Conception d'un système d'information géographique résilient pour l'environnement

Ba-Huy Tran, Christine Plumejeaud-Perreau, Alain Bouju, Vincent Bretagnolle

Conception d'un système d'information géographique résilient pour l'environnement

Ba-Huy Tran [1], **Christine Plumejeaud-Perreau** [2], **Alain Bouju** [1], **Vincent Bretagnolle** [3]

1. *Laboratoire Informatique Image et Interaction (L3I), Université de la Rochelle*
   av. Michel Crépeau
   17000 La Rochelle, France, France
   {ba-huy.tran,alain.bouju}@univ-lr.fr
2. *LIttoral ENvironnement et Sociétés (LIENSs), U.M.R. CNRS 7266*
   2 rue Olympe de Gouges
   17000 La Rochelle, France
   christine.plumejeaud-perreau@univ-lr.fr
3. *Centre d'Etudes Biologiques de Chizé (CEBC), U.M.R CNRS 7372 et Université de la Rochelle*
   79360 Villiers en Bois, France
   breta@cebc.cnrs.fr

ABSTRACT. *This paper presents the mapping between heterogeneous data sources as a solution for designing a resilient GIS for environment. From a case study conducted on Chizé environmental observatory, we summarize the difficulties encountered by biologists and ecologists experts when maintaining various environmental data acquisition devices and analyzing these data. In this example, crop rotation and wildlife data are recorded since 1994 in various spatio-temporal databases. We show how the implementation of a semantic mapping of these sources can solve challenges raised by the analysis and the maintenance of theses always evolving and heterogenous systems. In particular, the demonstration of the feasibility of such a system is made, and we measure its ability to answer complex queries combining several data sources and the spatial and temporal dimensions.*

RÉSUMÉ. *Cet article présente le mapping entre sources de données hétérogènes comme une solution aux problèmes que posent la conception d'un Système d'information géographique pour l'environnement qui soit résilient. A partir d'un cas d'étude mené sur un observatoire environnemental basé à Chizé, nous résumons les difficultés rencontrés par des experts biologistes et écologues lorsqu'ils maintiennent divers dispositifs d'acquisition de données environnementales et souhaitent analyser les données de façon conjointe. Dans cet exemple, des données*

d'assolement et de faune et flore sont enregistrées depuis 1994 dans diverses bases de données spatio-temporelles. Nous montrons comment la mise en oeuvre d'un mapping sémantique de ces sources peut résoudre les difficultés d'analyse et de maintenance qu'induisent ces systèmes, amenés à de constantes évolutions de leurs modèles. En particulier, la démonstration de la faisabilité d'un tel système est faite, et nous mesurons sa capacité à répondre à des requêtes complexes mêlant plusieurs sources de données et les dimensions spatiales et temporelles.

KEYWORDS: *Mapping, Ontology, Integration, Environment, Ecology, GIS, Semantic web*

MOTS-CLÉS : *Mapping, Ontologie, Intégration, Environnement, Ecologie, SIG, Web sémantique*

DOI:10.3166/HSP.1.1-15

1. Introduction

La nécessité de collecter des observations sur une longue durée pour la recherche sur les relations entre environnement et anthroposystème a entrainé la mise en place de Zones Ateliers[1] par le CNRS. Elles questionnent les interactions entre un milieu et les sociétés qui l'occupent et l'exploitent. La spécificité des Zones Ateliers réside dans la taille de l?objet d?étude, qui est un territoire de grande dimension.

Ainsi, le Centre d'Etudes Biologiques de Chizé (CEBC) a mis en place un observatoire des assolements et de la biodiversité sur la "Zone Atelier Plaine et Val de Sèvre". Cet observatoire collecte de nombreuses données sur la biodiversité faunistique et floristique, ainsi que l'assolement agricole, avec un suivi de cette même zone sur plus de 20 ans dans diverses bases de données spatio-temporelles. Cependant, l'observation de la rotation des cultures, des insectes de tout type (carabes, libellules, etc.), des plantes, ou des oiseaux ne requière ni les même moyens, ni les mêmes méthodes. Ainsi les protocoles de collecte de ces informations diffèrent forcément, ils font l'objet de projets distincts qu'assurent diverses équipes de recherche au CEBC, constituant à terme des bases de données très hétérogènes. Pourtant il existe un fort besoin d'être en mesure de croiser ces informations, de façon assez systématique et souple, pour mener une analyse spatio-temporelle fine des milieux écologiques.

A cet effet, les utilisateurs envisagent la migration des données et schémas dans un système d'information centralisé. Sinon, il faut mettre en oeuvre une médiation entre ces sources de données (Vargas-Solar, Doucet, 2002). En terme de médiation, une des solutions les plus avancées et prometteuse repose sur l'intégration par les techniques du Web sémantique (Hacid, Reynaud, 2004) ou (Raffaetà *et al.*, 2008).

Nous avons expérimenté ces deux approches dans le cadre opérationnel de la mise en place d'un Système d'Information Géographique pour l'Environnement (SIG-E) sur la zone atelier, ouvert et libre sur le plan logiciel, et libérant les données tout en respectant les restrictions juridiques afférantes à celles-ci. La première section expose en détail nos besoins en termes d'interopérabilité autour du SIG-E, ainsi que les avan-

1. http://www.za-inee.org

tages et inconvénients de la migration dans une base relationnelle spatiale centralisée. La seconde section propose le mapping dynamique entre sources hétérogènes comme une alternative qui s'inspire du Web sémantique. La troisième section démontre la faisabilité et les possibilités qu'offre cette nouvelle approche appliquée à notre cas d'étude. Elle en expose aussi les limites. La dernière section résume les enseignements tirés de ces expériences, et expose les perspectives de cette recherche.

2. Nécessité d'une solution par médiation souple autour du SIG-E

L'observatoire de la "Zone Atelier Plaine & Val de Sèvre" constitue notre cas d'étude. Il couvre une zone de 450 km^2 au sud de la ville de Niort, dans le département des Deux-Sèvres en Poitou-Charentes, France. Il s'agit essentiellement d'une plaine céréalière intensive: céréales, maïs, tournesol, pois et colza où l'élevage (bovins, caprins) est encore présent mais en forte baisse. Les parcelles agricoles sont encore de taille modeste (4-8 ha) et 15% d'entre elles sont occupées par des prairies (artificielles, permanentes ou temporaires), contre 60% en 1970. Il a pour objet de recherche la relation entre les pratiques agricoles et les processus écologiques, à travers l'étude de l'évolution de l'organisation spatiale du paysage. L'enjeu de notre recherche est d'offrir les moyens d'une analyse spatio-temporelle fine des données collectées.

2.1. Les bases de données de la Zone Atelier Plaine et Val de Sèvre

Nous présentons ici deux des bases de données qui doivent être croisées de façon prioritaire pour les écologues.

2.1.1. La base "Assolement"

L'organisation spatiale du paysage évolue dans le temps parce que les agriculteurs modifient l'assolement de leurs parcelles chaque année, mais également recomposent parfois les parcelles entre elles (par des scissions, fusions ou des redistributions), changeant ainsi les formes des parcelles. Depuis 1994, les occupations du sol sont donc relevées annuellement sur le terrain et numérisées sur les 19 000 parcelles agricoles par l'équipe Agripop du CNRS de Chizé. Ces données sont centralisées dans une base de données nommée "Assolement". Son modèle de données se fonde sur le paradigme Space-Time composite proposé par (Langran, Chrisman, 1998). L'idée consiste à ne pas stocker la géométrie de chaque parcelle pour chaque année, mais à utiliser dans le modèle de petites géométries (appelées ici "micro-parcelles") obtenues par l'intersection de toutes les parcelles au cours de la période d'observation. La géométrie de toutes les parcelles peut être reconstruite "à la volée" pour chaque année en utilisant une composition des micro-parcelles constituant la parcelle.

L'observation sur le long terme nécessite aussi un système opérationnel avec des interfaces conviviales pour les utilisateurs, similaire aux solutions proposées par les Systèmes d'Information Géographique (SIG). Les utilisateurs doivent enregistrer chaque année à la fois le nouveau type d'utilisation du sol de chaque parcelle, mais aussi

éventuellement des changements de formes pour chaque parcelle. La première solution logicielle qui a été développée pour le site en 2007 était basée sur la solution d'un SIG (ArcGIS) couplée avec une base de données Access, programmée avec des scripts VBA. Implémentant le modèle composite « Space-Time » décrit précédemment, le système a permis depuis l'acquisition et l'analyse des données de rotation des cultures.

2.1.2. La base "Faune et Flore"

Parallèlement, des données de faune et flore sont collectées sur le terrain depuis plusieurs années par l'équipe AGRIPOP de Chizé et centralisées dans une autre base de données, "Faune et Flore", qui est structurée suivant un schéma relationnel spatial, implémenté dans PostgreSQL[2] avec PostGIS. Ces données, ponctuelles et datées, proviennent des différents chercheurs qui rapportent leurs observations concernant 600 espèces, principalement des oiseaux, et certaines plantes, grâce à une interface Web. Pour les oiseaux, la base constitue une collection d'observations décrivant le comportement des espèces observées ainsi que leurs nids, et leur contexte (hauteur de végétation, date, heure, localisation, etc.).

2.1.3. La base "Insect" et les autres bases

Il existe par ailleurs d'autres données structurées sur différentes espèces, souvent dans des tableurs, ou bien des bases de données Access. Il serait souhaitable de pouvoir interroger et croiser aussi ces sources avec la connaissance de l'assolement et de la faune avicole. C'est le cas des données relatives à l'observation des carabes, petits coléoptères auxilliaires des champs et très sensibles à la qualité des milieux, qui bénéficient d'un suivi depuis 9 ans dans la base Access "Insect".

2.2. Expression des besoins autour de l'analyse de la biodiversité

Avec les données déjà disponibles, un nombre conséquent d'analyses peut-être mené. Ces analyses peuvent en premier lieu servir à vérifier le corpus de données. Or, concernant l'assolement, les experts décrivent un certains nombres de règles de succession permettant d'écarter ou de corriger des valeurs douteuses. Ainsi un tournesol ne succède jamais à du colza. Ce sont là des règles de succession temporelle sur un même espace. Ces règles peuvent être contrôlées par une requête sur "Assolement" renvoyant l'historique d'occupation des parcelles.

L'analyse cherche aussi à confirmer des heuristiques pressenties par l'observation, et à affiner leur paramètres. Elles concernent par exemple les préférences des animaux par type et forme d'assolement :

– l'outarde canepière aime une variété de petites parcelles alentours.

2. http://www.postgresql.org/

– le busard cendré met son nid dans des blés dont les parcelles proches contiennent de la luzerne ou ont contenu ou contiendront probablement de la luzerne ou des prairies dans les quelques années passés et à venir. Ainsi il peut chasser des mulots dans les prairies proches pour ses oisillons.

Ces règles impliquent de pouvoir raisonner sur la configuration spatiale (premier exemple) ou spatio-temporelle (second exemple) des observations. Elles exigent aussi de pouvoir facilement croiser les bases de données "Assolement" et "Faune & Flore". Pour l'instant, la seule solution connue des expérimentateurs est l'importation dans des systèmes d'analyse qui leur permettent de travailler sur des extractions des données au prix de nombreuses manipulations, et remise en forme pour les logiciels d'analyse ciblés comme R[3], ou Fragstat[4]. Ils ne peuvent donc mener systématiquement et aisément ces analyses.

2.3. Migration d'un système d'acquisition propriétaire vers un système libre

Aujourd'hui, en vue d'une ouverture du système d'information sur le Web, et afin de faciliter les analyses transversales entre évolution du paysage et variations de la biodiversité, il est envisagé d'intégrer dans la base de données "Faune et Flore" les informations d'Assolement. En effet, les nouveaux produits ESRI (ArcGIS depuis la version 9.0) ne sont plus programmables avec VBA mais avec Python, ce qui empêche toute évolution de la base de données Assolement. En effet, si les données sont stockées dans Access, l'application VBA gérait jusqu'ici un grand nombre de règles métier et de contraintes non spécifiées avec Access. Et il n'est pas envisageable de migrer les quelques 50 000 lignes de code VBA vers Python, pour utiliser, par exemple, le driver PYODBC sur la base Access. Par ailleurs, sur le plan de l'analyse, les performances de la solution actuelle sont plutôt médiocres. Par exemple, il faut 2 minutes pour extraire l'historique d'occupation des parcelles.

Suite à l'analyse de la solution existante, et un tour d'horizon des possibilités existantes, (Pinet, 2012), nous avons d'abord choisi de transférer les données à partir d'Access dans une base de données reposant sur PostgreSQL avec son extension spatiale PostGIS. A ce stade, l'intérêt d'utiliser une solution complètement libre était lié à la possibilité de modifier les programmes à loisir dans le futur, assurant ainsi la pérennité du logiciel. Quantum GIS[5] (QGIS) a été choisi comme équivalent d'ArcGIS libre et gratuit. Les développements réalisés à partir de QGIS 1.8 et de son API Python ont permis de retrouver l'ancien niveau de fonctionnalités offertes par ArcGIS associé avec Access en trois mois. L'extension permet aux utilisateurs d'obtenir une vue de toutes les parcelles de cultures par année, de modifier des informations sur les contrats en cours, les associations d'agriculteurs, les contrats associés aux parcelles et de mettre à jour les nouvelles cultures ou les géométries de chaque année dans la base

3. http://www.r-project.org/
4. http://www.umass.edu/landeco/research/fragstats/fragstats.html
5. http://www.qgis.org/fr/site/

de données via son interface. En ce qui concerne les capacités d'analyse de l'outil, le niveau de réactivité mesuré est assez élevé grâce aux performances de la base de données PostgreSQL avec son extension PostGIS. L'historique des cultures pour toutes les parcelles est calculé et affiché en 10 secondes, contre 2 minutes dans l'outil précédent dans les même conditions (PC standard de type Core i7 3740QM 2.7GHz avec 8Go RAM).

Si cette solution est satisfaisante sur le plan fonctionnel à court terme, on note cependant que la solution logicielle est très étroitement couplée avec le modèle de données. Si celui-ci évolue, pour intégrer par exemple la base de données "Insect", il faudra refaire l'ensemble des applicatifs qui existent à l'heure actuelle autour de cette base "Insect" pour son usage courant, ce qui implique un surcoût en temps de travail important. De plus, le modèle est actuellement implémenté dans une base de données relationnelle, optimisée pour des requêtes portant sur des agrégats (Robinson *et al.*, 2013). Il apparaît que si une vue de type graphe spatio-temporel est générée, comme cela a déjà été proposé (Mondo *et al.*, 2010), et que les interfaces utilisateurs s'adressent à cette vue plutôt que directement à l'implémentation, le système d'information pourrait gagner en souplesse et en performance.

– Souplesse car on pourrait intégrer au fur et à mesure de nouvelles sources de données (comme la base de données "Insect") et proposer grâce au graphe spatio-temporel différentes connexions aux nouvelles entités du modèle.

– Performance car un graphe est optimisé pour la navigation par connexion, ce qui correspond aux requêtes qui sont formulés sur le système d'information.

3. Mapping des sources d'information à l'aide d'une ontologie spatio-temporelle

Par conséquent, afin de faire évoluer notre SIG-E, nous nous tournons vers les technologies du Web sémantique conçu pour intégrer et raisonner sur des sources de données hétérogènes en constante expansion, capable de modéliser un graphe de relations entre ces données. Nous visons au développement d'une ontologie qui permettra la représentation de diverses entités dynamiques ainsi que l?analyse de leurs relations spatio-temporelles. Concernant l'architecture du système destiné à implémenter ce modèle avec l'information existante, nous discutons des avantages et inconvénients des deux options disponibles.

3.1. Les ontologies dans les travaux antérieurs

Une ontologie est une spécification explicite d'une conceptualisation. Elle aide à structurer la connaissance et améliorer la compréhension des concepts en mettant en évidence comment les entités relient les uns aux autres (Gruber, 1993). En définissant les entités et leurs relations, les ontologies peuvent résoudre le problème d'hétérogénétié présenté. Pour cette raison, nous souhaitons développer notre ontologie en nous appuyant sur celles déjà existantes et utiles pour nous, qui sont, en l'occurence, les ontologies de temps et les ontologies des fluents.

3.1.1. Ontologie de temps

L'ontologie OWL-Time[6] (Hobbs, Pan, 2004) développée au sein du consortium W3C se consacre aux concepts et relations temporelles comme définis dans la théorie d'Allen, et bénéficie d'une spécification précise formalisée en OWL, et est donc certainement appropriée. Cette ontologie de domaine est tout d'abord destinée à décrire le contenu temporel des pages Web et les propriétés temporelles des services Web. Étant recommandée par le W3C pour la modélisation des concepts temporels, cette ontologie fournit un vocabulaire pour exprimer des faits sur les relations topologiques entre les instants et les intervalles. Cependant, une ontologie de temps n'est pas suffisante pour représenter l'évolution d'un objet. Une ontologie de niveau supérieure telle que l'ontologie des fluents, basée sur les ontologies de temps, est strictement nécessaire.

3.1.2. Ontologie des fluents

Les ontologies traditionnelles sont synchroniques, c'est à dire qu'elles se réfèrent à un seul point dans le temps. Afin d?assurer le suivi de l?évolution spatiale et sémantique (diachronie) d?un objet, nous avons besoin d'incorporer la dimension temporelle dans l?ontologie. . En effet, les philosophes ont établi une distinction entre deux paradigmes : l'endurantisme (également appelé tridimensionnel ou 3D) et le perdurantisme (aussi appelé quatre dimensions ou 4D) pour représenter les identités diachroniques. L'endurantisme suppose que les objets (désignés comme "endurants" ou "continuants") ont trois dimensions spatiales et existent en totalité à chaque moment de leur vie. Ainsi, ces objets n'ont normalement pas de dimension temporelle. En revanche, l'approche perdurantiste considère que les objets (les "occurrents" ou "perdurants") ont quatre dimensions (spatiales et temporelles). Ces objets ont des "tranches de temps" (*time slices*) dans leur vie qui composent la dimension temporelle. Cette approche représente donc les différentes propriétés d'une entité dans le temps comme les "fluents" qui ne sont validés que pendant certains intervalles ou à des moments dans le temps. L'approche perdurantiste permet des représentations plus riches de phénomènes du monde réel grâce à sa flexibilité et son expressivité (Al-Debei *et al.*, 2012).

Les deux langages principaux du Web Sémantique, OWL et RDF, ne permettent que des relations binaires entre les individus, sans aucune considération de la relation temporelle entre eux. Néanmoins, plusieurs méthodes ont été proposées afin de surmonter cette limitation, la plus connue est le *4D-Fluents* (Welty, Fikes, 2006). Cette méthode utilise la classe *TimeSlice* qui représente les parties temporelles de l'entité tandis que *TimeInterval* constitue une classe du domaine temporel. L'entité est liée à une instance de la classe *TimeSlice* par la propriété *tsTimeSliceOf* et cette instance est connectée avec une instance de la classe *TimeInterval* par la propriété *tsTimeInterval*.

6. http://www.w3.org/2006/time

Plusieurs approches basées sur le *4D-Fluents* ont été introduites pour représenter la dimension temporelle. TOWL (Frasincar *et al.*, 2010) étend le langage OWL avec une dimension temporelle permettant la représentation du temps, des changements et des transitions. SOWL (Batsakis, Petrakis, 2011a; 2011b) par exemple étend l'ontologie OWL-Time en considérant les relations qualitatives entre les intervalles. Récemment, le modèle *Continuum*, présenté par (Harbelot *et al.*, 2013a; 2013b) permet l'inférence de informations qualitatives à partir d'informations quantitatives en reliant les diverses représentations dynamiques d'une entité.

3.2. Ontologie spatio-temporelle

Nous proposons une ontologie (Figure 1) basée sur l'approche *4D-fluents* qui s'adapte bien aux exigences de notre système, décrite en RDFS. Les principales entités spatio-temporelles dans notre recherche sont les parcelles, les routes, la faune et la flore (les observations d'espèce), en particulier les insectes et les oiseaux. Ces entités ont des "tranches de temps" qui correspondent à leurs caractéristiques, leur occupation spatiale différente à travers leur vie. De cette façon, la rotation des cultures de chaque parcelle ou la modification des limites de ces parcelles peuvent être représentées et analysées.

Dans l'intégration des bases données, nous somme confrontés au problème de l'hétérogénéité de leurs données, en particulier en ce qui concerne la dimension temporelle. En effet, tandis que la rotation des cultures ou les changements de géométrie des parcelles sont archivés avec des intervalles de validité temporels, les observations des chercheurs concernant les insectes et la faune ou la flore (position, comportement) sont datés par des instants. Notre solution étend alors le modèle *4D-Fluents* en généralisant la classe *Interval* à la classe *TemporalEntity* de l'ontologie OWL-Time qui a deux sous-classes *Interval* et *Instant*. Nous proposons aussi la classe *MicroparcelGeometry* correspondant à des micro-parcelles dans notre système comme une sous-classe de la classe *Polygone*, spécialisant *Geometry* (la partie spatiale reste ainsi conforme aux spécifications de l'Open Geospatial Consortium). Par conséquent, nous pouvons exploiter la structure de nos bases de données construites sur le paradigme *Space-Time composite* et bénéficier de son avantage majeur, qui est l'économie d'espace de stockage.

3.2.1. Raisonner sur le temps

Les relations qualitatives dans le domaine temporel sont basées sur des relations binaires et mutuellement exclusives. Les travaux d'Allen (Allen, 1983a) fondent une algèbre temporelle permettant de définir les relations topologiques entre objets datés. Pour deux intervalles temporels définis par leurs dates de début et fin, il existe les 13 relations suivantes : *avant*, *rencontre*, *chevauche*, *débute*, *pendant*, *termine* avec leurs réciproques et *égal*. Ces intervalles peuvent être considérés comme des instances de la classe *ProperInterval* de OWL-Time. Ils sont liés à deux instances de la classe *Instant* par l'attribut *hasBeginning* et *hasEnd* qui déterminent leurs dates de début et fin.

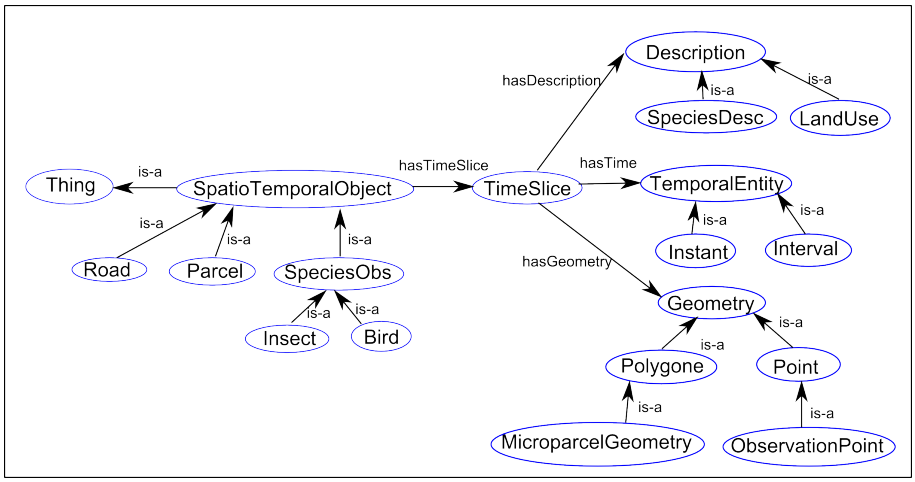

Figure 1. Modèle d'ontologie du système

Les 13 relations d'Allen permettent de répondre à des questions sur la proximité temporelle de deux phénomènes, mais à condition d'employer pour les intervalles la même granularité. Or, il est indispensable pour le croisement de nos sources de données, d'exprimer et de modéliser la relation "à l'intérieur de" entre un instant et un intervalle. Bien que OWL-Time (Frasincar *et al.*, 2009) définisse une relation *inside* entre un instant et un intervalle, nous ne pouvons pas l'utiliser directement. Le langage *Semantic Web Rule Language* (SWRL[7]) est une solution pour ajouter des règles d'inférences générales et il fournit des bibliothèques implémentant des prédicats ou *Built-In* réutilisables directement. Nous l'utilisons donc pour définir les règles d'inférence des relations topologiques dans la dimension temporelle. Ainsi, les relations temporelles qualitatives entre les objets spatiaux dans nos bases de données sont déduites par le moteur d'inférence Pellet[8] grâce à un ensemble de règles exprimées en SWRL. La règle en SWRL pour chercher la relation "à l'intérieur de" entre un instant et un intervalle est exprimée comme suit:
- Instant(?x), ProperInterval(?a), hasBeginning(?a, ?b), hasEnd(?a, ?c), inXSDDateTime(?b, ?d), inXSDDateTime(?c, ?e), inXSDDateTime(?x, ?y), greaterThanOrEqual(?y, ?d), lessThanOrEqual(?y, ?e) -> inside(?x, ?a)

3.3. Architecture du système

Afin de peupler l'ontologie présentée à partir de sources de données existantes, nous nous appuyons sur une technique de traduction définissant une correspondance entre les bases de données et notre ontologie. Dans la littérature, les outils de traduc-

7. http://www.w3.org/Submission/SWRL/
8. http://clarkparsia.com/pellet/

tion peuvent se conformer au language R2RML[9], ou proposer leur propre langage de mapping (Michel *et al.*, 2014). La plateforme D2RQ[10] (Bizer, 2004), qui appartient à la seconde catégorie, est adaptée à nos besoins car ce projet libre très actif autorise la traduction de plusieurs bases de données situées sur différents SGBD. Grâce à D2RQ, l'exploitation des données d'"Assolement" et de "Faunes et Flore" stockées au PostgreSQL ainsi que d'"Insect" enregistrées avec Access est directe et automatique.

Il existe deux options pour l'architecture de système. Dans la première option (Figure 2), les données relationnelles sont transformées en graphes RDF virtuels par D2RQ grâce au fichier de mapping, document décrivant la connexion aux bases de données ainsi que la correspondance entre notre ontologie et chaque schéma de nos bases de données.

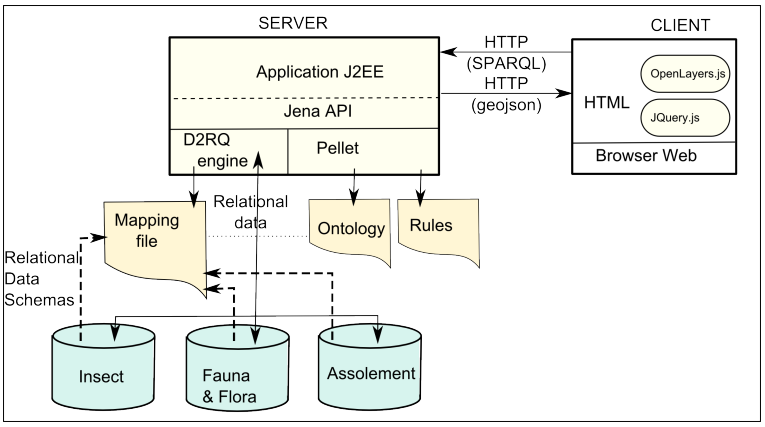

Figure 2. Architecture du système testée.

La seconde option consiste à migrer les données vers des triplestores qui sont les bases de données spécialement conçues et optimisées pour le stockage et la récupération de données RDF. Actuellement, plusieurs triplestores supportent le stockage et les requêtes des données spatiales en utilisant le GeoSPARQL[11] ou le stSPARQL[12], les extensions du langage SPARQL. Les triplestores gérant au mieux les calculs de contraintes spatio-temporelles sont uSeekM[13], Parliament[14] et Strabon[15]. D'autres triplestores comme OpenLink Virtuoso, OWLIM et AllegroGraph ne permettent que la représentation de géométries de type ponctuelles et fournissent quelques fonctions géospatiales (Garbis *et al.*, 2013). Cette architecture offre de bien meilleures perfor-

9. http://www.w3.org/TR/r2rml/
10. http://d2rq.org/
11. portal.opengeospatial.org
12. www.strabon.di.uoa.gr/stSPARQL
13. http://dev.opensahara.com/projects/useekm/
14. http://parliament.semwebcentral.org/
15. http://strabon.di.uoa.gr/

mances que la première et nous évite d'implémenter les règles pour inférer les relations topologiques entre objets spatiaux. Néanmoins, elle nécessite de procéder à la mise à jour régulière de ces triplestores qui ne sont que des copies non synchronisées des données.

Dans un premier temps, nous avons souhaité vérifier l'intérêt de notre proposition à travers la première architecture, plus simple à mettre en oeuvre, incluant une application que nous avons développée pour faciliter les requêtes de l'utilisateur et visualiser les résultats sur une carte.

4. Expérimentation et discussion

Afin de démontrer la pertinence de l'architecture proposée, nous avons développé un outil permettant à réaliser les requêtes SPARQL et visualiser les résultats sur demande. L'utilisateur peut aussi choisir une interface basique au cas où il ne soit pas familiarisé avec ce langage. L'intégration des requêtes et la visualisation repose sur le framework *Jena*[16] et la bibliothèque *OpenLayers*[17] avec le fond cartographique d'*OpenStreetMap*. Les données d'Assolement, de Faune et flore, d'Insect peuvent être interrogées en même temps. Les résultats obtenus sont stockés en plusieurs couches différentes pour faciliter la présentation et l'analyse.

4.1. L'apport du mapping

Le modèle de données sous forme de graphe sous-jacent à RDF facilite l'intégration des bases de données différentes se situant dans un même ou dans différents SGBD. En appliquant cette technique dans notre cas, nous pouvons analyser le croisement entre l'assolement et la biodiversité. Plus concrètement, les experts peuvent facilement analyser les références des animaux par type et forme d'assolement. Par exemple, ils peuvent consulter la corrélation entre les positions des Busards cendrés (*Circus pygargus*) et des parcelles dont la culture est Blé, Luzerne ou Prairie (Figure 3) pour une année donnée.

4.2. Capacités de raisonnement temporel

Grâce aux relations temporelles qualitatives inférées par le moteur d'inférence Pellet, les chercheurs peuvent aussi vérifier la qualité les données stockées. En effet, les règles du domaine ou les expériences d'expertise liées à la rotation des cultures, à l'apparition ou la disparition d'une culture d'une parcelle peuvent être représentées en langage SPARQL, afin de trouver des anomalies dans les données retenues. La Figure 4 montre un exemple de recherche de parcelles ayant une succession Colza-Tournesol qui, selon l'expertise, est très peu probable.

16. http://jena.apache.org/
17. http://openlayers.org/

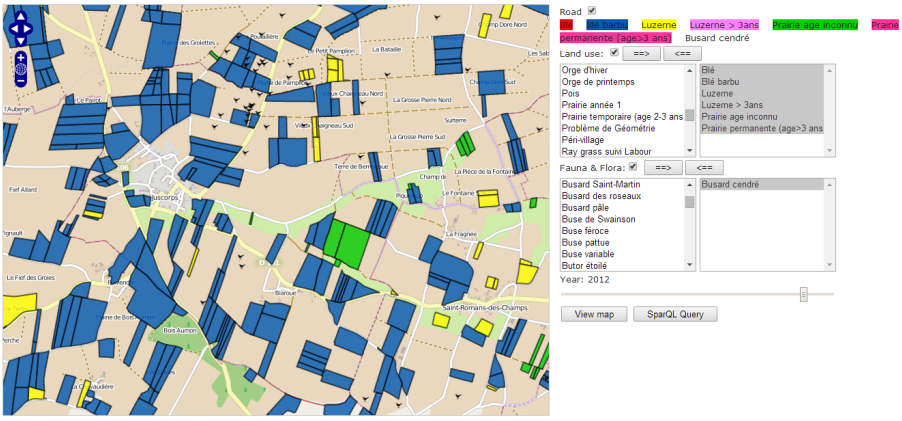

Figure 3. Recherche de corrélations entre les positions des Busards cendrés et la nature des cultures sur les parcelles en 2012. En rouge et bleu les parcelles avec blé, en jaune et magenta les luzernes, en vert, les prairies. La position des nids est figurée par un oiseau.

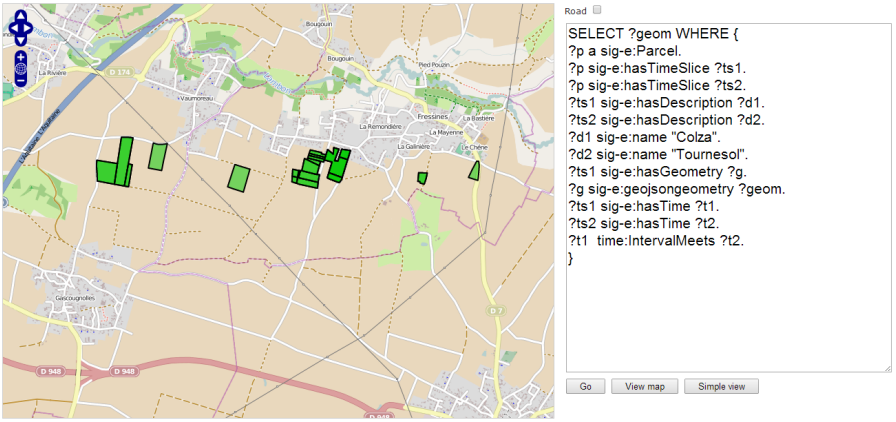

Figure 4. Recherche de la succession improbable Colza-Tournesol : parcelles en vert.

4.3. Performances

Dans l'expérimentation menée sur une machine Core i7 3740QM 2.7GHz avec 8Go RAM, plus de 1.8 million de triplets sont générés grâce à D2RQ. Pour une recherche des parcelles suivant leur culture et une année donnée, le temps de réponse varie entre 2 et 10 secondes côté serveur, ce qui est acceptable. L'affichage de ces résultats côté client sur le fond OpenStreetMap prend 2 secondes en plus (la géométrie des parcelles est transformée en format geojson).

Pour diminuer le temps de réponse, les relations temporelles qualitatives sont stockées dans un fichier puis unies avec les données RDF obtenues par le mapping

lors d'une requête SPARQL. La machine accomplit cette étape en 2 minutes pour l'ensemble des bases. Malgré cela, les requêtes sur les rotations de cultures nécessitant un raisonnement temporel sont résolues en 25 minutes. Cette faiblesse des performances est liée aux fait que les requêtes réécrites ne sont pas optimisées et les clauses de sélection sont calculées au niveau du système RDB2RDF au lieu d'être poussées vers la base de données relationnelle (Gray *et al.*, 2009). Elles pourraient cependant être largement améliorées par la mise en oeuvre d'une migration des données vers un triplestore dédié, tel que Strabon.

5. Conclusion

Les travaux exposés dans cet article s'inscrivent dans le projet interdisciplinaire "GéoConnaissances des milieux naturels" visant à améliorer l'exploitation des informations collectées depuis 1994 par l'observatoire de la "Zone Atelier Plaine & Val de Sèvre". Nous cherchons à proposer une plateforme ouverte et libre pour manipuler et analyser des données environnementales. Nous avons montré que l'intégration de ces bases par la migration dans une unique base de donnée relationnelle et spatiale est une approche coûteuse qui ne peut répondre aux besoins d'évolutivité de ce type de système d'observation.

Nous proposons une architecture permettant l'interrogation de ces bases à travers une modélisation intégrant les composantes spatiales, temporelles et thématique des données. Appliquée à notre cas d'étude, cette approche facilite la mise en relation des cultures et des observations. Ainsi, l'usage du moteur d'inférence de Pellet autorise le calcul des relations temporelles (Allen, 1983b), et nous permet, par exemple, de rechercher efficacement des successions de cultures improbables. Cependant les performances sont encore insuffisantes. Notre approche pour manipuler les relations temporelles et spatiales est similaire à celle utilisée pour des trajectoires d'animaux (Mefteh *et al.*, 2012) et notre objectif est d'améliorer les performances pour l'utilisation des relations temporelles (Wannous *et al.*, 2013b) et spatiales (Wannous *et al.*, 2013a). Dans ce but nous envisageons d'étendre D2RQ vers le support des données spatiales comme proposé par (Valle *et al.*, 2010) ou d'appliquer les règles d'inférence pour la dimention spatiale comme introduites par (Karmacharya *et al.*, 2010). L'usage d'un mécanisme de persistance, un triplestore, adapté à la représentation des données sous la forme de graphe, avec indexation spatio-temporelle est aussi envisagé.

Dans nos perspectives, il sera alors possible d'utiliser ces résultats directement pour l'enrichissement et la qualification des sources de données. En effet, les règles d'expertise vérifiées de façon automatique nous permettraient d'annoter les entrées improbables et peu fiables des sources de données, les écartant ainsi de procédures ultérieures d'analyse. Par ailleurs, si l'une de nos perspectives les plus évidentes concerne l'ouverture de nos données sémantiquement annotées sur le Web, il existe encore à l'heure actuelle des obstacles juridiques qui doivent être levés, car, par exemple,

la base de données Assolement fait l'objet d'une déclaration à la CNIL, puisque les informations concernant les exploitants agricoles sont tout à fait confidentielles.

References

Al-Debei M. M., Asswad M. M. al, Cesare S. de, Lycett M. (2012). Conceptual modelling and the quality of ontologies: Endurantism vs. perdurantism. *CoRR*.

Allen J. F. (1983a). Maintaining knowledge about temporal. *Intervals CACM*, Vol. 26, pp. 198–3.

Allen J. F. (1983b). Maintaining knowledge about temporal intervals. *Commun. ACM*, Vol. 26, No. 11, pp. 832–843.

Batsakis S., Petrakis E. (2011a). Representing temporal knowledge in the semantic web: The extended 4d fluents approach. In I. Hatzilygeroudis, J. Prentzas (Eds.), *Combinations of intelligent methods and applications*, Vol. 8, p. 55-69. Springer Berlin Heidelberg.

Batsakis S., Petrakis E. (2011b). Sowl: A framework for handling spatio-temporal information in owl 2.0. In N. Bassiliades, G. Governatori, A. Paschke (Eds.), *Rule-based reasoning, programming, and applications*, Vol. 6826, p. 242-249. Springer Berlin Heidelberg.

Bizer C. (2004). D2rq - treating non-rdf databases as virtual rdf graphs. In *In proceedings of the 3rd international semantic web conference (iswc2004).*

Frasincar F., Milea V., Kaymak U. (2009). towl: Integrating time in OWL. In R. D. Virgilio, F. Giunchiglia, L. Tanca (Eds.), *Semantic web information management - A model-based perspective*, pp. 225–246. Springer. Retrieved from http://dx.doi.org/10.1007/978-3-642-04329-1_11

Frasincar F., Milea V., Kaymak U. (2010). towl: Integrating time in owl. In R. de Virgilio, F. Giunchiglia, L. Tanca (Eds.), *Semantic web information management*, p. 225-246. Springer Berlin Heidelberg.

Garbis G., Kyzirakos K., Koubarakis M. (2013). Geographica: A benchmark for geospatial rdf stores. *CoRR*, Vol. abs/1305.5653.

Gray A. J., Gray N., Ounis I. (2009). Can rdb2rdf tools feasibily expose large science archives for data integration? In *Proceedings of the 6th european semantic web conference on the semantic web: Research and applications*, pp. 491–505. Berlin, Heidelberg, Springer-Verlag.

Gruber T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, Vol. 5, No. 2, pp. 199 - 220.

Hacid M.-S., Reynaud C. (2004, jun). L'intégration de sources de données. *Revue Information - Interaction & Intelligence (I3) Une Revue en Sciences du Traitement de l'Information*, Vol. 4, No. 2. Retrieved from http://liris.cnrs.fr/publis/?id=1979

Harbelot B., Arenas H., Cruz C. (2013a). Continuum: A spatiotemporal data model to represent and qualify filiation relationships. In *Proceedings of the 4th acm sigspatial international workshop on geostreaming*, pp. 76–85. ACM.

Harbelot B., Arenas H., Cruz C. (2013b, February). The spatio-temporal semantics from a perdurantism perspective. In *In Proceedings of the Fifth International Conference on Advanced Geographic Information Systems, Applications, and Services GEOProcessing.*

Hobbs J. R., Pan F. (2004). An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing*, Vol. 3, pp. 66–85.

Karmacharya A., Cruz C., Boochs F., Marzani F. (2010). Use of geospatial analyses for semantic reasoning. In R. Setchi, I. Jordanov, R. Howlett, L. Jain (Eds.), *Knowledge-based and intelligent information and engineering systems*, Vol. 6276, p. 576-586. Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-15387-7_61

Langran G. E., Chrisman N. R. (1998). A framework for temporal geographic information. *Cartographica: The International Journal for Geographic Information and Geovisualization*, Vol. 25, No. 3, pp. 1-14.

Mefteh W., Bouju A., Malki J. (2012). Une approche ontologique pour la structuration de données spatio-temporelles de trajectoires : Application à l'étude des déplacements de mammifères marins. *Hermes-Lavoisier*, Vol. 22, No. 1, pp. 55–75.

Michel F., Montagnat J., Faron-Zucker F., Catherine. (2014, May). *A survey of RDB to RDF translation approaches and tools*. Technical report. (ISRN I3S/RR 2013-04-FR 24 pages)

Mondo G. D., Stell J. G., Claramunt C., Thibaud R. (2010, jun). A graph model for spatio-temporal evolution. *Journal of Universal Computer Science*, Vol. 16, No. 11, pp. 1452–1477.

Pinet F. (2012). Entity-relationship and object-oriented formalisms for modeling spatial environmental data. *Environmental Modelling & Software*, Vol. 33, No. 0, pp. 80 - 91. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364815212000151

Raffaetà A., Ceccarelli T., Centeno D., Giannotti F., Massolo A., Parent C. *et al. (2008). An application of advanced spatio-temporal formalisms to behavioural ecology.* GeoInformatica*, Vol. 12, No. 1, pp. 37-72. Retrieved from http://dx.doi.org/10.1007/s10707-006-0016-6*

Robinson I., Webber J., Eifrem E. (2013). Graph databases*. O'Reilly Media, Incorporated.*

Valle E. D., Qasim H. M., Celino I. (2010). Towards Treating GIS as Virtual RDF Graphs. In Proceedings of 1st international workshop on pervasive web mapping, geoprocessing and services (webmgs 2010).

Vargas-Solar G., Doucet A. (2002). Médiation de données: solutions et problèmes ouverts. Actes des deuxièmes assises nationales du GdRI3.

Wannous R., Malki J., Bouju A., Vincent C. (2013a). Modelling mobile object activities based on trajectory ontology rules considering spatial relationship rules. In A. Amine, A. M. Otmane, L. Bellatreche (Eds.), Modeling approaches and algorithms for advanced computer applications*, Vol. 488, p. 249-258. Springer International Publishing.*

Wannous R., Malki J., Bouju A., Vincent C. (2013b). Time integration in semantic trajectories using an ontological modelling approach. In M. Pechenizkiy, M. Wojciechowski (Eds.), New trends in databases and information systems*, Vol. 185, p. 187-198. Springer Berlin Heidelberg.*

Welty C., Fikes R. (2006). A reusable ontology for fluents in owl. In Proceedings of the 2006 conference on formal ontology in information systems*, pp. 226–236. IOS Press.*