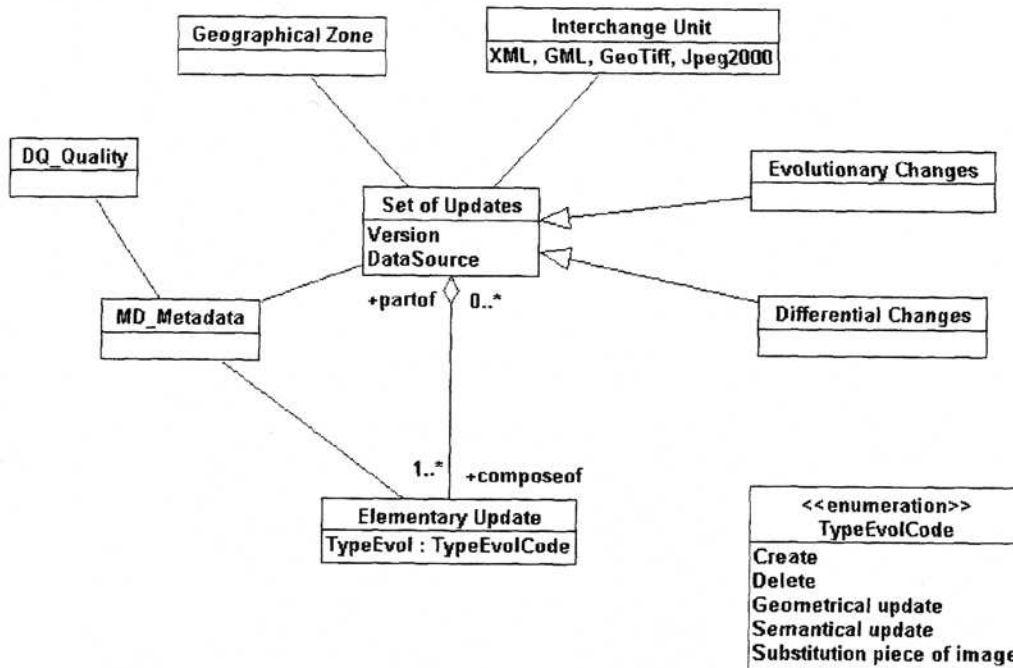


Figure 5 : Modélisation des évolutions

Conclusions et perspectives

Nous avons présenté un modèle de métadonnées permettant de gérer les jeux de données et leurs évolutions, manipulés par plusieurs acteurs distribués sur des sites distants. Ce modèle est basé sur la norme ISO 19115 et garantit de fait un certain niveau d'interopérabilité, facilitant sa mise en oeuvre dans d'autres contextes. Il permet aux acteurs de connaître aisément la provenance et la qualité des jeux de données et des évolutions. Cette connaissance est primordiale pour intégrer efficacement les données dans un système d'information géographique. Il permet également de relier les jeux de données aux évolutions et vice versa afin d'avoir une vue générale sur l'évolution de l'infrastructure globale.

Nos travaux futurs porteront sur le raffinement de ce modèle en nous inspirant des travaux en ingénierie coopérative afin de modéliser au mieux les interactions entre tous les acteurs et sur les recherches en infrastructure de données spatiales afin de gérer au mieux les échanges de données géographiques et de leurs évolutions.

Ce travail de thèse commencé en juin 2004 est financé par un contrat de collaboration de recherche et d'encadrement CIFRE entre la société EADS Defense and Security Systems SA, l'Université Paul Sabatier de Toulouse et le laboratoire Cogit de l'Institut Géographique National. Il intervient dans le cadre du second volet «Dynamique et Cohérence» du projet ENVOL effectué pour la Délégation Générale pour l'Armement. Bibliographie

Bibliographie

- Badard T., 2000, Propagation des mises à jour dans les bases de données géographiques multi-représentations par analyse des changements géographiques, Thèse de doctorat, Université de Marne la vallée.
- Badard T. et Richard D., 2001, Using XML for the exchange of updating information between geographical information systems, Computers, Environment and Urban Systems, 2001, Vol 25, Oxford, Elsevier Science Ltd., p. 17-31.
- Cellary W. et Jomier G., 1990, Consistency of Versions in Object Oriented Databases, Proceedings of the 16th International Conference on Very Large Data Bases, Brisbane, Queensland, Australia, August 1990, p. 432-441.
- Christensen A., 2001, Issues in the Conceptual Modeling of Geographic Data, PhD Dissertation, The Danish Research Agency and The National Survey and Cadastre.
- Delavar, M., Rajabifard, A., and Rezayan, H., 2003, NSDI and IT Evolution. Proceedings of the Map Asia Conference, 2003.
- Devogèle T., 1997, Processus d'intégration et d'appariement des Bases de Données Géographiques : Application à une base de données multi-échelles, Thèse de doctorat, Université de Versailles.
- Ding J, Ahearn S. et Cooper E., 2004, An incremental Geographic Update System for a large Geographic Database in NY City, Proceedings of the Agile Conference.
- Gançarski S., 1994, Versions et bases de données : modèle formel, support de langage et d'interface utilisateur, Thèse de doctorat, Université Paris Sud.
- Gançarski S. et Jomier G., 1994, Un formalisme pour la Gestion de Versions D'entités dans leur Contexte Actes des journées Bases de Données Avancées, Clermont Ferrand, France.
- ISO 19115 : 2003, ISO/TC 211 Geographic Information — Metadata. 2003.
- Jomier G., Cellary C., Gançarski S. et Manouvrier M., 2001, Bases de données et Internet : Modèles, langages et systèmes. Chapitre 8 : les versions, Paris, Hermès Sciences, Éditions Lavoisier, p. 235-255.
- Kilpeläinen T., 2000, Maintenance of Multiple Representation Databases for Topographic Data, The Cartographic Journal, Vol 37 N°2, December 2000, p.101-107.
- OpenGis Consortium, 1999, Geographic Markup Language (GML). OGC RFC 1113 December 1999, edited by Lake, R., 37 pages.
- Peerbocus A., 2001, Gestion de l'évolution spatio-temporelle dans une base de données géographiques, Thèse de doctorat, Université de Paris Dauphine, 2001.
- Raynal L, Badard T. et Braun A., 2001, Synthèse de l'étude sur la conception d'un serveur géographique multi-échelles, rapport SGME/2500/017, v2.0, octobre 2001.
- Vangenot C., C.Parent et Spaccapietra S., 2002, Modeling and Manipulating Multiple Representation of Spatial Data, in Proceedings of 10th Spatial Data Handling Symposium, Ottawa

LA RECHERCHE GÉOGRAPHIQUE D'INFORMATION SUR LE WEB: BESOINS ET ÉVALUATION

Bénédicte BUCHER¹, Paul CLOUGH², Hideo JOHO², Ross PURVES³ et Awase Khirni SYED³

¹ Laboratoire COGIT - Institut Géographique National {benedicte.bucher@ign.fr}

² Department of Information Studies, University of Sheffield, UK {p.d.clough;h.joho@sheffield.ac.uk}

³ Department of Geography, University of Zürich, Switzerland {rsp;sak@geo.unizh.ch}

Résumé

La recherche géographique d'information (RGI) sur le Web est un domaine récent qui évolue rapidement. Le développement de ces applications nécessite une analyse des besoins auxquels elles doivent répondre et ainsi que des méthodes permettant d'évaluer comment elles y répondent. Cet article décrit l'analyse des besoins ayant conduit au développement de l'application SPIRIT et la méthode proposée pour son évaluation. Cette évaluation doit porter d'une part sur le comportement absolu du système et d'autre part sur ce comportement perçu par l'utilisateur final. Nous développons une collection particulière de documents pour faciliter la mesure des performances de SPIRIT ainsi qu'une grille de notation de la pertinence spatiale et thématique d'un document en réponse à une requête. Nous soulignons l'importance d'intégrer dans l'évaluation de systèmes RGI les interactions de l'utilisateur avec le système autant que les performances absolues du système.

Introduction

La recherche géographique d'information (RGI) sur le Web est définie par (Larson, 1996) comme "l'aide à l'accès à des sources d'information localisée". Actuellement, l'expression *recherche d'information* (RI) renvoie généralement à la recherche de documents répondant à une requête parmi une collection importante et non structurée de documents textuels stockée sur le Web. Dans ce contexte, nous limitons la recherche géographique d'information à la recherche de documents répondant à une requête de la forme <thème,

relation spatiale, localisation> parmi une collection importante et non structurée de documents textuels stockée sur le Web. Dans cette requête, le thème et la localisation sont liés par une relation spatiale d'inclusion, topologique ou directionnelle, par exemple "châteaux dans le pays de Galle", "châteaux proches du pays de Galle" ou "châteaux au nord du pays de Galle". Les travaux en RGI sont conduits par des universitaires et également par des sociétés commerciales. Par exemple, Google a développé récemment un moteur de recherche dit "local" qui s'appuie sur des annuaires commerciaux (<http://local.google.co.uk/>).

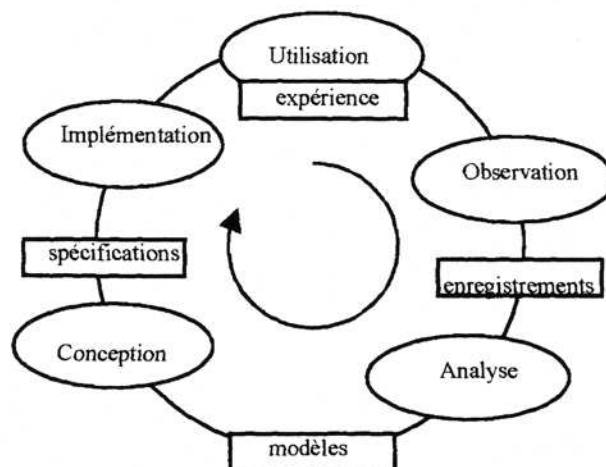


Figure 1: Activités et objets dans un processus de développement (d'après Henderson 1991, p262)

Le projet européen SPIRIT (Spatially aware information retrieval on the internet) veut développer une application de RGI qui exploite les éléments thématiques et géographiques contenus dans les documents Web. Il adopte une démarche itérative similaire à celle illustrée sur la figure 1 qui accorde une grande importance à l'utilisateur. Ce cycle est parcouru plusieurs fois lors du développement de l'application. Avant l'implémentation, les étapes d'utilisation, observation, analyse et conception correspondent à une *analyse des besoins* conduisant à l'*établissement de spécifications fonctionnelles et de spécifications de conception*. Dès lors qu'un prototype est disponible, ces mêmes étapes correspondent à l'*évaluation*. Cet article présente l'analyse des besoins ayant présidé au développement de l'application de RGI SPIRIT et les spécifications de cette application. Nous détaillons ensuite l'évaluation de telles applications et en particulier de l'application SPIRIT.

Analyse des besoins pour SPIRIT

La méthode d'analyse des besoins utilisée dans le projet SPIRIT se compose de deux éléments. D'une part, des maquettes ont été construites et associées à des scénarios d'utilisation imaginés par les membres du projet. Ces maquettes ont été présentées à des utilisateurs potentiels qui ont ensuite exprimé leurs besoins en termes d'interactions et de fonctionnalités d'un système similaire à celui illustré par les maquettes. D'autre part, une analyse des applications existantes qui fournissent certaines de ces fonctionnalités a été conduite. Par exemple, nous avons analysé les modalités d'expression d'un lieu dans des applications comme Local Google et Multimap (www.multimap.com). Ces analyses d'une part des besoins et d'autre part des solutions partielles disponibles ont résulté en un ensemble de fonctionnalités appelant des solutions innovantes de la part du projet SPIRIT.

Un besoin clé est celui d'effectivement trouver des documents sur le Web qui soient spatialement et thématiquement pertinents en réponse à une requête composée d'un thème, d'une relation spatiale et d'une localisation. Ce besoin renvoie à plusieurs fonctionnalités. La première est l'interprétation d'un nom de lieu exprimé par l'utilisateur. Cela peut demander une levée d'ambiguïté, par exemple entre London, UK, et London, Ontario, ou encore la gestion de lieux vagues comme le sud de l'Angleterre (Purves et al., 2005). La fonctionnalité suivante consiste à transmettre la requête à un moteur de recherche qui intègre des techniques pour gérer les composantes thématiques et géographiques de la requête et ordonner les documents y répondant (V an Kreveld et al., 2004).

Les utilisateurs ont également insisté sur l'importance d'associer une carte à la présentation des réponses, surtout lorsque la pertinence spatiale d'un document n'est pas claire

pour l'utilisateur. Ce besoin surgit généralement lorsque l'utilisateur ne connaît pas en détail le lieu sur lequel porte sa requête, par exemple une personne recherche des maisons d'hôte au nord de Leicester mais ne connaît pas suffisamment la région pour savoir si une localité se trouve effectivement au nord de Leicester. Fournir à l'utilisateur une carte localisant sa requête ainsi que les documents Web obtenus en réponse lui permet de juger de la *pertinence spatiale* des résultats, c'est-à-dire d'évaluer visuellement la relation spatiale spécifiée dans sa requête entre le lieu exprimé et les *empreintes géographiques* des documents. Les utilisateurs ont aussi exprimé le besoin de préciser leur requête de façon itérative en spécifiant sur la carte une nouvelle zone d'intérêt.

Certains utilisateurs recherchent enfin non seulement des documents Web mais aussi des données géographiques. Par exemple, un utilisateur recherchant des informations sur "la randonnée dans les Alpes" aimerait obtenir des données altimétriques sur cette région.

L'application SPIRIT

L'analyse des besoins menée dans SPIRIT a conduit à la définition de spécifications initiales pour notre application. L'application se compose de plusieurs modules. L'interface utilisateur supporte l'expression de la requête et la présentation des résultats. Le moteur de recherche et le module de classement recherchent et classent les documents répondant à une requête de la forme <thème, relation spatiale, localisation>. Enfin, une ontologie géographique stocke et gère les connaissances concernant la sémantique et la géométrie des lieux (Jones et al., 2004). L'utilisateur interagit avec cette architecture uniquement via l'interface utilisateur (Purves et al., 2005).

L'interface permet à l'utilisateur d'exprimer un besoin de deux façons : soit en formulant une requête structurée soit en dessinant une requête graphique. Dans le cas d'une requête structurée, illustrée sur la figure 2, l'utilisateur spécifie un nom de lieu, comme Edinburgh. Dans le cas d'une requête graphique, il dessine un polygone sur la carte proposée par l'interface. Les autres modules composant SPIRIT, comme le module de classement, implémentent également plusieurs méthodes. Par exemple, le module de classement implémente plusieurs mesures de pertinence et peut classer les résultats soit uniquement selon leur pertinence thématique soit selon diverses combinaisons de la pertinence spatiale et de la pertinence thématique.

Lors de l'évaluation, il sera nécessaire non seulement d'évaluer la qualité des résultats obtenus avec ces différentes modalités ou méthodes, mais aussi de les comparer pour déterminer quelles techniques sont les plus efficaces.



SPIRIT

Structured Query

About SPIRIT

Search for

castles

In

wales United Kingdom

Town or City name Region name Country name parameters in components

search

You are searching for [castles](#) in [wales](#) [united kingdom](#)

- 1 [Data Wales: Some pictures of the Sealed Knot Socie...](#)
- 2 [Medjugorje Cards](#)
- 3 [Paintings to Order](#)
- 4 [Medjugorje Churches](#)
- 5 [Untracost](#)
- 6 [The Castles of Wales](#)
- 7 [Pentecost Cards in Welsh](#)
- 8 [landscape pictures of snowdonia national park nord...](#)
- 9 [Medjugorje Tapes CD's](#)
- 10 [Castles in Wales](#)

Figure 2: Interface de SPIRIT supportant l'expression de la requête structurée <castles, in , wales>. SPIRIT propose en réponse une liste ordonnée de documents associée à une carte les localisant.

L'évaluation d'applications de recherche géographique d'information

La recherche géographique d'information (RGI) est un domaine en constante évolution. Peu de systèmes de RGI sont basés sur les techniques de recherche d'information (RI) car ils sont pour la plupart des systèmes de « recherche géographique d'information géographique » et non de recherche de documents Web. A notre connaissance, il n'y a pas de méthode d'évaluation proposée pour des applications comme SPIRIT qui soit adaptée à ces deux aspects : la recherche d'information parmi une collection non structurée de documents textuels et le traitement spécifique de l'information géographique lors de cette recherche. Cela dit, comme de tels systèmes émergent peu à peu, il sera de plus en plus crucial de disposer de méthodes d'évaluation suffisamment génériques pour pouvoir s'appliquer à tous, permettre de les comparer et de déterminer les techniques les plus efficaces.

Par contre, l'évaluation a été beaucoup étudiée dans le domaine de la recherche d'information (RI). La complexité de l'évaluation dans ce domaine réside essentiellement dans la difficulté d'appréhender et de mesurer la pertinence d'un document pour un utilisateur formulant une requête. En effet, cette pertinence est souvent subjective, voire controversée (Saracevic, 1975). Deux stratégies d'évaluation ont été

peu à peu développées dans la littérature en RI (Spark Jones and Willett, 1997:167-174): l'évaluation orientée système et l'évaluation orientée utilisateurs.

L'évaluation orientée système vise à mesurer les performances du système de façon la plus standard et objective possible (Borlund, 2003, Van Rijsbergen, 1979). Cette évaluation permet de comparer des applications RI ou différentes implémentations de modules d'une même application. Elle met en place un banc d'essai simulant des tâches de recherche d'information en l'absence d'utilisateurs. Ce banc d'essai consiste en une ressource standard appelée collection test, qui se compose des éléments suivants :

- un ensemble de documents Web représentant un domaine ou le Web, D,
- un ensemble de requêtes représentant les tâches utilisateurs de façon réaliste et contrôlable (Peters, 2001: 1069), R,
- des notes de pertinence pour chaque document Web et chaque requête, P. La construction de ces notes s'appuie sur une grille de notation qui doit guider la définition de notes de pertinence de façon la plus objective et générique possible. L'attribution de ces notes est l'étape la plus longue de la constitution d'une collection test.

Ainsi faite, une collection test simule le résultat idéal de tâches utilisateurs. Ses deux premiers éléments, D et R, peuvent être utilisés pour reproduire ses tâches à l'aide d'une application de RI dont on veut mesurer les performances.

Les mesures sont obtenues en comparant les documents trouvés par l'application au sein de D pour répondre à chaque requête de R avec les notes de pertinences appartenant à P.

Deux grandes mesures résument les performances d'une application de RI et servent de base au calcul d'autres mesures : la précision et le rappel. Ces mesures sont classiquement définies pour une collection dans laquelle les notes de pertinence sont binaires : pour une requête et un document donnés, le document est pertinent ou ne l'est pas. Dans ce contexte, la précision mesure le nombre de documents pertinents retournés par l'application parmi les documents retournés par cette application. A ce stade, il faut souligner qu'une application peut retourner un nombre très important de documents et que les documents les moins bien classés seront rarement consultés par l'utilisateur. Aussi, une mesure plus intéressante est la "précision à n", où n vaut par exemple 10. La précision à n vaudra donc : le nombre de documents pertinents parmi les 10 premiers documents trouvés par l'application. Le rappel mesure lui le nombre de documents retournés parmi les documents pertinents. Des mesures de précision et rappel ont été proposées pour s'adapter à des notes de pertinence plus détaillées (Kekäläinen and Järvelin, 2002). Cela permet une plus grande expressivité, en particulier dans le domaine de la recherche d'information parmi une collection de documents XML structurés (Kazai et al., 2004), ou dans le domaine de la recherche d'images (Clough et al., 2005).

La conception de collections test standard a commencé il y a 40 ans sous l'impulsion de Cleverdon (1967) et est depuis lors un élément de référence de l'évaluation d'applications de RI, mis en oeuvre par exemple dans les campagnes de la Text REtrieval Conference (TREC). Dans de telles campagnes d'évaluation, les participants utilisent la même collection test –ensemble de documents Web et ensemble de requêtes– pour mesurer les performances de leurs systèmes de façon comparable. Ces campagnes étudient également les interactions utilisateurs.

L'évaluation orientée utilisateurs impliquent les utilisateurs, au contraire de l'évaluation orientée système. Elle prend son sens lorsque l'application a une interface utilisateurs. Ces dernières années, le développement d'applications de RI offrant une plus grande interactivité a remis en cause la stratégie purement orientée système pour évaluer ces applications (Borlund, 2003). Il est maintenant conseillé d'adopter une stratégie complémentaire orientée utilisateurs pour mesurer en quoi l'application, globalement, favorise le processus de recherche d'information mené par l'utilisateur. Cette évaluation étudie d'une part des éléments spécifiques de l'interface et d'autre part l'utilisabilité globale du système par les utilisateurs, c'est-à-dire la possibilité pour les utilisateurs de réaliser leurs tâches de recherche d'information et la facilité avec laquelle ils la réalisent. Borlund (2003) propose d'évaluer des systèmes de RI interactifs en s'appuyant sur un scénario d'utilisation qui couvre les principales tâches utilisateurs. L'utilisateur suit le scénario et évalue la pertinence des documents qu'il obtient. Cette pertinence est relative au contexte défini par le scénario. Il est également possible de mettre en place des interviews pour approfondir par exemple l'évaluation de l'utilisabilité du système. Une littérature importante existe dans ce domaine, que ce soit sur l'évaluation de logiciels en général ou de systèmes de RI (e.g. Bawden, 1990). En définitive, concernant l'évalua-

tion orientée utilisateurs, il n'existe pas de forte spécificité liée au domaine de la RI et cette évaluation se fait comme pour les logiciels en général.

L'évaluation dans SPIRIT

L'évaluation dans SPIRIT se compose de l'évaluation orientée système et de l'évaluation orientée utilisateurs. L'évaluation orientée utilisateur s'appuie sur des scénarios comme proposé par (Borlund 2003) ainsi que sur un questionnaire. Les scénarios et le questionnaire donnent une place importante aux activités de l'utilisateur relatives à l'information spatiale, comme l'expression des aspects spatiaux de sa requête, l'appréhension des aspects spatiaux d'un document Web ou l'interprétation du classement spatial.

L'évaluation orientée système présente plus de spécificités. Pour la mettre en place, il faut définir une collection test c'est-à-dire : un ensemble de documents Web, un ensemble de requêtes, une grille de définition de notes de pertinence d'un document relativement à une requête et les notes de pertinence affectées à chaque document pour chaque requête. La conception de cette collection test doit prendre en compte plusieurs facteurs :

- l'évaluation doit mesurer tout particulièrement la gestion des aspects spatiaux de la recherche d'information,
- l'évaluation doit être adaptée aux limites du prototype SPIRIT en termes de documents Web et de données géographiques. En effet, les connaissances géographiques de SPIRIT sont organisées dans l'ontologie géographique construite à partir de données géographiques. L'alimentation de cette ontologie a été limitée aux données géographiques que nous avons pu acquérir dans le contexte du projet. Par ailleurs SPIRIT ne fonctionne pas sur le Web mais sur une copie partielle du Web acquise en début de projet auprès de Google. Nous détaillons ci-dessous la définition des éléments composant la collection test de SPIRIT.

Les documents Web retenus dans la collection test, c'est-à-dire l'élément D introduit dans la définition d'une collection test, doivent d'une part être représentatifs des documents sur lesquels SPIRIT fonctionnera et d'autre part être en nombre suffisamment restreint pour que la production des notes de pertinence soit possible. Ces documents ont été extraits d'une copie partielle du Web d'1 Terabyte acquise auprès de Google au début du projet SPIRIT (Joho and Sanderson, 2004). En prévision de l'évaluation des capacités spatiales de SPIRIT, nous avons pris soin de garder dans cet ensemble de documents Web les réponses à des requêtes impliquant une relation spatiale. Par exemple, concernant la requête « pubs near Glencoe », nous avons intégré dans cette collection les documents trouvés par un moteur de recherche classique fonctionnant sur cette copie pour des requêtes textuelles « pub, x » où x sont les noms de localités proches de Glencoe. Les requêtes utilisées pour construire cette collection test sont bien plus nombreuses que celles utilisées pour l'évaluation. Pour les générer, nous avons dans un premier temps utilisé les noms des 200 plus grandes villes du Royaume Uni et gardé pour chacune les 50 premiers documents renvoyés par le moteur de recherche textuel. Le résultat consistait en 9 010 documents et 85 MB de texte. Dans un deuxième temps, nous avons utilisé des requêtes correspondant à des besoins d'utilisateurs tels "Arts festivals dans Edinburg" qui nous semblaient repré-

sentatifs des besoins pour un système comme SPIRIT . Nous avons choisi de privilégier des localisations dans le Royaume Uni tout en utilisant d'autres localisations comme Montreux en Suisse.

La définition des requêtes composant la collection test de SPIRIT, c'est-à-dire l'élément R introduit dans la définition d'une collection test, a été faite en considérant les requêtes pour lesquelles les moteurs classiques n'apportent pas une satisfaction suffisante à l'utilisateur, c'est-à-dire les requêtes correspondant à un besoin de moteur de recherche géographique d'information. Il s'agit de requêtes de la forme <thème, relation spatiale, localisation> ayant les propriétés suivantes :

- Le nom de lieu est aussi un mot commun davantage utilisé, comme Battle (nom de lieu anglais signifiant par ailleurs bataille).

- Le nom de lieu est ambigu, comme Paris au Texas.

- Le nom de lieu utilisé dans la requête est peu susceptible d'être utilisé dans les documents réponse, comme dans <blog, près de, Vélieux>.

- Le nom de lieu renvoie à une région imprécise, comme le Sud de la France. Dans ce cas, l'application doit avoir une intelligence spatiale complexe pour formaliser l'empreinte géographique de la requête.

- Le thème de la requête est aussi un nom de lieu, comme Forêt noire.

- La relation spatiale n'est pas une relation d'inclusion, comme <maison d'hôte, au nord de, Auxerre>.

Une première grille de définition de notes de pertinence d'un document relativement à une requête a été proposée. Elle est présentée ci-dessous. Une note est composée de deux variables : la pertinence thématique et la pertinence spatiale. Chaque variable est ternaire pour distinguer plusieurs niveaux de pertinence. Notons que la pertinence spatiale est dépendante non seulement de la localisation mais aussi du thème car l'échelle d'une carte « locale » sur laquelle situer le document est dépendante de la granularité du thème. Ainsi selon que je recherche des informations de randonnée ou de météorologie, le niveau de détail des informations de localisation dont l'utilisateur a besoin ne sera pas le même.

Pertinence thématique

1. Le document contient de l'information pertinente relativement à la requête et permet à l'utilisateur d'appréhender cette pertinence sans l'aide d'autres connaissances. Autrement dit, le lien du document avec le thème est suffisamment clair et ne nécessite pas d'expertise pour être établi.
2. Le document contient de l'information pertinente mais il faut consulter d'autres ressources pour former un jugement sur cette pertinence. Par exemple, le document mentionne le thème mais se contente de pointer vers une ressource sur ce thème. Ou encore le document a un lien avec le thème mais ce lien n'est pas simple à appréhender pour un non expert et il faut s'aider de ressources plus simples pour le comprendre.
3. Le document n'a pas de lien avec le thème de la requête.

Pertinence spatiale

1. Le document fait référence à une localisation répondant aux critères de la requête et cette localisation est suffisamment détaillée dans le document pour que l'utilisateur puisse la retrouver sur une carte locale.
2. Le document fait référence à une localisation répondant aux critères de la requête mais il n'y a pas suffisamment d'information de localisation pour que l'utilisateur puisse la retrouver sur une carte locale.
3. Le document ne fait pas référence à une localisation répondant aux critères de la requête.

Tableau 1: Grille initialement définie dans SPIRIT pour noter la pertinence d'un document Web en réponse à une requête de la forme <thème, relation spatiale, localisation>.

Nous avons testé l'utilisabilité de cette grille en nous appuyant sur 5 requêtes décrites dans le tableau 2. Pour chacune, les 10 premiers documents renvoyés par SPIRIT ont

été proposés à 11 évaluateurs qui ont dû les noter conformément à la grille. Précédemment, des exemples de jugements ont été donnés aux évaluateurs pour illustrer la grille.

1. Caving in Derbyshire (UK)
2. Castles in Wales (UK)
3. Skiing near Glencoe (UK)
4. Art festivals in Edinburgh (UK)
5. Music in Montreux (Switzerland)

Tableau 2: Requêtes utilisées pour tester la grille de notation de pertinence.

Ces évaluateurs ont répondu à un questionnaire pour indiquer leur compréhension de la grille et commenter l'utilisation qu'ils en ont faite. En général, ils ont trouvé la grille simple à comprendre mais ont eu des difficultés à attribuer des notes de pertinence spatiale. Il est ainsi apparu logiquement qu'un évaluateur doit avoir une bonne connaissance de la localisation de la requête pour attribuer une note de pertinence spatiale lorsque cette note vaut 2. Par exemple, il existe une piste de ski près de Glencoe appelée « White Corries ». Si la localisation Glencoe ne figure pas sur un document relatif à cette piste, il faudra lui attribuer une note de pertinence spatiale de valeur 2 mais seul un évaluateur connaissant le voisinage de Glencoe et en particulier cette piste de ski le saura. Ainsi, pour construire les notes de pertinence il faut que les évaluateurs aient une bonne connaissance de la localité sur laquelle ils effectuent leur requête. Il faut également tenir compte des indices de localisation disponibles dans le document, qui ne sont pas toujours sous la forme d'un nom de lieu ou d'une adresse. Par ailleurs, les évaluateurs ont trouvé inutile la distinction entre une pertinence thématique de 2 et de 1.

Nous avons par ailleurs mesuré la cohérence entre les notes attribuées par les 11 évaluateurs en utilisant une mesure de Cohen Kappa. La valeur de cohérence obtenue pour les notes spatiales et pour les notes thématiques est de 95%, ce qui est largement satisfaisant.

Conclusion et perspectives

Cet article a décrit l'analyse des besoins et une démarche de mise en place d'une méthode d'évaluation pour une application de recherche géographique d'information, SPIRIT. L'évaluation de telles applications nécessite l'adaptation de techniques existantes d'évaluation d'application en RI. Cela comprend en particulier le développement de nouvelles grilles de notation de pertinence. La définition de ressources standard dans ce domaine aiderait à déterminer quelles sont les techniques les plus efficaces et donc faciliterait le progrès en RGI. Dans le contexte de SPIRIT, une première grille de notation de pertinence a été proposée et testée. Nous avons utilisé pour l'évaluation finale une nouvelle grille ne conservant que deux niveaux de pertinence thématique (pertinent / non pertinent) et conservant les 3 niveaux de pertinence spatiale de la grille initiale. Des mesures de précision et rappel de SPIRIT ont pu être faites par la suite à l'aide de cette collection test et ont montré l'apport des techniques de SPIRIT à la RGI par rapport à un moteur de recherche purement textuel. De plus, ces applications impliquent généralement une forte interactivité avec l'utilisateur de sorte que leur évaluation ne peut faire l'économie d'une évaluation orientée utilisateurs. Celle-ci peut être menée en s'appuyant sur des scénarios d'utilisation et un questionnaire. A l'avenir, nous pensons que la formalisation de la notion de pertinence d'une information en RGI devra s'appuyer sur la modélisation d'activités spatiales et de l'« affordance » des objets géographiques dans ces activités (Jordan et al., 1998).

Remerciements

Ces travaux sont financés en partie par le projet européen No. IST -2001-35047 (SPIRIT) et par le BBW Suisse (01.0501).

Références

- Bawden, D. 1990. *User-orientated evaluation of information systems and services*, Gower Publishing company .
- Borlund, P. 2003. *The IIR evaluation model: a framework for evaluation of interactive information retrieval systems*. In: *Information Research*, vol. 8, no. 3, paper no. 152.
- Cleverdon, C. The Cranfield test on index language devices: In Sparck Jones, K. & Willett, P. eds. (1997) *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann. Pp. 47-59.
- Clough, P., Mueller, H. and Sanderson, M. (2005), *The CLEF 2004 Cross Language Image Retrieval Track, In Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, Eds (Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M. and Magnini, B.), *Lecture Notes in Computer Science (LNCS)*, Springer, Heidelberg, Germany (in print).
- Henderson A. 1991. *A Development Perspective on Interface Design and Theory* in J.M. Carroll (Ed): *Designing Interaction: Psychology at the Human Computer Interface*. Cambridge, Cambridge University Press: 254-268.
- Joho, H., and Sanderson, M. 2004. *The SPIRIT Collection: an overview of a large web collection*. *SIGIR Forum*, 38(2).
- Jones, C.B., Abdelmoty, A.I., Finch, D., Fu, G. & Vaid, S. 2004. *The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing*. In *Proceedings of the 3rd International Conference on Geographic Information Science (GIScience 2004)*, Maryland, USA. LNCS.

Jordan, T. M. Raubal, B. Gartrell, and M. Egenhofer . 1998. An Affordance-Based Model of Place in GIS. in: T. Poiker and N. Chrisman (Eds.), 8th Int. Symposium on Spatial Data Handling, SDH'98, Vancouver, Canada, pp. 98-109.

Kazai, G., Lalmas, M., De Vries, P. 2004. The overlap problem in content-oriented XML retrieval evaluation. SIGIR 2004: 72-79

Kekäläinen, J. & Järvelin, K. 2002. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology* 53(13), 1120-1129.

Larson, R.R. 1996. Geographic Information Retrieval and Spatial Browsing. In *GIS and Libraries: Patrons, Maps and Spatial Information*, Linda Smith and Myke Gluck, Eds., University of Illinois.

Purves, R.S., Clough, P. and Joho, H. 2005. Identifying imprecise regions for geographic information retrieval using the web. In *Proceedings of GISRUUK*, 2005.

Saracevic, T. 1975. Relevance: a review of and a framework for the thinking on the topic. *Journal of the American Society for Information Science*, vol. 26: 321-343

Sparck Jones, K. & Willett, P. eds. 1997. *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann.

Van Kreveld, M., Reinbacher, I., Arampatzis, A. and Van Zwol, R. 2005. Multi-Dimensional Scattered Ranking Methods for Geographic Information Retrieval. *Geoinformatica*, 9,1, 61-84.

Van Rijsbergen, C. J. *Information Retrieval*, Butterworth-Heinemann, Newton, MA, 1979 (<http://www.dcs.gla.ac.uk/Keith/Preface.html>)

Voorhees, E.M. 2001. Overview of TREC 2001. In *Proceedings of TREC 2001*, NIST.
