Atelier IN-OVIVE - 4ème édition

"INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement"

IC 2016

L'objectif de l'atelier IN-OVIVE adossé à la conférence IC est de dresser un panorama des recherches et expérimentations francophones traitant de l'intégration de sources/masses de données hétérogènes notamment à l'aide d'ontologies, dans le domaine des sciences du vivant et de l'environnement. Les quatre principaux thèmes de l'atelier IN-OVIVE 2016 sont : modélisation et représentation des connaissances, ontologie et données liées, évaluation et qualification des sources d'informations et des données extraites, raisonnement et connaissances imparfaites.

Comité d'initiative

- Patrice Buche, Ingénieur de Recherche HDR UMR INRA IATE
- Stéphane Dervaux, Ingénieur d'Etude UMR INRA MIA-Paris
- Juliette Dibie, Professeur AgroParisTech & UMR INRA MIA-Paris
- Liliana Ibanescu, Maitre de conférences AgroParisTech & UMR INRA MIA-Paris
- Claire Nédellec, Directrice de Recherche UMR INRA MaIAGE
- Pascal Neveu, Ingénieur de Recherche UMR INRA MISTEA

Comité de programme

- Robert Bossy, Ingénieur de Recherche UMR INRA MaIAGE
- Julie Bourbeillon, Maître de conférences AgroCampus Ouest
- Sylvie Despres, Professeur Université Paris 13 & unité INSERM LIMICS
- Brigitte Grau, Professeur ENSIIE & LIMSI
- Ollivier Haemmerlé, Professeur Université Le Mirail Toulouse III & UMR CNRS IRIT
- Mouna Kamel, Maître de Conférences Université de Perpignan & UMR CNRS IRIT
- Nathalie Pernelle, Maître de Conférences Université Paris-Sud 11 & LRI IASI
- Mathieu Roche, Maître de Conférences, HDR Université Montpellier II & LIRMM
- Catherine Roussey, Chargée de Recherche, IRSTEA unité TSCF
- Fatiha Saïs, Maître de Conférences Université Paris-Sud 11 & LRI IASI
- Maguelonne Teisseire, Professeur TETIS IRSTEA
- Konstantin Todorov, MCF Université Montpellier 2 & LIRMM
- Haïfa Zargayouna, Maître de Conférences Université Paris 13 Sorbonne Paris Cité & UMR CNRS LIPN

Table des matières

Représentation et structuration efficiente de la connaissance de la bio-raffinerie lignocellulosique du bois

BAUDRIT Cédric, FERNANDEZ Christophe, NDIAYE Amadou

I2M, Institut de Mécanique et d'Ingénierie – INRA Bordeaux
cedric.baudrit@bordeaux.inra.fr, christophe.fernandez@bordeaux.inra.fr, amadou.ndiaye@bordeaux.inra.fr

1.1 Introduction

Il existe une ressource durable en biomasse lignocellulosique forestière provenant de systèmes « socio-écologiques » (couplant production primaire forestière et transformation industrielle du bois) et gérée par de multiples agents à l'échelle locale (territoire, région,…) pour des marchés globaux. L'enjeu de la bio-raffinerie, aujourd'hui, est d'optimiser la valorisation de toutes les composantes et propriétés intrinsèques du bois (cellulose, lignine, hémicellulose, extractibles, minéraux, …) pour les transformer en énergie et une gamme de co-produits à hautes valeurs ajoutées (synthons, polymères, biocarburant …) [1]. Il est capital de comprendre et maîtriser la déconstruction moléculaire du bois afin de rendre accessible la cellulose, les hémicelluloses et les lignines pour la synthèse de nouveaux produits via des bioprocédés socio-économiquement et écologiquement viables. Une des difficultés que rencontre la filière de la biomasse réside dans le fait que la conception intégrée de produits bio-sourcés diversifiées et durables nécessite une coopération entre de nombreuses disciplines (biologie, physique, économie, ingénierie, chimie, informatique, écologie …) et l'assemblage de différentes expertises qui demeurent encore peu structurée. La fabrication d'un produit ou la conduite de projet industriel ne se limite plus à sa dimension technique mais doit intégrer, dans sa conception, la dimension de durabilité de l'échelle locale à l'échelle mondiale. Les innovations, modifications et le développement de nouveaux projets seront indéniablement facilités pour ceux qui sauront avoir une vue d'ensemble sur leur système. La mutualisation des connaissances et des savoirs par et entre champs disciplinaires, générera des transferts de méthodes entre disciplines et fera indéniablement émerger de l'innovation. Pour ce faire, il est capital de proposer des approches conceptuelles qui permettent de représenter l'ensemble des acteurs et des éléments fonctionnels en interaction à différents niveaux d'échelles. Le comportement collectif des éléments fonctionnels et leurs interactions engendrent des structures organisées qui influencent en retour des comportements individuels. L'objectif de ce travail est de créer une interface capable de fédérer dans un cadre formel et unificateur la connaissance des scientifiques et le savoir-faire des industriels relatifs à la bio-raffinerie de la biomasse lignocellulosique du bois.

1.2 Modèle de la bioraffinerie lignocellulosique bois

Une modélisation graphique du modèle de bioraffinerie lignocellulosique bois est proposée (Fig. 1). Chaque entité du modèle, à l'exemple de *Produits ciblés* (voir Fig. 1), peut être reliée par un hyperlien à une carte conceptuelle (appelée aussi graphe sémantique) [2] et/ou une carte de processus [3] permettant de décrire de manière formelle les procédés. Chaque carte se compose de concepts interconnectés par des relations ontologiques (taxonomique, méréologique et du domaine) et des hyperliens permettant de descendre dans la granularité de la description des concepts et procédés déclinés. Le modèle a été implémenté avec des outils Web 2.0 [4] permettant de parcourir l'hypergraphe résultant jusqu'à un niveau de détail souhaité [5].

1.3 Conclusion et perspectives

Le modèle permet l'assemblage de connaissances hétérogènes multi-sources et multi-échelles dans un cadre formel homogène. Etant muni d'une structure d'algèbre, l'exploitation du graphe permettra, au delà de son aspect descriptif, de faire des choix parmi l'ensemble des résultats possibles suite à

l'interrogation des modèles sous-jacents dans le but de faire émerger un potentiel d'éco-innovation (stratégies innovantes de fabrication, conduites de procédés ou de lever des verrous technico-scientifiques).

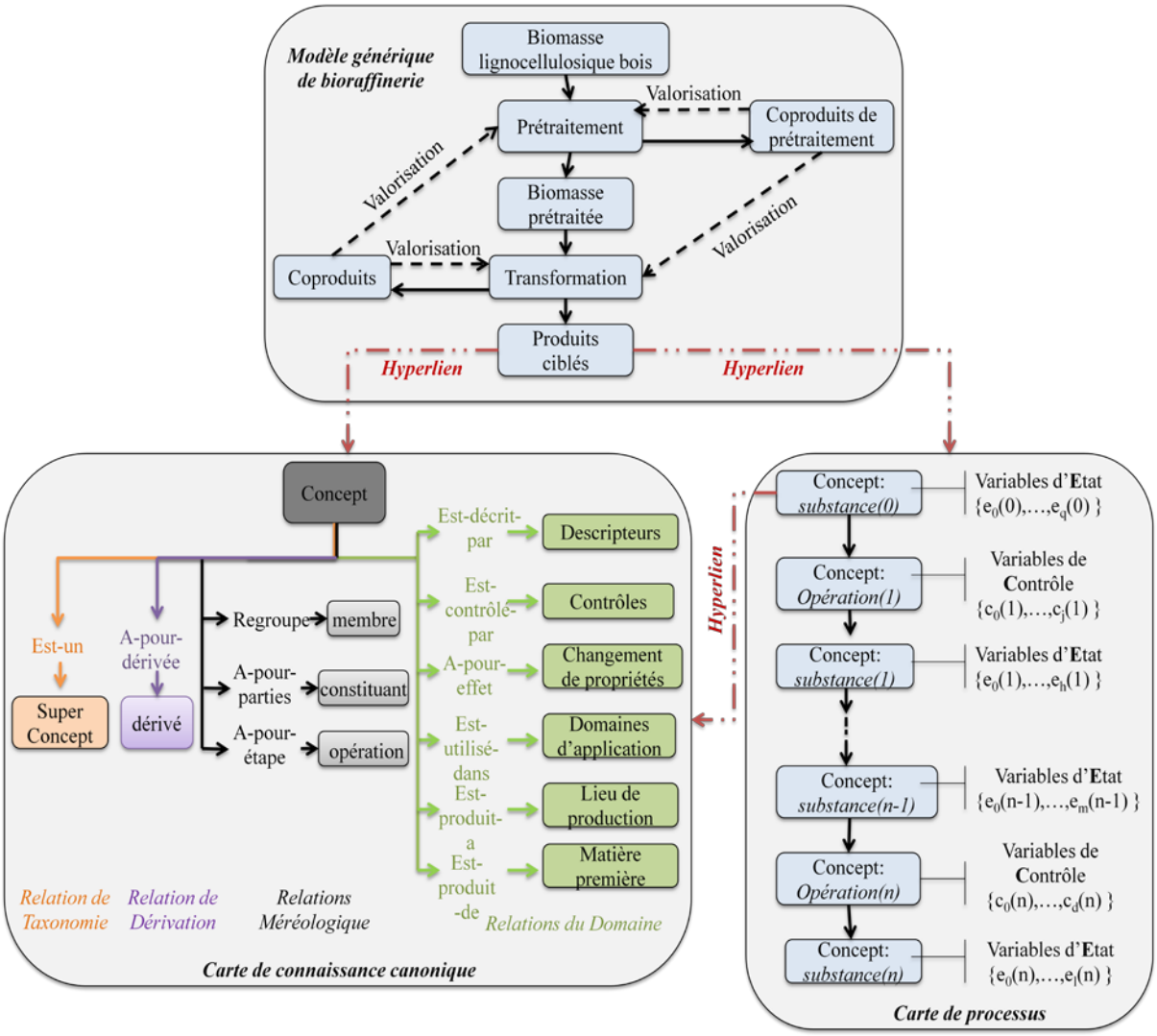


Figure 1 : Modèle de représentation de la bioraffinerie lignocellulosique du bois.

Références

[1] Gabenisch, A., Maës, J., Mandret, N. (2012) *Marché actuel des nouveaux produits issus du bois et évolutions à échéance 2020*. MAAPRAT- DGPAAT / MINEFI – DGCIS / PIPAME / Cabinet Alcimed.

[2] Milton, N.R. (2007) *Knowledge acquisition in practice: A step-by-step Guide*. Springer.

[3] Ndiaye, A., Della Valle, G., Roussel, P. (2009) Qualitative modelling of a multi-step process: The case of French breadmaking. *Expert Systems with Applications*, 36(2): 1020-1038.

[4] Ndiaye, A., Fernandez, C. (2010). MakeBook: Make a Book of Knowledge [Application]. http://prodinra.inra.fr/record/329581.

[5] Baudrit, C. *Califorest*. http://147.210.201.248/califorest/ (login:califorest, password:califorest@2015)

Modélisation et analyse de données environnementales à travers une ontologie spatio-temporelle

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

[1] L3I, Laboratoire Informatique Image et Interaction (L3I), Université de la Rochelle
ba-huy.tran@univ-lr.fr, alain.bouju@univ-lr.fr

[2] LIENSs, LIttoral ENvironnement et Sociétés, U.M.R. CNRS 7266
christine.plumejeaud-perreau@univ-lr.fr

[3] CEBC Centre d'Etudes Biologiques de Chizé, U.M.R CNRS 7372 et Université de la Rochelle
breta@cebc.cnrs.fr

Résumé : Cet article présente les travaux pour le projet interdisciplinaire GEMINAT (GéoConnaisances des milieux naturels) qui a pour but d'intégrer et d'exploiter des données environnementales hétérogènes par l'application du web sémantique. À partir d'un cas d'étude mené sur un observatoire environnemental basé à Chizé, nous résumons les besoins d'analyse spatio-temporelle des experts biologistes et écologues envers les bases de données de l'assolement et de la biodiversité. Nous montrons comment la mise en œuvre d'un framework avec une ontologie spatio-temporelle jouant le rôle d'un médiateur sémantique 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.
Mots-clés : Ontologies, Environnement, Spatio-temporel, Intégration de données.

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 tous types (carabes, libellules, etc.), des plantes, ou des oiseaux ne requièrent ni les mêmes 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.

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

A cet effet, les utilisateurs envisagent la migration des données et schémas dans un système d'information centralisé. Pour cela, il faut mettre en œuvre une médiation entre ces sources de données. En terme de médiation, une des solutions les plus avancées et prometteuse repose sur l'intégration par les techniques du Web sémantique.

La première section expose en détail notre cas d'étude avec les bases de données hétérogènes disponibles et les besoins d'analyse spatio-temporelle envers ces sources de données. La seconde section propose un framework sémantique avec une ontologie spatio-temporelle en tant que médiateur sémantique pour l'intégration et l'exploitation des données hétérogènes. La troisième section démontre la faisabilité et les possibilités qu'offre cette nouvelle approche appliquée à notre cas d'étude. La dernière expose les perspectives de cette recherche.

1.1 Les bases de données de la Zone Atelier Plaine & Val de Sèvre

L'observatoire de la "Zone Atelier Plaine & Val de Sèvre" (Figure 1a) 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 rechercher 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. Depuis plus de vingt ans, plusieurs bases de données ont été recueillies par l'équipe AGRIPOP (CNRS Chizé) (Figure 1b). Ces données peuvent être catégorisées en deux groupes : la base d'assolement et les bases de biodiversité (plus de détails peuvent être trouvés sur le site de l'atelier [2]).

La base *Assolement*

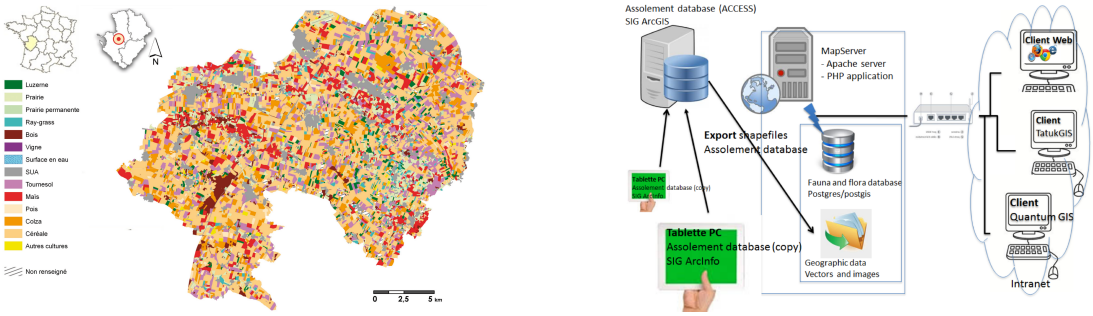
L'organisation spatiale du paysage évolue dans le temps parce que les agriculteurs modifient l'assolement de leurs parcelles chaque année, mais également recomposent parfois les parcelles entre elles (par des scissions, fusions ou des redistributions), changeant ainsi les formes des parcelles.

Depuis 1994, les occupations du sol sont donc relevées annuellement sur le terrain et numérisées sur les 19 000 parcelles agricoles. Ces données sont centralisées dans une base de données nommée "Assolement".

Une parcelle agricole est définie comme une unité de gestion, un polygone entouré par des entités ayant différentes cultures dans les années successives (généralement quatre). Une parcelle est délimitée par des limites physiques telles qu'une route, une rivière, un chemin de champ ou une limite d'un seul champ. Elle ne contient qu'un seul type de culture. Elle diffère de la parcelle cadastrale, mais également des blocs stockés dans le RPG (Registre Parcellaire Graphique) qui sont mis à jour tous les deux ans et distribués par IGN, disponibles ici [3]. Le RPG contient des informations pertinentes pour le suivi des propriétaires des parcelles, mais les limites des blocs ne correspondent pas aux parcelles

2. http ://za-geminat.cnrs.fr/

3. http ://www.geoportail.gouv.fr/donnee/251/registre-parcellaire-graphique-rpg-2012



(a) La zone Atelier Plaine et Val de Sèvre

(b) Architecture du système de collection des données

FIGURE 1: Observatoire de la Zone Atelier Plaine & Val de Sèvre

observées sur le terrain. En fait, les agriculteurs souvent divisent leur bloc en plusieurs parcelles, mais seulement la surface de chaque bloc est déclarée avec sa nature de la culture en pourcentage.

Le modèle de données de la base 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 *microparcelles*, 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 microparcelles 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 le nouveau type d'utilisation du sol de chaque parcelle et éventuellement des changements de forme pour chaque parcelle.

Les bases *Biodiversité*

— 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 qui est structurée suivant un schéma relationnel spatial, implémenté dans PostgreSQL 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.).

— Il existe également les données des micromammifères. Il s'agit d'un ensemble de 10.000 observations de trois espèces de micro-mammifères qui sont de bons indicateurs de pratiques agricoles : Campagnol des champs, Mulot sylvestre et Musaraigne musette. Dans chaque observation, les taux de capture et les caractéristiques individuelles sont enregistrés.

— 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 auxiliaires des champs et très sensibles à la qualité des milieux, qui bénéficient d'un suivi depuis 9 ans dans une base Access.

1.2 Expression des besoins autour de l'exploitation des données hétérogènes

Avec les données déjà disponibles, un nombre conséquent d'analyses peut-être mené :

1. L'analyse peut être utilisée en premier lieu pour vérifier l'ensemble de données collecté. Lors de la rotation des cultures, les experts peuvent décrire un certain nombre de règles de succession afin d'éliminer ou de corriger les valeurs douteuses. Par exemple, la succession des cultures peu probable comme "Tournesol- Tournesol" ou "Tournesol-Colza", ainsi que la disparition de bois dans la zone de l'atelier peut être détectée et examinée. Principalement, ce type d'analyse a besoin du raisonnement sur les relations temporelles entre les durées relevées de l'assolement.
2. D'autre part, les événements territoriaux, tels que la fusion, l'intégration, la scission, l'extraction, la réaffectation et de rectification (Plumejeaud *et al.*, 2011) sont souhaités être détectés. L'analyse de ces événements permet de découvrir la corrélation entre la prise de l'utilisation des terres et la fragmentation des terres ou de l'agrégation dans la pratique agricole. Ces événements peuvent être détectés à travers des requêtes avec le raisonnement spatio-temporel.
3. Enfin, les experts veulent aussi rechercher la corrélation entre les observations des espèces et les utilisations des terres des parcelles. Ils pourraient concerner les préférences des animaux par le type et la forme de la rotation des cultures. Ils peuvent également vérifier l'apparition d'espèces en fonction de la chaîne alimentaire et de la saison. Les requêtes croisées de différentes bases de données avec le raisonnement sur les relations spatio-temporelles sont nécessaires pour sélectionner les observations qui se reproduisent à l'intérieur des durées relevées de l'assolement.

2 Une ontologie spatio-temporelle pour l'exploitation des données environnementales

Dans le but d'intégrer et d'exploiter l'ensemble de données présentées, nous introduisons une approche ontologie spatio-temporelle globale qui agit en tant que médiateur sémantique. L'ontologie est formée sur la base de l'ontologie du temps et de l'ontologie de l'espace. Un mécanisme de raisonnement sur les relations spatio-temporelles entre entités est également proposé.

2.1 Ontologie de temps et représentation du temps dans le web sémantique

L'ontologie OWL-Time [4] (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

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

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. Ces ontologies 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 (relation diachronique) 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 et le perdurantisme pour représenter les identités diachroniques. L'endurantisme suppose que les objets 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 ont quatre dimensions (spatiales et temporelles). Ces objets ont des *tranches de temps* 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. Selon les études de comparaison de (Al-Debei *et al.*, 2012), l'approche perdurantiste permet des représentations plus riches de phénomènes du monde réel grâce à sa flexibilité et son expressivité.

Les deux langages principaux du Web Sémantique, OWL et RDFS, ne permettent que des relations binaires entre les individus, sans aucune considération de la relation temporelle entre eux. Plusieurs méthodes ont été proposées afin de surmonter cette limitation, la plus connue est le modèle perdurantist 4D-Fluents (Welty & Fikes, 2006). Le modèle considère que l'existence d'une entité peut être exprimée avec plusieurs représentations, chacune correspond à un intervalle de temps défini, appelée un *tranche de temps* (*time slice*). Lorsque la propriété d'un objet change, une nouvelle tranche de temps est établie, tenant la nouvelle propriété de l'objet. De cette manière, les changements se produisent sur les propriétés de la partie temporelle de l'ontologie en gardant les entités de la partie statique inchangée. Cette méthode introduit 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*.

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, 2011; Batsakis & Antoniou, 2014) par exemple étend l'ontologie OWL-Time en considérant les relations qualitatives entre les intervalles. Récemment, le modèle *Continuum*, présenté par (Harbelot *et al.*, 2013, 2014) permet l'inférence de l'information qualitative à partir d'informations quantitatives en reliant les diverses représentations dynamiques d'une entité. Le modèle représente des entités dynamiques comme constituées par des tranches de temps avec sémantique, spatiale, composantes temporelles et d'identité. Donc, il est capable de relier les diverses représentations d'une entité et permet l'inférence de l'information qualitative. Les résultats d'inférence sont ensuite ajoutés à l'ontologie afin d'améliorer les connaissances sur le phénomène. Le modèle a été appliqué avec succès dans les études sur l'évolution urbaine (Harbelot *et al.*, 2013) ou d'un processus de décolonisation (Harbelot *et al.*, 2014).

Possédant la capacité de présentation du monde réel des objets dynamiques, le modèle de

4D-fluent avec l'ontologie OWL-Time sont choisis pour le développement de notre ontologie spatio-temporelle qui sera présentée plus tard.

2.2 Ontologie de l'espace

Les entités spatiales sont représentées par des points, des lignes (polygonale, lignes) qui renferment les objets ou les régions et leurs relations. De nombreuses ontologies spatiales ont été introduites pour différentes applications. Selon l'étude de (Ressler *et al.*, 2010), parmi 45 ontologies géospatiales et temporelles, sept ontologies spatio-temporelles sont recommandés pour la réutilisation. Les deux versions de GeoRSS [5] : GeoRSS Simple et GeoRSS GML sont les plus utilisées. GeoRSS a été conçu comme un standard destiné à inclure les coordonnées géographiques dans un flux RSS pour les applications géographiques comme la cartographie interactive. Dans le standard GeoRSS, les contenus sont des points géographiques, des lignes, des zones d'intérêt et leurs descriptions. Ces emplacements peuvent être codés soit dans une chaîne littérale de latitude et de longitude, appelé GeoRSS Simple, ou avec une représentation plus robuste en utilisant GML, appelé GeoRSS GML, qui est formellement défini comme un profile d'application de GML et prend en charge une plus grande gamme de fonctionnalités, notamment les systèmes de référence autres que la latitude et la longitude du WGS84. Le modèle GeoRSS simple est réutilisé dans le développement de notre ontologie en raison de sa simplicité, ses trois concepts principaux : le point, la ligne et le polygone sont suffisants pour représenter la dimension spatiale de nos entités spatio-temporelles.

2.3 Raisonnement spatio-temporel

Les objets perdurants évoluent dans le temps. En conséquence, les objets spatio-temporels peuvent changer leur position ou leurs occupations au cours de leur vie. Le raisonnement spatio-temporel est utilisé pour détecter les relations spatio-temporelles entre ces entités. Ce dernier est réalisé par la combinaison dynamique du mécanisme de raisonnement spatial et temporel qui sont présentés ci-après.

Les relations qualitatives dans le domaine temporel sont basées sur des relations binaires et mutuellement exclusives. Les travaux d'Allen (Allen, 1983) 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 : *before, meets, overlaps, during, starts, finishes* et leur réciproque, respectivement *after, met-by, overlapped-by, contains, started by, finished-by*, et *equals*. 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.

Les 13 relations d'Allen permettent de répondre à des questions sur la proximité temporelle de deux phénomènes, à condition d'employer pour les intervalles la même granularité. En sus, nous devons donc exprimer et modéliser la relation "à l'intérieur de" entre un instant et un intervalle qui est indispensable pour le croisement de nos deux bases de données. Par ailleurs, le langage OWL lui-même n'a pas d'opérateurs temporels pour manipuler des valeurs de temps. Le langage Semantic Web Rule Language (SWRL) est une solution pour ajouter des règles

5. http ://www.georss.org/

d'inférence générales. Il fournit des prédicats facilitant surtout la comparaison des valeurs temporelles. Ce dernier peut être utilisé par un moteur d'inférence, comme par exemple Pellet [6] pour déduire des relations temporelles entre les entités. Dans notre cas, le même résultat est obtenu en représentant ces règles par des requêtes SPARQL CONSTRUCT ou SPARQL UPDATE [7]. L'utilisation du SPARQL comme langage de règle est proposée dans SPIN [8] (Knublauch *et al.*, 2011) ou Corese/KGRAM (Corby *et al.*, 2004). Par exemple, la déduction de la relation *à l'intérieur de* entre un instant et un intervalle de temps peut se représenter par la requête (Code 1) ou par une règle SWRL comme le Code 2.

La dimension spatiale des objets dans les bases de données environnementales en général, et dans nos bases de données en particulier est représentée par des points et des polygones qui sont définis par un ensemble des points. Afin de découvrir leurs relations spatiales, les relations qualitatives doivent être déduites de ces informations quantitatives. Dans la littérature, l'analyse topologique entre les objets spatiaux est souvent réalisée par le modèle 9IM (Egenhofer & Herring, 1991) ou le Modèle RCC8 (Randell *et al.*, 1992). Dans les deux cas, on obtient un ensemble équivalent de huit relations topologiques disjointes qui sont mutuellement exhaustives : *equals, disjoint, intersects, touches, within, contains* et *overlaps*. Malheureusement, ces relations ne peuvent être déduites par les règles SWRL simples. Plusieurs études (Karmacharya *et al.*, 2010; Vandecasteele *et al.*, 2012) ont introduit les SWRL built-ins pour le traitement spatial et la représentation des relations spatiales, mais il existe encore des limitations en ce qui concerne principalement la performance et la capacité de réutilisation.

Par conséquent, le raisonnement sur l'information spatiale complexe est réalisé par un triplestore spatial. Ainsi, le raisonnement spatio-temporel est effectué en combinant les relations temporelles déjà déduites et insérées à la base de connaissances et les fonctions spatiales du triplestore.

Code 1: Infercence de la relation *à l'intérieur de* entre un instant et un intervalle de temps par requête SPARQL Update

```
INSERT {?x time:inside ?a.}
WHERE
{
 ?x rdf:type time:Instant.
 ?x time:inXSDDateTime ?dt.
 ?a rdf:type time:Interval.
 ?a time:hasBeginning ?be.
 ?a time:hasEnd ?end.
 ?be time:inXSDDateTime ?dt1.
 ?end time:inXSDDateTime ?dt2.
FILTER(?dt>=?dt1 && ?dt<=?dt2)
 }
```

Code 2: Infercence de la relation *à l'intérieur de* entre un instant et un intervalle de temps par règle SWRL

```
Instant(?x), ProperInterval(?a),
    hasBeginning(?a,?b), hasEnd(?a,?c
    ), inXSDDateTime(?b,?d),
    inXSDDateTime(?c,?e),
    inXSDDateTime(?x,?y),
    lessThanOrEqual(?y,?e),
    greaterThanOrEqual(?y,?d) ->
    inside(?x,?a)
```

6. http ://clarkparsia.com/pellet/
7. https ://www.w3.org/TR/sparql11-update/
8. http ://spinrdf.org/

2.4 Une ontologie spatio-temporelle pour l'environnement

Nous proposons une ontologie basée sur l'approche 4 D-fluents qui joue le rôle d'un médiateur sémantique pour l'intégration et l'exploitation des données environnementales présentées ci-dessous.

Objet spatio-temporel

Les objets spatio-temporels (*sige :STObj*) principaux dans notre recherche sont les parcelles, les nids et les individus (appelés ici les individus spatio-temporels) appartenant aux différents types d'insectes d'oiseaux ou de micromammifères. Ils peuvent être connus avec un numéro suivi ou inconnus.

Élément spatio-temporel

Les objets spatio-temporels ont un ou plusieurs éléments spatio-temporels (*sige :STElement*) qui correspondent à leurs différentes caractéristiques et occupations spatiales à travers leur vie. La classe *sige :STElement* a deux sous-classes : *sige :Obsv* pour les observations des individus et *sige :TimeSlice* pour les relevés des parcelles. Chaque élément spatio-temporel a trois composants : sémantique (*sig :Description*), temporel (*time :TemporalEntity*) et spatial (*georss :_geometry*). De cette façon, la rotation de cultures de chaque parcelle, les événements territoriaux ou les différentes observations d'une espèce donnée peuvent être représentés et analysés.

Composant temporel

Tandis que les informations des parcelles (leur culture, leur géométrie) sont relevées par des intervalles de validité temporelles, les observations d'espèces sont enregistrées de manière aléatoire 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*.

Composant spatial

Dans ce composant, nous proposons deux classes *sig :MPGeom* et *sig :ParcelGeom* qui correspondent respectivement à des micro-parcelles et parcelles comme une sous-classe de la classe *georss :Polygon*, spécialisant *georss :_geometry*. Par conséquent, nous pouvons conserver la structure initiale de nos bases de données construites sur le paradigme *Space-Time composite* et améliorer la performance du système grâce à la géométrie précalculée des parcelles.

Composant sémantique

Ce composant vise à décrire les observations des individus ou les relevés des parcelles. Il peut être le contexte de ces observations ou la culture appliquée à ces parcelles.

2.5 Vers un système sémantique pour l'intégration et exploitation des données environnementales

Puisque l'application de règles d'inférence sur des sources de données distribués peut s'avérer extrêmement coûteuse (Seye *et al.*, 2014) et qu'il n'existe pas un bon mécanisme pour raisonner sur les relations spatiales qualitatives, la centralisation de différents sources de données

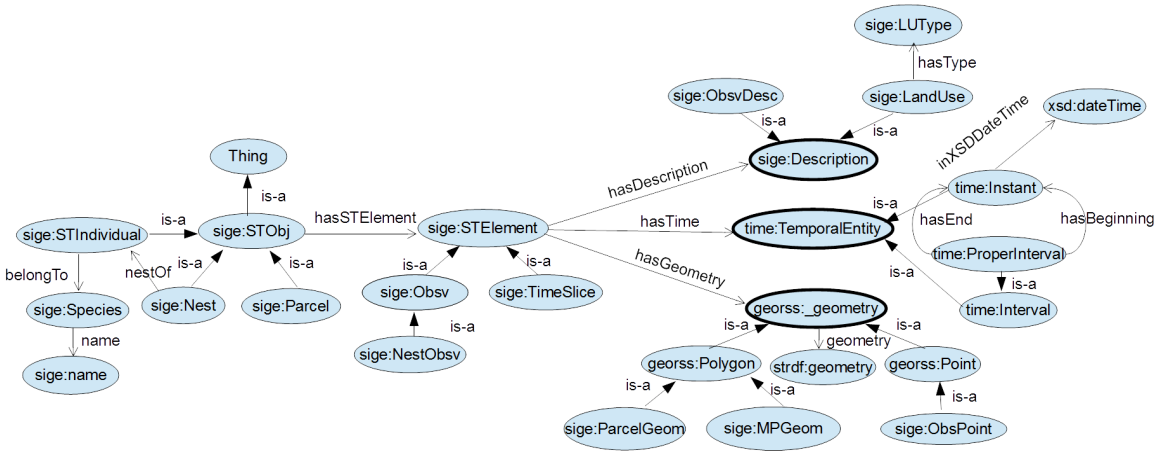


FIGURE 2: Une ontologie spatio-temporelle pour l'environnement.

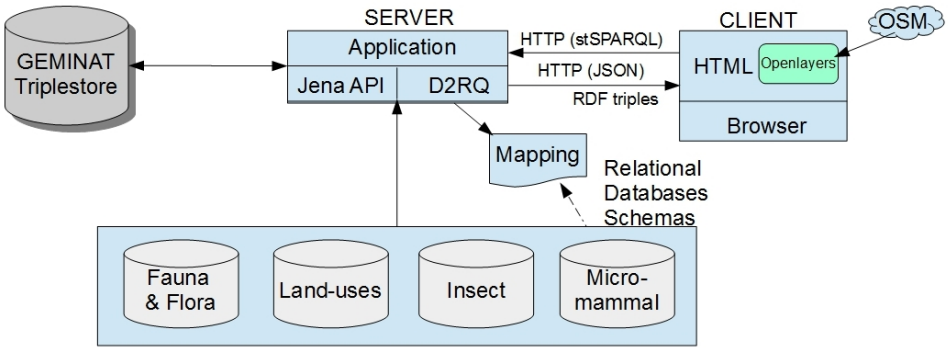


FIGURE 3: Un framework pour l'intégration et l'exploitation des données environnementales

à un triplestore spatial est appliquée. Un framework (Figure 3) est développé dans lequel un serveur web est hébergé pour recevoir les requêtes stSPARQL de l'utilisateur. Le framework se compose de quatre parties : la translation de données, le chargement du triplestore, l'enrichissement spatio-temporel et l'analyse sémantique.

1. **Translation de données** : Afin de peupler l'ontologie avec des sources de données existantes, nous nous appuyons sur la technique de translation qui définit une correspondance entre ces sources de données et l'ontologie. L'outil D2RQ [9] (Bizer, 2004) est choisi en raison de son support de différents SGBD. Celui-ci transforme les données relationnelles en graphe RDF virtuelle à travers un fichier de mapping qui décrit comment se connecter à ces bases de données et mettre en correspondance l'ontologie au schéma de ces sources

9. http ://d2rq.org/

de données. Ce graphe RDF est alors géré par le framework *Jena* [10].

2. **Chargement du triplestore** : Les données transformées en RDF sont insérées dans la base de connaissance *GEMINAT* qui est gérée par le triplestore Strabon. Le chargement se réalise grâce à l'interface web, ou aux requêtes SPARQL ou à la fonction du triplestore.
3. **Enrichissement spatio-temporel** : Les relations spatio-temporelles sont inférées par des requêtes SPARQL 1.1 en utilisant des fonctiones spatiales incorporées dans le triplestore. Ces nouvelles relations sont insérées dans la base de connaissance pour l'enrichissement.
4. **Analyse sémantique** : Le résultat des requêtes SPARQL retourné par Strabon est traité et ensuite préparé par Jena pour la visualisation. Ce dernier est visualisé à travers la bibliothèque *OpenLayers* [11] avec la base de données géographiques d'*OpenStreetMap* [12]. Les résultats sont stockés dans plusieurs couches différentes afin de faciliter la présentation et l'analyse.

3 Exploitation des données environnementales hétérogènes

Le framework proposé ainsi que l'utilisation d'une ontologie spatio-temporelle en tant que médiateur sémantique peut remplir les trois principaux besoins de l'analyse spatio-temporelle. En effet, le modèle de données sous forme de graphe RDF sous-jacent facilite l'intégration des différentes bases de données. En outre, grâce au triplestore Strabon, les relations spatio-temporelles entre les objets peuvent être déduites. De nouvelles déclarations peuvent se déduire de la base de connaissances à travers des règles spatio-temporelles et des règles métiers représentées sous forme de requêtes SPARQL Update.

— Pour analyser les corrélations entre la rotation des cultures et la biodiversité, les experts peuvent visualiser la corrélation des espèces par type et la forme de la rotation des cultures. Par exemple, ils peuvent consulter la corrélation entre les positions des Busards cendrés (Circus pygargus) et la nature des cultures des parcelles pour une année donnée (Figure 4a). Dans cette analyse, la relation *inside* entre l'instant de l'observation et la durée de déclaration de l'assolement ; et la relation *within* entre le point d'observation et le polygone de la parcelle enregistrée sont utilisées (Code 3).

— Grâce aux relations temporelles qualitatives inférées, les chercheurs peuvent également vérifier la qualité de leurs données enregistrées. En effet, les règles de domaine ou des connaissances d'experts sur la rotation des cultures, l'apparence ou la disparition de certaines plantes cultivées peuvent être représentées par des requêtes SPARQL pour détecter des anomalies dans les données recueillies. Dans cette étude, la relation temporelle *intervallMeets* entre deux intervalles de relevée de l'assolement de la même parcelle est utilisée.

10. http ://jena.apache.org/
11. http ://openlayers.org/
12. www.openstreetmap.org

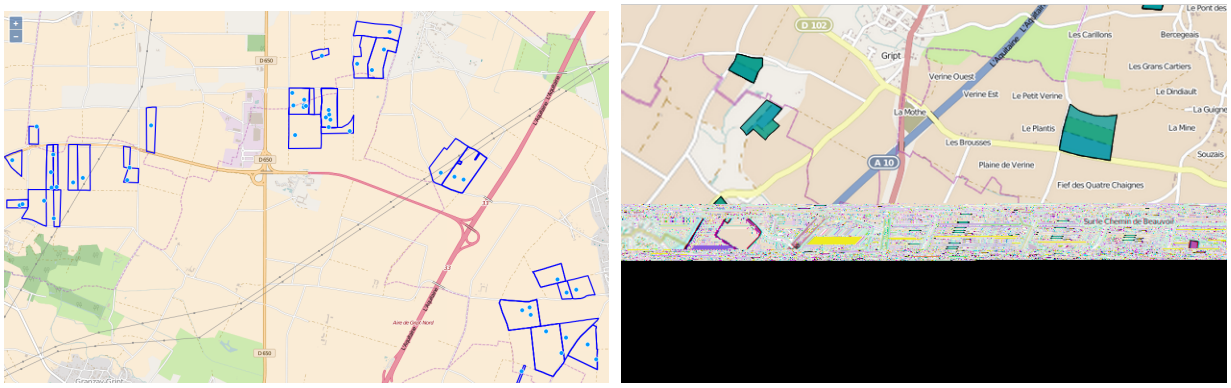
— Les événements territoriaux appliqués sur les parcelles agricoles peuvent être découverts en combinant les relations spatio-temporelles qualitatives. Ils se détectent en incorporant la relation *intervallMeets* entre deux intervalles de temps et la relation *within* entre la géométrie des parcelles de chaque relevée de l'assolement. Par exemple, des événements d'intégration, dans laquelle une parcelle a été absorbée par une autre, en 2009, peuvent être récupérées et affichées dans la carte comme (Figure 4b) grâce à la requête (Code 4). Une analyse de ces événements ainsi que l'utilisation du sol des parcelles concernées peuvent révéler les préférences de la pratique des agriculteurs. Les données sur les sols et les données de journalisation des prix du marché pourraient être utilisés dans les analyses.

4 Conclusion et perspectives

Les travaux exposés 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". Ce projet reçoit des aides financières de la Fédération de Recherche en Environnement pour le Développement Durable (FREDD [13]) du PRES Limousin Poitou Charentes. Nous cherchons à proposer une plateforme ouverte et libre pour analyser et exploiter des données environnementales hétérogènes. Nous proposons une méthode 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, qui répond aux besoins d'analyse et d'exploitation des experts. L'approche introduite pourrait être réutilisée pour effectuer la gestion et l'analyse des données environnementales à long terme pour d'autres observatoires.

Dans nos perspectives, d'un part, nous considérons à intégrer d'autres ensembles de données de la zone d'atelier, tels que les insectes et les données botaniques, ou les données satellites. Il sera alors possible d'utiliser le système pour enrichir et qualifier nos sources de données. D'autre part, nous cherchons à montrer la faisabilité et la réutilisabilité de l'approche proposée en l'appliquant aux autres données spatio-temporelles hétérogènes. Nous prévoyons également de publier une partie de ces données sur le Web sous forme des données liées dans le but de faciliter les échanges avec d'autres ensembles de données disponibles, notamment météorologiques. Enfin, l'application de l'extraction de règles d'association sur la base de connaissances est également envisagée pour découvrir de nouvelles règles d'association et d'examiner celles qui sont déjà connues (par exemple : la relation entre les espèces et les cultures des parcelles et entre les espèces elles-mêmes dans leur chaîne alimentaire).

13. http ://fredd.cue-lpc.fr/



(a) Corrélations entre les positions des Busards cendrés et la nature des cultures des parcelles en 2009.

(b) Événement d'intégration des parcelles en 2009

FIGURE 4: Analyses spatio-temporelles

Code 3: Corrélation entre la nidification des Busards cendrés et la nature des cultures

```
SELECT *
WHERE
{
?nest sige:nestOf ?indv.
?indv sige:belongTo ?species.
?species sige:name ?name.
?nest sige:hasNestObsv ?nobsv.
?nobsv sig:hasTime ?inst.
?inst time:inside ?intv.
?ts sige:hasTime ?intv.
?ts sige:hasLandUse ?lu.
?nobsv sige:hasGeometry ?geon.
?geon sige:geometry ?geomn.
?ts sige:hasGeometry ?geo.
?geo sige:geometry ?geom.
FILTER (strdf:within(?geomn,?geom)
&& ?name="Busard cendre")

}
```

Code 4: Détection des évenements d'intégration des parcelles

```
SELECT *
WHERE
{
?tsa sige:hasTimeSlice ?pc.
?tsb sige:hasTimeSlice ?pc2.
?tsc sige:hasTimeSlice ?pc.
FILTER(?pc!=?pc2)
?tsa sige:hasTime ?intva.
?tsb sige:hasTime ?intvb.
?tsc sige:hasTime ?intvc.
?intva time:intervallMeets ?intvc.
?intvb time:intervallMeets ?intvc.
?tsa sige:hasPGeometry ?geoa.
?geoa sige:geometry ?geoma.
?tsb sige:hasPGeometry ?geob.
?geob sige:geometry ?geomb.
FILTER(strdf:intersects(?geoma,?
    geomb))
?tsc sige:hasPGeometry ?geoc.
?geoc sige:geometry ?geomc.
FILTER(strdf:within(?geoma,?geomc)
&& strdf:within(?geomb,?geomc))
}
```

Références

AL-DEBEI M. M., AL ASSWAD M. M., DE CESARE S. & LYCETT M. (2012). Conceptual modelling and the quality of ontologies : Endurantism vs. perdurantism. *CoRR*.

ALLEN J. F. (1983). Maintaining knowledge about temporal intervals. *Commun. ACM*, **26**, 11.

BATSAKIS S. & ANTONIOU G. (2014). Representing and reasoning over spatial relations in owl : A rule-based approach. *Lightning Talks, W*, **3**.

BATSAKIS S. & PETRAKIS E. G. M. (2011). Sowl : A framework for handling spatio-temporal information in owl 2.0. In *Proceedings of the 5th International Conference on Rule-based Reasoning*, and Applications, RuleML'2011 : Programming.

BIZER C. (2004). D2rq - treating non-rdf databases as virtual rdf graphs. In *In Proceedings of the 3rd International Semantic Web Conference (ISWC2004)*.

CORBY O., DIENG-KUNTZ R. & FARON-ZUCKER C. (2004). Querying the semantic web with corese search engine. In R. L. DE MÁNTARAS & L. SAITTA, Eds., *ECAI*, p. 705–709 : IOS Press.

EGENHOFER M. & HERRING J. (1991). *Categorizing Binary Topological Relationships Between Regions, Lines, and Points in Geographic Databases*. Department of Surveying Engineering, University of Maine.

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.

HARBELOT B., ARENAS H. & C. C. (2013). Continuum : A spatio-temporal data model to represent and qualify filiation relationships. In *85. ACM* : Proceedings of the 4th acm sigspatial international workshop on geostreaming.

HARBELOT B., ARENAS H. & CRUZ C. (2014). Un modèle sémantique spatio-temporel pour capturer la dynamique des environnements". 14 ème conférence extraction et gestion des connaissances, rennes, france.

HOBBS J. R. & PAN F. (2004). An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing*, **3**.

KARMACHARYA A., CRUZ C., BOOCHS F. & MARZANI F. (2010). Use of geospatial analyses for semantic reasoning. In I. J. SETCHI, R. HOWLETT & L. JAIN, Eds., *R*, p. 576–586. Vol. 6276, p.. Springer Berlin Heidelberg : Knowledge-based and intelligent information and engineering systems.

KNUBLAUCH H., JAMES A. H. & KINGSLEY I. (2011). Spin - overview and motivation.

LANGRAN G. E. & CHRISMAN N. R. A. (1998). framework for temporal geographic information. *Cartographica : The International Journal for Geographic Information and Geovisualization*, **25**(3), 1–14.

PLUMEJEAUD C., MATHIAN H., GENSEL J. & GRASLAND C. (2011). Spatio-temporal analysis of territorial changes from a multi-scale perspective. *International Journal of Geographical Information Science*, **25**(10), 1597–1612.

RANDELL D. A., CUI Z. & COHN A. G. (1992). A spatial logic based on regions and connection. In *Proceedings 3rd International Conference On Knowledge Representation And Reasoning*.

RESSLER J., DEAN M. & KOLAS D. (2010). Geospatial ontology trade study. In *Proceedings of the 2010 Conference on Ontologies and Semantic Technologies for Intelligence*, p. 179–211, Amsterdam, The Netherlands, The Netherlands : IOS Press.

SEYE O., FARON-ZUCKER C., CORBY O. & GAIGNARD A. (2014). Publication, partage et réutilisation de règles sur le Web de données. In C. FARON-ZUCKER, Ed., *IC - 25èmes Journées francophones d'Ingénierie des Connaissances*, p. 237–248, Clermont-Ferrand, France.

VANDECASTEELE A., & NAPOLI A. (2012). Spatial ontologies for detecting abnormal maritime behaviour. In S. K. YEOSU, Ed., *OCEANS 2012 MTS/IEEE Yeosu Conference : The Living Ocean and Coast - Diversity of Resources and Sustainable Activities*.

WELTY C. & FIKES R. A. (2006). reusable ontology for fluents in owl. In *Proceedings of the conference on formal ontology in information systems*, p. 226–236 : p.. IOS Press.

Gestion Sémantique des Bulletins de Santé du Végétal dans le projet Vespa

Catherine ROUSSEY*, Stephan BERNARD*, François PINET *, Xavier Reboud**, Vincent CELLIER***
* Irstea de Clermont-Ferrand, 9 avenue Blaise Pascal CS 20085 63178 AUBIERE
** INRA Dijon UMR 1347 AGROECOLOGIE, 17 rue Sully, BP 86510 21065 DIJON Cédex
*** INRA, Centre de Dijon, UE 0115 Domaine Expérimental d'Époisses. Bretenière,

1 Introduction

Dans cet article, nous présentons le système que nous avons conçu et développé afin de faciliter l'accès à l'information et la recherche au sein des nombreux Bulletins de Santé du Végétal dans une optique de comparaison ou de meilleure visibilité d'une dynamique temporelle. Notre système vise les différents acteurs des filières agricoles, comme premiers utilisateurs de notre contribution. Les BSV mis en ligne sur les sites Web des organismes n'étaient pas toujours pérennes ; ainsi les BSV des années antérieures ne sont souvent plus accessibles sur leurs sites. Dans notre système, nous avons collecté et archivé les BSV des différents sites. Le système que nous avons mis en place permet de stocker et de rendre accessible de manière pérenne des archives des BSV. Il offre ainsi un point d'accès unique aux BSV et le système rend possible la recherche dans ce corpus. Afin de rendre possible cette recherche, nous avons du décrire le contenu de chaque BSV par des annotations. Ces annotations ont été extraites semi automatiquement à partir des sites Web des organismes. Nous avons publié ces annotations sur le Web de données liées. Nos annotations sont des données structurées associées aux BSV. Ces annotations permettent des recherches selon différents critères dans le corpus. Elles sont publiées sur le Web de données liées afin de pouvoir être réutilisées par d'autres. Ainsi, grâce à notre système, il est par exemple possible de rechercher des BSV de régions différentes portant sur la même culture et la même période. Nous pourrions aussi compléter les annotations des BSV en les liants vers d'autres sources, tels les bulletins météo. Etablir un tel lien aurait du sens dans la mesure ou beaucoup de processus épidémiques de maladies ou de ravageurs des cultures sont très dépendant des conditions météorologiques telles que température ou humidité. Permettre d'accéder facilement à ces données supplémentaires peut grandement faciliter les interprétations et prévision sur l'état sanitaire des cultures dans les régions pour des périodes données, etc.

Avec notre système, un utilisateur peut par le biais d'un seul point d'accès, interroger l'intégralité du corpus, pour se constituer son propre corpus de travail ne contenant que les BSV qui répondent à son besoin d'information. Trois classes d'annotations ont été utilisées :

1. Spatiale: la région de publication des BSV
2. Temporelle: la date de publication des BSV
3. Thématique: la culture principale du BSV mentionné sur le site Web de l'organisme.

Les utilisateurs peuvent par exemple rechercher les BSV par région, par période (dates, intervalles de dates, mois, année, etc.), par cultures ou familles de cultures, etc.

Pour satisfaire les besoins d'information des utilisateurs, nous avons mis en place dans notre système des annotations de qualité. Comme les BSV sont disponibles sur le Web de

données liées, tout organisme peut rajouter ses propres annotations pour compléter notre description des BSV.

2 Présentation des Bulletins de Santé du Végétal (BSV)

En France, le Grenelle de l'environnement et le plan Ecophyto ont renforcé les réseaux nationaux de surveillance sur les cultures et les pratiques agricoles. Les Bulletins de Santé du Végétal sont une des modalités mises en place par ces réseaux de surveillance. Le Bulletin de Santé du Végétal (BSV) est un document d'information technique et réglementaire, rédigé sous la responsabilité d'un représentant régional du ministère de l'agriculture, tel qu'une Chambre Régionale d'Agriculture ou encore la Direction Régionale de l'Alimentation, de l'Agriculture et de la Forêt (DRAAF). La figure 2 présente un exemple de BSV de la région Midi-Pyrénées. Ce représentant est tenu de mettre ses bulletins à disposition du public sur son site internet. Depuis quelques années, tous les BSV sont accessibles au format PDFdirectement sur le site dédié pour chaque région. La conséquence est que les BSV sont répartis sur différents sites web (un par région). Les BSV sont rédigés en collaboration étroite avec de nombreux partenaires impliqués dans la protection des cultures, réunis au sein d'un comité de rédacteurs. Ils ont pris le relais des avertissements agricoles. La liste des auteurs des BSV varie en fonction de la région et de la filière agricole, ce qui a pour conséquence que leur contenu et leur présentation ne sont pas uniformes et varient en fonction des auteurs. Les BSV diffusent des informations relatives à la situation sanitaire des principales productions végétales de la région et proposent une évaluation des risques encourus pour les cultures. Des données générales concernant les arrêtés de lutte obligatoire (notes nationales, . . .) ou les évolutions de la réglementation peuvent aussi figurer dans les BSV. Selon l'actualité sanitaire et/ou la culture, le rythme de parution des BSV est variable, allant d'une parution hebdomadaire à mensuelle. Les BSV sont donc une synthèse interprétée des observations effectuées en amont sur les cultures par différents organismes collecteurs. Les auteurs des BSV décident lors de leur réunion éditoriale si une observation doit être considérée comme un phénomène unique localisé ou bien comme relevant d'un phénomène d'ampleur potentielle importante et suffisamment représentatif pour être signalé. Comme de nombreux phénomènes sanitaires sont d'autant plus gérables qu'ils sont pris précocement, l'exercice s'avère souvent délicat. Ainsi, les BSV ne sont pas une agrégation automatique de données mesurées mais bien une synthèse humaine la plus consensuelle possible des jugements sur des observations.

BULLETIN DE SANTE DU VEGETAL
MIDI-PYRENEES

Grandes Cultures - n°28 — 8 juillet 2010

A retenir

MAÏS : Pyrale:
- 1ère génération : vol très étalé avec un pic enregistré entre le 13 juin et le 1er juillet selon les secteurs.
- 2ème génération: pic de vol prévu au cours de la 1ère décade d'août.

Sésamie :
2ème génération : le début du vol devrait intervenir entre le 20 et le 25 juillet, avec un pic prévu entre le 26 juillet et le 03 août.

MAÏS

• Stades phénologiques et état des cultures

Les températures de ces derniers jours ont fortement accéléré le développement du maïs. Aujourd'hui, le stade moyen est situé entre 14 feuilles à sortie paniculé, la floraison est proche pour la majorité des parcelles. Les derniers semis arrivent à 10 feuilles.

• Sésamie

Le vol de première génération est terminé. L'observation de pieds de ponte se poursuit. Toutefois ces attaques ne dépassent jamais 1 à 2 % de pieds touchés dans la parcelle. Les piégeages sont peu nombreux pour l'instant.

D'après le modèle, le pic de vol de deuxième génération devrait se situer entre le 26 juillet et le 03 août.

Évaluation du risque :

Le risque faible en premier génération se confirme. Toutefois, si le climat reste sec et chaud, les attaques de deuxième génération pourraient être significatives dans les secteurs les plus touchés en 2009.

• Pyrale

L'étalement du vol de première génération se confirme. Les piégeages restent significatifs sur la majorité des secteurs. Le pic de vol est dépassé dans l'ensemble de la région; du 13 juin au 1er juillet selon les secteurs.

Cependant, les conditions optimales pour les pontes et leur survie se situaient autour du 15 juin, ce qui devrait donner un maximum de vol de la deuxième génération lors de la première décade d'août.

Le début de vol devrait être significatif au 20 juillet dans les secteurs les plus chauds et au 25 juillet pour les secteurs les plus frais.

Figure 1 : Un bulletin de santé du Végétal de la région Midi-Pyrénées catégorie grande culture

3 Les vocabulaires RDF défini dans le projet Vespa

Pour stocker nos annotations des BSV nous avons défini un schéma d'annotation. Pour renseigner ce schéma, nous avons aussi défini plusieurs vocabulaires: un vocabulaire pour les régions et un vocabulaire pour les cultures. Concernant les cultures nous nous sommes rendu compte que chaque site web avait sa propre typologie de cultures en fonction des cultures principales de la région concernée. Nous avons donc défini un vocabulaire des cultures commun à toutes les régions, intitulé FrenchCropUsage. L'ensemble des annotations est stockée dans un *triplestore* RDF accessible en SPARQL (sous l'url ontology.irstea.fr/bsv/snorql)

3.1 FrenchCropUsage: Un vocabulaire hiérarchique pour décrire les types de cultures

À notre connaissance, il n'existait pas de ressource structurée française permettant de décrire les cultures par leurs usages ou leur destination. Les grandes classes d'usage de l'agriculture sont l'alimentation humaine ou l'alimentation animale. Certaines productions sont destinées à être transformées pour faciliter leur consommation. Par exemple, la production de houblon est destinée à la fabrication de la bière. Très peu de productions agricoles sont destinées à l'industrie sans avoir un but alimentaire. Nous pouvons citer par exemple le chanvre, qui est utilisé pour la fabrication de textile.
Notre but était de construire une hiérarchie des cultures en fonction de leur usage. Les liens hiérarchiques représentaient des relations de généralisation/spécialisation entre cultures (céréale/blé). Pour construire notre référentiel intitulé FrenchCropUsage, nous avons étudié les termes contenus dans des documents disponibles librement. Les documents étudiés sont :

- Les statistiques agricoles annuelles publiées sur le site de l'Agreste. Le document intitulé "la statistique agricole annuelle : présentation générale" décrit la hiérarchie des cultures pour répertorier l'ensemble de la production agricole [Agreste].
- Les métadonnées du registre parcellaire graphique présente une nomenclature des cultures ou groupes de culture [Registre Parcellaire].
- Les listes des noms de rubriques utilisées pour organiser les BSV sur chacun des sites web des chambres agricoles (une liste contient les rubriques "Arboriculture", "Grandes cultures", …).
- Le classement des cultures par groupe d'usage proposé par Wikipédia [Wikipédia France Culture].
- Pour compléter par des définitions chacun des éléments de la hiérarchie nous avons recherché les définitions dans le Larousse Agricole [Larousse Agricole].
- en cas d'absence d'information dans le Larousse Agricole, nous avons utilisé le portail français de l'agriculture de Wikipédia [Portail Agricole]. Des absences de définition sont à noter surtout pour tous les fruits tropicaux.

L'ensemble de la hiérarchie a été modélisée à l'aide du vocabulaire SKOS proposé par le W3C [SKOS]. SKOS est un vocabulaire RDF permettant de décrire des référentiels de type thésaurus. Il permet de décrire des concepts représentés par des termes (en utilisant la classe principale, skos:Concept) et d'exprimer les relations entre ces concepts. Par exemple il existe une relation hiérarchique qui exprime une relation de spécificité (ou de généricité) entre concepts.
Notre vocabulaire de type de culture est disponible sur le web de données liées. Il contient 272 concepts. La profondeur maximale de la hiérarchie est de 6 niveaux. Chaque concept est défini par les propriétés suivantes:

- skos:prefLabel contient le terme préféré utilisé comme étiquette du concept en français. En général, le terme est le nom usuel de la plante cultivée suivi de son usage.
- skos:altLabel contient les autres termes qui peuvent être utilisés comme étiquettes du concept.
- skos:definition contient la définition en français du concept justifiant sa position dans la hiérarchie.
- skos:inScheme exprime l'appartenance du concept au référentiel.
- rdfs:seeAlso contient un lien web vers une définition retenue lors de la construction du référentiel, comme par exemple les définitions du Larousse Agricole
- skos:note contient au moins une définition trouvée dans une autre source comme l'agreste ou Wikipédia.

- skos:editorialNote contient la définition du Larousse Agricole. Pour des raisons de propriété intellectuelle cette propriété est supprimée dans la version en ligne.
- skos:broaderindiquelelien vers le concept plus générique.
- skos:narrower indique le lien vers le concept plus spécifique

Plusieurs requêtes SPARQL ont été utilisées pour contrôler et valider automatiquement le contenu du référentiel. Par exemple, une requête permet de contrôler les liens skos:narrower et skos:broader. Une requête vérifie que chaque skos:Concept est rattachée à la racine, possède au moins un skos:prefLabel et un skos:definition.

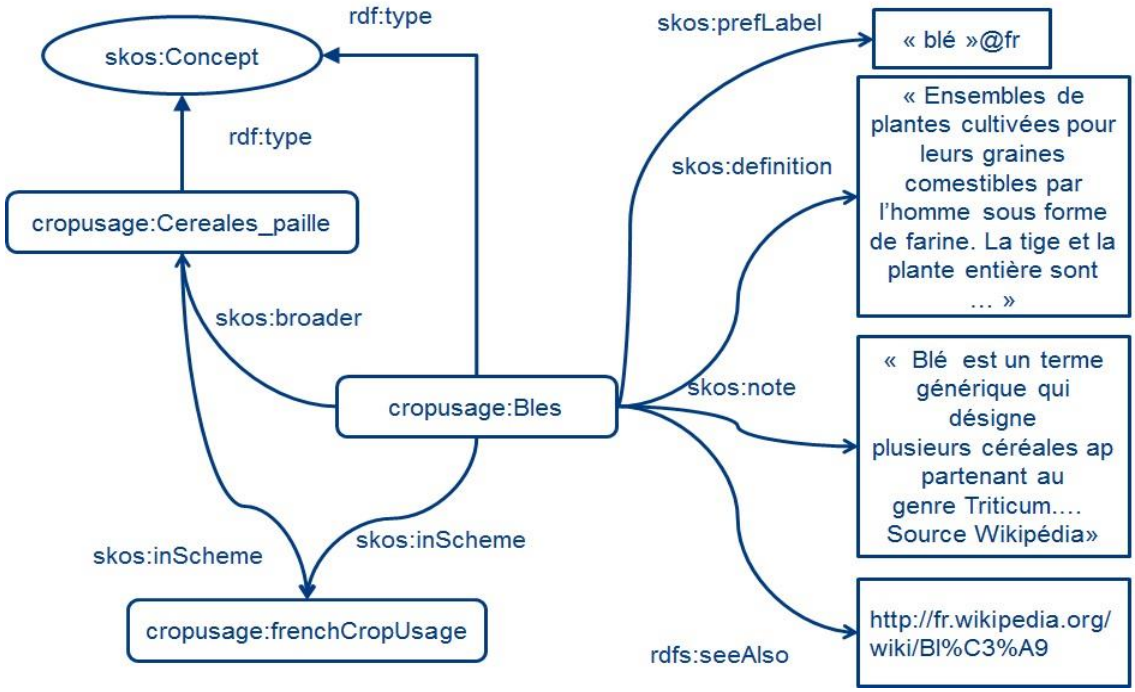


Figure 2 : Un exemple du contenu RDF de FrenchCropUsage

3.2 Un vocabulaire pour décrire les régions

Nous avons voulu associer à chaque bulletin sa région de publication: C'est à dire conserver l'information que ce bulletin a été disponible sur le site web de la chambre d'agriculture de telle région.
Pour simplifier l'interrogation des BSV, nous avons dupliqué une description des régions de France en réutilisant les jeux de données publiés sur le LOD de l'IGN, de l'INSEE et de DBPedia.
La description des régions est constitué de:

- une URI qui suit le patron de nommage suivant "http://ontology.irstea.fr/irstea/places#NumeroDeLaRegion".
- rdf:type : Cette propriété indique qu'une région est une instance de irstea:Region. Cette classe est définie comme équivalent à la classe Région de l'IGN et à celle de l'Insee.
- rdfs:label : Cette propriété stocke le nom de la région en toutes lettres. Cette donnée a été extraite automatiquement des jeux de données du LOD.

- owl:sameAs : Cette propriété indique au moins un lien d'équivalence vers l'un des jeux de données du LOD: IGN, INSEE et DBPedia.

Le jeu de données publié sur le LOD de l'IGN n'est pas complet. Il ne contient pas les départements et territoires d'outre-mer. C'est pour cette raison que nous avons dû compléter notre jeu de données avec celui de l'INSEE. Pour le moment, aucun de ces jeux de données ne décrit les nouvelles régions issues de la réforme des collectivités territoriales. Pour les bulletins de l'année 2016, nous avons défini ces nouvelles régions et indiqué par la propriété prov:wasDerivedFrom de quelle ancienne région elles sont issues.

3.3 Vespa: Un vocabulaire pour décrire le schéma d'annotations des BSV

Dans un premier temps nous avons stocké des informations extraites des sites web sur lesquels les bulletins ont été téléchargés (chambres d'agriculture, DRAAF, …). Ces informations sont structurées à l'aide des métadonnées du schéma d'annotation du Dublin Core (dcterms).
La figure 3 représente notre schéma d'annotation. Les propriétés que nous avons créées spécifiquement pour l'annotation des BSV sont préfixées par *vespa.*

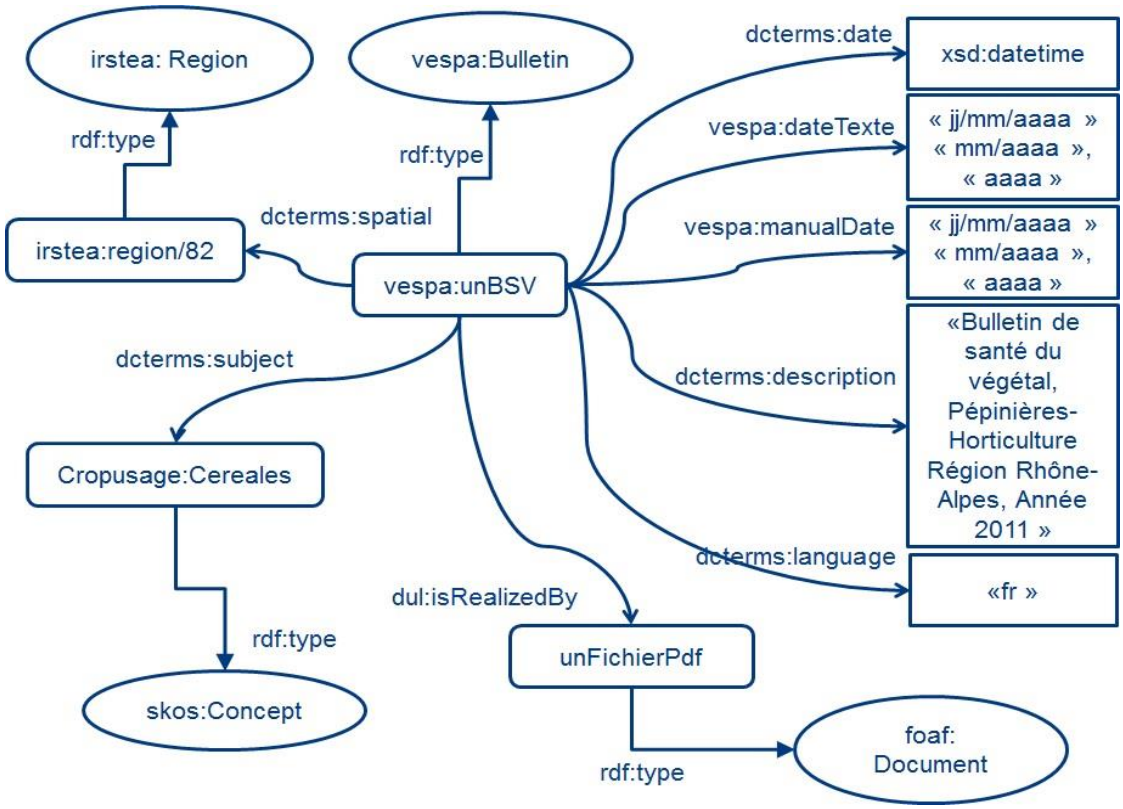


Figure 3 : Le schéma d'annotation général des BSV

Un bulletin est représenté par une instance de la classe vespa:Bulletin. Cette classe est une sous classe d'Objets d'Information. C'est à dire une entité abstraite qui regroupe l'ensemble des informations relatives à un objet indépendamment de comment cet objet est

physiquement. Par exemple un objet d'information est l'œuvre de Victor Hugo intitulé "les Misérables" et cet objet est indépendant du livre que vous avez sur votre étagère. Nous retrouvons cette notion dans la classe Œuvre de data.bnf.fr ou dans la classe Information Object de l'ontologie Dolce Ultra Light. Un objet d'information peut avoir plusieurs réalisations concrètes distinctes: un fichier PDF, une page html etc...Le lien entre l'objet d'information et sa réalisation (le fichier PDF) est indiqué par la propriété vespa:isRealizedBy.
Les annotations vont être portées par l'instance de la classe vespa:Bulletin.

Les propriétés utilisées pour décrire les BSV sont :

- vespa:textExtractionDate : contient une chaîne de caractères stockant la date de publication du bulletin (jj/mm/aaaa). Dans le cas de bulletin mensuel ou annuel, cette propriété contient le mois (mm/aaaa) ou l'année (aaaa). Cette propriété contient le résultat des processus d'extraction automatique.
- vespa:manualDate : contient une chaîne de caractères représentant la date de publication saisie manuellement au même format que vespa:textExtractionDate. Cette propriété, si elle existe, est considérée comme contenant une information exacte. Ces valeurs ont été renseignées lors de la validation de jeux de tests (moins de 4% des BSV).
- dcterms:date : contient la date de publication du bulletin, au format xsd:datetime. Dans le cas d'un bulletin mensuel ou annuel, la date est celle du premier jour de la période. Cette propriété contient la date de vespa:manualDate si elle est renseignée, ou sinon la date contenue dans la propriété vespa:textExtractionDate. Cette propriété est à utiliser en priorité pour l'interrogation des BSV.
- dcterms:description : contient une description du BSV (région, type de culture, année) qui correspond aux rubriques du site web où il a été téléchargé.
- vespa:isRealizedBy est le lien vers le fichier PDF associé.
- dcterms:spatial est le lien vers le nœud rdf représentant la région dans le jeu de données.
- dcterms:subject est le lien vers le skos:Concept du référentiel FrenchCropUsage. Cette propriété peut être utilisée plusieurs fois car un bulletin peut faire référence à différentes cultures.
- dcterms:language : est la propriété qui stocke la langue du bulletin, dans notre cas uniquement le français (fr).

4 Les processus de construction des annotations

Les annotations sont construites à partir des données issues des sites web (les sites web des chambres d'agriculture ou des DRAAF) donnant accès aux BSV. Chacun de ces sites propose un classement des bulletins de santé du végétal au minimum par année et par type de culture. Le type de culture est généralement repris dans le titre du bulletin. Comme le montre la figure suivante ces informations constituent la base du processus d'annotation des BSV. Une fois les annotations RDF spatio-temporelles construites elles sont publiées sur le web à l'aide d'un SPARQL end point. Un serveur apache dispose d'une adresse URL pour

chacun des fichiers PDF. Dans les sections suivantes nous allons détailler les différents processus d'annotation spatiale, thématique et temporelle.

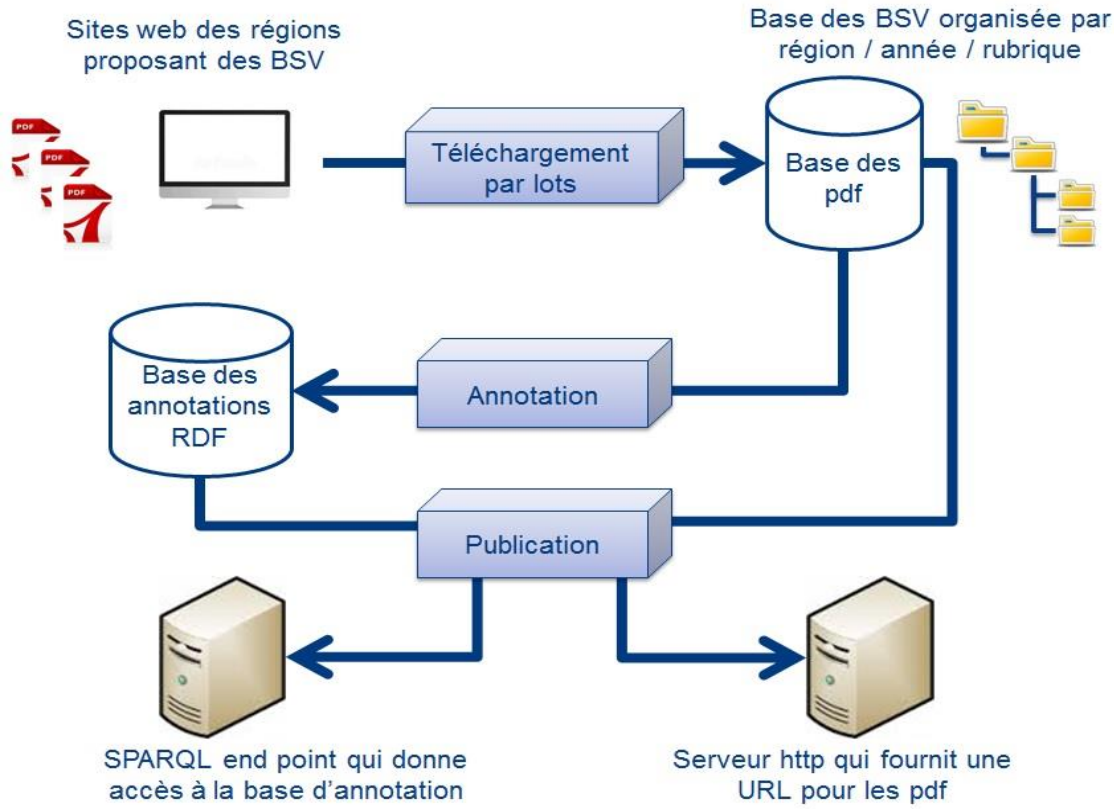


Figure 4 : La combinaison des processus utilisés pour publier les annotations des BSV sur le Web de données liées.

4.1 Annotations spatiales

La région est déterminée par le site web qui met les BSV à disposition. Par exemple, les BSV Auvergne se téléchargent sur le site web de la chambre régionale d'agriculture d'Auvergne. Donc le nom du site web est associé à la région concernée et tous les bulletins extraits du même site sont annotés par la même région.

Il arrive que des bulletins soient le fruit de collaborations inter-régionales. Dans ce cas, on retrouvera un même bulletin dans deux régions différentes. La détection de ces doublons n'a pas encore été réalisée. Par conséquent dans notre jeu de données le même BSV sera dupliqué et associé à des URI différentes, une URI par région.

4.2 Annotations thématiques sur les cultures

Chaque site web organise l'accès à ses BSV de manière différente mais au moins une des rubriques est relative à la culture. Certains noms de rubriques se retrouvent dans plusieurs sites web (comme par exemple "Grandes cultures"). D'autres sont spécifiques à la région. Par exemple, le site web de la région Midi Pyrénées découpe ses bulletins viticoles en plusieurs rubriques "Viticulture - Cahors, Lot", "Viticulture - Fronton", Etc. Le site web de la

région Île-de-France a défini une rubrique qui lui est propre “Grandes cultures - Pommes de Terre - Légumes industriels”.

Donc pour chacune des rubriques en lien avec le type de culture, nous leur avons attribuée manuellement un ensemble de concepts de notre référence FrenchCropUsage.

4.3 Annotations temporelles

Pour extraire la date de publication d’un bulletin, nous avons mis en place 3 processus d’extraction de date et réutiliser les sorties de l’outil PestObserver [Turenne et al., 2015].

Processus Nom de fichier

Notre premier processus d’extraction de date travaille sur les noms de fichiers téléchargés sur le site web. En effet, nous nous sommes rendu compte que ces noms de fichier portent parfois une indication de leur date de publication. Nous avons recherché plusieurs patrons de nommage de date dans ces noms de fichier pour extraire une date suivant le format jj/mm/aaaa. Un exemple de patron de nommage de date que nous avons utilisé est deux chiffres + “_” + mois écrit en lettre +”_” + année écrit avec 4 chiffres.

Processus Métadonnées

Notre deuxième processus allait chercher la date la plus ancienne stockée dans les métadonnées du fichier PDF. Cette date correspond le plus souvent à la clé CréationDate.

Processus Gate

Notre troisième processus recherche les dates dans le contenu du fichier PDF. Pour se faire nous avons utilisé la plateforme Gate et réutilisé le processus d’extraction de date standard. Vu le nombre de dates dans un bulletin, ce processus a été configuré pour ne rechercher que les dates complètes composées d’un jour d’un mois et d’une année. Comme ce processus retourne quand même plusieurs dates, notre processus ne conserve que la date qui apparait le plus souvent, car la date de publication est souvent répétée dans les bas de page.

Processus PestObserver

Enfin nous avons réutilisé les sorties de l’outil PestObserver [Turenne et al,. 2015] qui implémente un processus de reconnaissance des dates dans le contenu textuel des fichiers. Cet outil est capable de reconnaitre une date incomplète. Par exemple il ne peut trouver qu’une année ou un mois suivi d’une année. Par contre cet outil retourne la première date découverte. Il fait l’hypothèse que la date est toujours indiquée dans les premières lignes du bulletin.

Processus Fusion

Pour combiner les sorties de ces 4 processus distincts d’extraction de date, nous avons évalué leur résultats sur 500 bulletins pris au hasard. Pour se faire nous avons déterminé manuellement leur date de publication en lisant le contenu du bulletin. Cette date est donc renseignée dans la propriété vespa:manualDate.

Comme le montre le tableau récapitulatif suivant, aucun des processus n'est capable de trouver une date correcte pour l'ensemble des 500 bulletins.

- 53% des fichiers ont une date exprimée dans leur nom de fichier. Sur ces fichiers notre processus basé sur les noms de fichiers a extrait une date de publication correcte dans 92% des cas.
- 98% des fichiers ont bien une métadonnée qui donne une date de création. Notre processus de recherche de date dans les métadonnées a retourné une date de publication correcte dans 72% des cas. Ce taux parait faible mais souvent la date de création n'est éloignée que de quelques jours de la date de publication indiquée dans le contenu du bulletin.
- Le processus Gate a pu retourner une date pour 91% des fichiers. Cette date est correcte dans 85% des cas.
- L'outil PestObserver a fourni une date pour 82% des BSV de notre échantillon. Ces dates sont justes dans 86% des cas.

Méthode	Nb de bulletins retournés	Taux de bulletins retournés	Nb de bulletins où le processus à trouver la date correcte	Taux de réussite	Score attribué au processus
Nom de fichier	*263*	53%	242	92%	27
Gate	*454*	91%	384	85%	25
Métadonnées	491	98%	353	72%	22
PestObserver	411	82%	354	86%	26
Fusion des processus	500	100%	451	90%	

Tableau : résultat des différents processus d'extraction de date

Les taux de réussite nous ont permis d'attribuer un score à la sortie de chaque processus, en normalisant sur cent les taux de réussite des quatre processus. Par exemple, le score du processus nom de fichier est calculé à l'aide de l'opération suivante: 92*100/(92+85+72+86)=27,46. Si une date est trouvée par plusieurs processus, son score est la somme des scores de processus concernés. La date de publication retenue par la fusion des processus est celle qui a obtenu le score le plus élevé.
Cette méthode nous a permis de trouver une date de publication exacte pour 90% des BSV de notre échantillon. Si aucune date n'est trouvée par aucun des processus nous récupérons l'année qui est indiqué sur le site web.

Sauvegarde des résultats

Dans un souci de traçabilité, les sorties des différents processus sont stockées dans le schéma d'annotation des BSV. Les propriétés utilisées pour stocker les résultats des

processus d'extraction sont toutes des chaines de caractères qui suivent un format « jj/mm/aaaa » :

- vespa:filenameDate : contient le résultat du processus travaillant sur les noms de fichier.
- vespa:gateContentDate : correspond à la date de publication trouvée par le processus d'extraction de la plateforme Gate.
- vespa:dateMetadata : la date trouvée dans les métadonnées du fichier PDF.
- vespa:pestObserverDate : la date de publication trouvée par le processus d'extraction de date de l'outil PestObserver.
- vespa:dateExtractionQuality : sauvegarde le score obtenu par la date extraite automatiquement. Cette date est renseignée dans la propriété vespa:textExtractionDate. Une valeur de 100 indique que tous les processus automatiques ont renvoyé la même date.

5 Travaux connexes

Il existe de nombreux systèmes dédiés à l'extraction d'annotations spatiales temporelles et thématiques dépendant des sources de données utilisées. Nous pouvons par exemple citer les systèmes d'extraction d'évènementsutilisant des données disponible sur le webcomme EventMedia[Khrouf et al. 2012] ou LODE [Shaw et al., 2009]. Concernant l'extraction à partir de documents textuels, le projet Pyrénées Itinéraires Virtuels a développé des traitements linguistiques poussés dédiés à chaque type d'annotations[Enjalbert et Gaio, 2006], [Le Parc-Lacayrelle et al., 2008].

6Conclusion

Dans cet article nous avons décrit le système donnant accès à une archive annotée des Bulletins de Santé du Végétal publiée sur le Web de données liées. Nous avons décrit le schéma d'annotation et les deux vocabulaires utilisés pour construire les annotations spatiales et thématiques. Les annotations actuelles spatiales et thématiques sont générées à partir des informations des sites web. Seules les annotations temporelles utilisent des processus d'extraction d'information à partir du contenu textuel des bulletins. Pour obtenir des annotations plus fines nous devrons compléter les annotations thématiques par des processus d'extraction d'information.

Bibliographie

[Agreste] la statistique agricole annuelle présentation générale. Disponible à l'url http://www.agreste.agriculture.gouv.fr/IMG/pdf_methosaa.pdf

[Enjalbert et Gaio, 2006] Enjalbert, P.,Gaio, M. :Géosem. Traitements sémantiques pour l'information géographique Revue Internationale de Géomatique, 16 (2) (2006), pp. 181–194

[Khrouf et al. 2012] Khrouf, H. Milicic, V.,Troncy, R..EventMedia Live: Exploring Events Connections in Real-Time to Enhance Content. In 11th International Semantic Web Conference (ISWC'12), Semantic Web Challenge, Boston, USA, November 11-15, 2012.

[Larousse Agricole] Larousse agricole Édition 2002. Disponible à l'url http://www.larousse.fr/archives/agricole/

[Le Parc-Lacayrelle et al., 2008] Le Parc-Lacayrelle, A., Gaio, M., Sallaberry, C. : La composante temps dans l'information géographique textuelle. Document numérique, Vol 10, N°2, p129-148, 2008.

[Portail Agricole] Portail:Agriculture et agronomie. Disponible à l'url https://fr.wikipedia.org/wiki/Portail:Agriculture_et_agronomie

[Registre Parcellaire]Description de la couche Registre parcellaire graphique 2012 (îlots PAC)Métadonnée du 24/09/2013. Disponible à l'url http://piece-jointe-carto.developpement-durable.gouv.fr/DEPT063A/METADONNEES/N_RPG_2012_S_063_metadonnees.pdf

[Sanderson et al. 2013] Sanderson R., Ciccarese P., Van de Sompel H., « Designing the W3C open annotation data model », Proceedings of the 5th Annual ACM Web Science Conference, ACM, p. 366-375, 2013

[Shaw et al., 2009] Shaw, R Troncy, L Hardman. LODE: Linking Open Descriptions of Events in fourth Asian Conference, ASWC 2009, Shanghai, China, December 6-9, 2009. P 156-167.

[SKOS] SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. available at https://www.w3.org/TR/skos-reference/

[Turenne et al., 2015] Turenne N., Andro M., RoselyneCorbière R., Phan T.T. Open Data Platform for Knowledge Access in Plant Health Domain : VESPA Mining (2015) arXiv:1504.06077 http://arxiv.org/abs/1504.06077.

[Wikipedia France Culture] page de Wikipédia sur les classements des cultures disponible à l'url:
https://fr.wikipedia.org/wiki/Classement_en_France_des_cultures_par_groupes_d'usage

Exposing French agronomic resources as Linked Open Data

Aravind Venkatesan[1], Nordine El Hassouni[2], Florian Phillipe[3], Cyril Pommier[3], Hadi Quesneville[3], Manuel Ruiz[12], Pierre Larmande[145]

[1] Institut de biologie Computationelle, Montpellier, France
Aravind.Venkatesan@lirmm.fr
[2] UMR AGAP, CIRAD, Montpellier, France
{nordine.el_hassouni,manuel.ruiz}@cirad.fr
[3] URGI, INRA, Versailles, France
{fphilippe,Cyril.pommier,hadi.quesneville}@versailles.inra.fr
[4] UMR DIADE, IRD, Montpellier, France
[5] Equipe Zenith, INRIA et LIRMM, Montpellier, France
pierre.larmande@ird.fr

Abstract. The advancements in empirical technologies has generated vast amounts of heterogeneous data. This situation has created a need to integrate the data to understand the system of interest in its entirety. Therefore, information systems play a crucial role in managing these data, enabling the biologists in the extraction of new knowledge. The plant bioinformatics node of the Institut Français de Bioinformatique (IFB) maintains public information systems that houses domain specific data. Currently, efforts are being taken to expose the IFB plant bioinformatics resources as Linked Open Data, utilising domain specific ontologies and metadata. Here, we present the overview and the initial results of the project.

Keywords: Data integration, Data interoperability, Knowledge management, Linked Data, RDF, Bioinformatics application, Agronomic research

Introduction

Agronomy is an overarching field that encompasses various research areas such as genetics, plant molecular biology, and agro-ecology. The last several decades has seen the successful implementation of high-throughput technologies that have revolutionised research in agronomy. These technological advancements have resulted in a number of initiatives been taken to systematically store and share information over the web such as Gramene (Monaco et al., 2014), TAIR (Lamesch et al., 2012), OryzaBase (Kurata et al., 2006), Plant Reactome (Croft et al., 2014), GnpIS (Steinbach et al., 2013) and the South Green bioinformatics platform (http://www.southgreen.fr), to name a few.

However, using these resources comprehensively, taking advantage of the associated cross-disciplinary research opportunities poses a major challenge to both domain scientists and information technologists. Effective data integration and management allows a broader perspective across many disciplines, than is possible from one or a series of individual studies. A solution for the data integration challenges is offered by the Semantic Web (SW) technologies (Berners-Lee & Hendler 2001). The objective of the current effort is to develop RDF knowledge bases that integrates existing domain specific ontologies and data from the respective regional portals, promoting data interoperability between the resources. To this end, we have developed the Agronomic Linked Data knowledge base (www.agrold.org) that is representative of the data housed in the southern region portal of France, the SouthGreen Bioinformatics platform (SG) (http://www.southgreen.fr/).

Semantification of the IFB plant bioinformatics nodes

Institut Français de Bioinformatique (IFB) is a French national node (http://www.elixir-europe.org/about/elixir-france) that is focused on providing integrated services for the life science community. The IFB platform provides access to databases, tools and services that covers three main domains namely, microbial, plant and health sciences. The IFB IT infrastructure is linked to six regional bioinformatics centers, the ReNaBi (French Bioinformatics Platforms Network), representing various regions of the French territory (ReNaBi-NE, North-East; PRABI, Rhône-Alpes region; ReNaBi-GS, Great South; ReNaBi-SO, South-West; ReNaBi-GO, Great West and APLIBIO, Paris area). These six regional centers are consists of regional bioinformatics platforms (PFs). Taken together, IFB will represent France in the ELIXIR European infrastructure initiative. To this end, the plant bioinformatics PFs maintain public data repositories that ranges from 'omics' to genetic data (genetic markers, maps and phenotypes) for various crop species.

Currently, the plant-centric PFs are working towards exposing their resources as linked data. The objective of the current effort is to develop RDF knowledge base that integrates existing domain specific ontologies and data from the respective PFs. This will promote interoperability between the databases. In the initial phase, two representative PFs are involved in this semantification process, namely:

a) The *Unité de Recherche Génomique-Info* (URGI) platform (https://urgi.versailles.inra.fr/) associated with the *Institut National de la Recherche Agronomique* (INRA), dedicated to maintain curated information on plants and crop parasite. The platform is part of the APLIBIO ReNaBi and plays a key role in the Wheat Initiative (http://wheatis.org/).

b) The South Green Bioinformatics platform (SG) part of the ReNaBi GS mainly associated with *Centre de coopération internationale en recherche agronomique pour le développement* (CIRAD) and *Institut de recherche pour*

le développement (IRD) among other regional institutes. SG provides tools and databases dedicated for genomic resource analysis of southern and Mediterranean plants.

AgroLD for SG resources

Currently, SG consists of 12 databases covering various plant species such as Banana, Cocoa, Maize and Rice. AgroLD is being developed in phases to expose all of these databases as Linked Data. Currently, Phase I of AgroLD includes data from:

1. TropGeneDB (Hamelin et al. 2013), a database that hosts genetic, molecular and phenotypic information on tropical crop species.
2. OryGenesDB (Droc et al. 2006), a database that serves as a repository on functional genomics for rice.
3. Oryza Tag Line (Larmande et al. 2008), a database that contains sequence information (Flanking Sequence Tags) that are based on molecular categorisation of mutagen insertion sites for rice.
4. GreenPhylDB (Conte et al. 2008), provides sequence homology information for the members of kingdom *plantae*.

Additionally, domain specific ontologies, ontology annotations, proteomics and genomics information from a variety of publically available data sources have been integrated, this includes Gene Ontology, Plant Ontology, UniprotKB, Gramene (Gene, ontology annotation, gene, Quantitative Trait Loci (QTL) and Metabolic Pathway information) (Monaco et al. 2014). The objective of this is to provide the critical mass required to implement real world use cases. Currently, AgroLD includes data pertaining to selected species namely, *Oryza* species (*O.sativa, O.barthii, O.brachyantha, O. glaberimma* and *O.meridionalis*), *Arabidopsis thaliana*, *Sorghum bicolor*, *Zea mays* and *Triticum* species (*T.aestivum* and *T. uraruta*). In the subsequent phases information pertaining to other species and SG databases will be considered. The AgroLD effort will be further extended set-up RDF knowledge bases to host data from other regional portals.

References

Berners-Lee T. & Hendler J. 2001. Publishing on the semantic web. Nature, 410, 1023-4.

Barrell, D. et al., 2009. The GOA database in 2009 - An integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, 37(SUPPL. 1).

Conte, M.G. et al., 2008. GreenPhylDB: A database for plant comparative genomics. *Nucleic Acids Research*, 36(SUPPL. 1).

Croft D. et al. (2014). The Reactome pathway knowledgebase. Nucleic Acids Research. 42(D1), D472-D477.

Droc, G. et al., 2006. OryGenesDB: a database for rice reverse genetics. *Nucleic acids research*, 34(Database issue), pp.D736–D740.

Hamelin, C. et al., 2013. TropGeneDB, the multi-tropical crop information system updated and extended. *Nucleic Acids Research*, 41(D1).

Kurata N. & Yamazaki Y. (2006). Oryzabase. An integrated biological and genome information database for rice. Plant physiology. 140(1), 12-17.

Lamesh P. et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research. 40(D1), D1202-D1210.

Larmande, P. et al., 2008. Oryza Tag Line, a phenotypic mutant database for the Génoplante rice insertion line library. *Nucleic Acids Research*, 36(SUPPL. 1).

Monaco, M.K. et al., 2014. Gramene 2013: Comparative plant genomics resources. *Nucleic Acids Research*, 42(D1).

Steinbach D., et al., 2013. GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database, 2013. bat058.

Sensible characterization of datasets : A dissimilarity approach

William Raynaut
IRIT UMR 5505, UT1, UT3
Universite de Toulouse
william.raynaut@irit.fr

Chantal Soule-Dupuy
IRIT UMR 5505, UT1, UT3
Universite de Toulouse
chantal.soule-dupuy@irit.fr

Nathalie Valles-Parlangeau
IRIT UMR 5505, UT1, UT3
Universite de Toulouse
nathalie.valles-parlangeau@irit.fr

ABSTRACT

Characterizing datasets has long been an important issue for algorithm selection and meta-level learning. Most approaches share a potential weakness when aggregating informations about individual features of the datasets. We propose a dissimilarity based approach avoiding this particular issue, and show the benefits it can yield in characterizing the appropriateness of classification algorithms.

Keywords

Dataset characterization, Dissimilarity, Meta-attributes, Meta-learning, Algorithm appropriateness

1. INTRODUCTION

In the traditionnal meta-learning framework, the dataset characterization problem consists in the definition of a subset of dataset properties (meta-level features of the dataset) that should allow a fine grain characterisation of datasets, while still complying to the requirements of the meta-level learner employed. However, to fit most learners requirements, dataset properties have to be aggregated into fixed-length feature vectors, which results into an important loss in information [1]. Relating in a way to "anti-essentialist" approaches, we investigate the possibility that limitations in the classical representations of datasets are among the main obstacles to well performing algorithm selection. We are thus focusing our efforts toward the definition of a representation that would allow the use of all available information to characterize the datasets.

2. MOTIVATION

The dataset characterization problem has been addressed along two main directions. In the first one, the dataset is described through a set of statistical or information theoretic measures as in the STATLOG project [2], and in most studies afterwards [5]. The second direction of approach to dataset characterization focuses, not on computed properties of the dataset, but on the performance of simple learners over the dataset. It was introduced as landmarking in [4], where the accuracies of a set of simple learners are used as meta-features to feed a more complex meta-level learner and further developments introduced more complex measures over the models generated by the simple learners, such as structural properties of decision trees [3].

The dataset characterization problem has thus already received quite some attention in previous meta-learning studies, but the aggregation of meta-features into fixed-length vectors processable through the meta-level learner has been a constant source of information loss.

3. APPROACH

Let us consider two datasets, A and B depicted in Figure 1. A describes 12 features of 100 individuals, and B, 10 features of 200 individuals. Let us say we want to compare the results of a set of 5 statistical or information theoretic measures over each individual feature, like mean, variance, standard deviation, entropy, and kurtosis (as illustrated over the second feature of A in Figure 1).

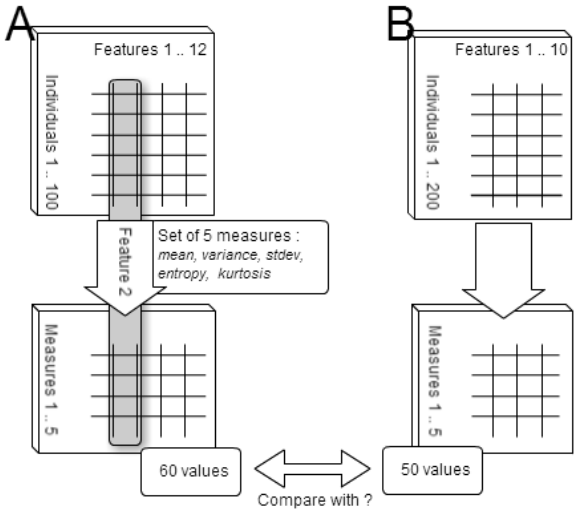


Figure 1: Measures over individual features

The complete information we want to compare is then a 60-values vector for A, and a 50-values vector for B. The standard approach would have been to average the measures over the different features, thus losing the characterization of the individual features (Figure 2)).

Our stance on the matter is to compare those features by most similar pairs, while comparing A's two extra features with empty features (features with no value at all). The assumption taken here is that a feature with absolutely no value is equivalent to no feature at all. To get back to our example, we end up comparing the 5 measures taken on the two closest (according to these very measures) features in A and B, then of the second closest, and so on, to finish on comparing the measures taken over the two extra features of A with measures taken over an artificial empty feature.

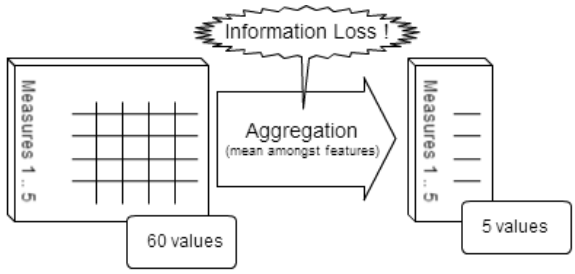


Figure 2: Averaging over individual features

These different comparisons sum up to an accurate description of how different A and B are, according to our set of measures. These pairwise comparisons would allow to ignore the presentation order of the features (which holds no meaningful information), focusing on the actual topology of the datasets.

4. VALIDATION

In [6], Wang & al. propose an intuitive definition of the goodness of dissimilarity functions in the context of learning. They define a dissimilarity function $d(x, x')$ to be *strongly* (ϵ, γ)-*good* for a given binary learning problem, if at least $1 - \epsilon$ probability mass of examples $z = (x, y)$ satisfy :

$$P(d(x, x') < d(x, x'') \mid y' = y,\ y'' = -y) \geq \tfrac{1}{2} + \tfrac{\gamma}{2}$$

In other words, the higher the chance the dissimilarity has to put examples of the same class closer together than those of different class, the greater the margin it will provide for separating the classes. This interpretation leads us to the definition of a binary problem that the proposed dissimilarity should be able to address.

Consider a set D of classification datasets, and a set A of classifiers. We execute every classifier of A on every dataset of D and measure a performance criterion c of the resulting model. Next, for each dataset x, we define the set A_x of the algorithms that are appropriate on this dataset along our performance criterion as those at most one standard deviation away from the best :

$$A_x = \{a \in A \text{ such that } |\max_{a' \in A}(c(a', x)) - c(a, x)| \leq \sigma_x\}$$

We can then consider, for each algorithm $a \in A$, the binary classification problem where instances are the datasets $x \in D$, and their class label stating whether a is appropriate on them. These problems thus characterize the appropriateness of the different algorithms on the datasets, which is an intuitive goal of the proposed dissimilarity. We can therefore compute for each algorithm $a \in A$ and dataset $x \in D$, the probability from which directly flows the (ϵ, γ)-goodness of d_ω^{ubr} :

$$P(d_\omega^{ubr}(x, x') < d_\omega^{ubr}(x, x'') \mid a \in A_x,\ a \in A_{x'},\ a \notin A_{x''})$$

The next result in [6] states that if d is a strongly (ϵ, γ)-good dissimilarity function, then there exists a simple classifier based on d that will, with probability at least $1 - \delta$ over the choice of $n = \frac{4}{\gamma^2} ln \frac{1}{\delta}$ pairs of examples of opposite class, have an error rate of no more than $\epsilon + \delta$. This result provides an easily understandable assessment of the dissimilarity adequateness to the problem.

We realised these measures over sets of datasets and classifiers from the OpenML meta-database, using in turn the full proposed dissimilarity, and the classic euclidean and Manhattan distances on the datasets meta-attributes. The γ parameter was brought as high as possible while keeping $\epsilon \leq 0.05$. Table 1 presents the δ and bound of error rate achievable for different numbers of examples and dissimilarity function.

	1000 examples		5000 examples	
	δ	error bound	δ	error bound
Proposed	0,871	0,921	0,501	0,551
Euclidean	0,945	0,995	0,755	0,805
Manhattan	0,952	1,002	0,783	0,833

	10000 examples		50000 examples	
	δ	error bound	δ	error bound
Proposed	0,251	0,301	0,001	0,051
Euclidean	0,570	0,620	0,060	0,110
Manhattan	0,613	0,663	0,086	0,136

Table 1: Error bound achievable with probability $1 - \delta$ by dissimilarity based classifiers for different numbers of examples

As we can see, the proposed dissimilarity seems to provide an improvement in characterizing the appropriateness of the different algorithms studied, giving good error bounds with much fewer examples. Yet this result is highly dependant on the choice of datasets and algorithms used to construct the appropriateness problems, and no assumption can be made toward its generalisability. What does stand, is that for *certain* algorithms, the use of the proposed dissimilarity will yield a significant improvement over classic distances in characterizing their appropriateness. Among the algorithms where the proposed dissimilarity most outperforms the other ones, we can note a majority of tree based classifiers. One can then postulate that the proposed dissimilarity characterizes well the appropriateness of tree-based classifiers, and thus that this appropriateness depends in a good part on the feature-specific meta-attributes it makes use of.

5. REFERENCES

[1] A. Kalousis and M. Hilario. Model selection via meta-learning: a comparative study. *International Journal on Artificial Intelligence Tools*, 10(04):525–554, 2001.

[2] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Upper Saddle River, NJ, USA, 1994.

[3] Y. Peng, P. A. Flach, P. Brazdil, and C. Soares. Decision tree-based data characterization for meta-learning. *IDDM-2002*, page 111, 2002.

[4] B. Pfahringer, H. Bensusan, and C. Giraud-Carrier. Tell me who can learn you and i can tell you who you are: Landmarking various learning algorithms. In *Proceedings of the 17th international conference on machine learning*, pages 743–750, 2000.

[5] R. Vilalta and Y. Drissi. A perspective view and survey of meta-learning. *Artif. Intell. Rev.*, 18(2):77–95, 2002.

[6] L. Wang, M. Sugiyama, C. Yang, K. Hatano, and J. Feng. Theory and algorithm for learning with dissimilarity functions. *Neural computation*, 21(5):1459–1484, 2009.

Explanation Dialogues in the Service of Durum Wheat Sustainability Improvement

Abdallah Arioua[1,2], Patrice Buche[1], Madalina Croitoru[3]

[1] UMR IATE, MONTPELLIER, FRANCE

[2] LIRMM, UM2, MONTPELLIER FRANCE

Résumé : We consider the application setting where a domain-specific knowledge base about Durum Wheat has been constructed by knowledge engineers who are not experts in the domain. This knowledge base is prone to inconsistencies and contradictions. The goal of this work is to propose a dialogue model of explanation to facilitate knowledge acquisition in inconsistent knowledge bases in order to (1) reduce inconsistencies ; and (2) acquire more domain knowledge to enrich the knowledge base.

Mots-clés : Argumentation, Explanation Dialogue, Inconsistency.

1 Introduction

The Dur-Dur research project [1] aims at reducing the use of pesticide and fertilizer in Durum Wheat cultivation while providing a protein-rich Durum Wheat. In the project we construct a multidisciplinary *Datalog*± (Calì *et al.* (2012)) knowledge base which will be used as a reference for decision making. Since this knowledge base is built by non-expert knowledge engineers, it is potentially prone to *inconsistencies*. In the litterateur, inconsistencies are addressed by reparation techniques : one only keeps the contradiction-free subsets of knowledge (Lembo *et al.* (2010)). In the project, this approach is too drastic as it **removes** a lot of expert knowledge.

To circumvent such problem we consider the following solution. We start with a system that uses a *prototypical* knowledge base [2] that has been constructed manually by non-experts, then on top of such knowledge base we allow for querying under inconsistency using the semantics of Lembo *et al.* (2010) to be able to reason with the available knowledge, then we allow for another facility called *explanation dialogues*. It works as follows, when the `Expert` queries the knowledge base and gets an unexpected result he/she can be engaged in a dialogue with the `System` to understand why the `System` has answered with such result. The `Expert` can ask further questions and the `System` can respond accordingly. The main salient point is that the `Expert`, when he/she perceives inconsistencies, contradictions, or errors can give *argumentative feedback* where he/she opposes to the result obtained by the `System`, and proposes a correction. This approach helps in reducing inconsistencies without removing possibly important pieces of knowledge.

To understand the approach consider the following query $Q =$ "*Is there any technical itinerary where do we use Maize as precedent* ?" which has been asked by the `Expert` and to which the `System` has answered *yes*. The `Expert` wants an explanation about the reason behind Q's entailment :

1. `http://www.agence-nationale-recherche.fr/?Projet=ANR-13-ALID-0002.`
2. The Durum Wheat Knowledge Base can be found in : `http://www.lirmm.fr/~arioua/dkb`

1. `Expert` : Explain why do we use Maize as a precedent ?
2. `System` : The burial of Maize residue will reduce the dose of nitrogen fertilizer.
3. `Expert` : Could you elaborate ?
4. `System` : The buried residues of Maize will enhance the soil by organic matter which in turn will give sufficient nitrogen to the plant, consequently we will use a lesser dose of nitrogen fertilizer.
5. `Expert` : I don't agree. Using the Maize is risky because burying its residues exposes the plant to a toxin contamination.

After stage (5) the `System` can either : (6) respond with another explanation or ($6'$) declares its inability of providing another explanation. If the `System` opts for (6) then the `Expert` can respond by (7) a feedback stating that he doesn't understand the new explanation, or ($7'$) a feedback stating that he doesn't agree thus advancing an argument or ($7''$) declaring understanding. If the `System` opts for ($6'$) the dialogue ends and all arguments advanced by the `Expert` are stored. Note that the `System` doesn't attack `Expert`'s arguments because we assume that the human expert has the authority. An alternative to stage (5) would be ($5'$) the `Expert` acknowledges *understanding* of using a Maize precedent. Then the dialogue ends.

This dialogue exposes the content of the knowledge base in a goal-directed manner. As been highlighted the importance of such dialogue (besides making the `Expert` understand the entailment of Q) is to have an *argumentative feedback*. In fact, that is what happened at stage 5 where we come to know that the knowledge base lacks an important piece of knowledge that has been highlighted by the `Expert`.

The explanation dialogue is formalized within the frameworks of McBurney & Parsons (2002); Walton (2011), where locutions as "explain", "providing explanation", "argumentation" , etc have been considered. Moreover, the generation of arguments and explanations is based on a logical instantiation of Dung's abstract argumentation frameworks Dung (1995). The dialogue has been implemented in a system called DALEK (**DiAL**ectical **E**xplanation in **K**nowledge-bases) [3]. A preliminary evaluation with two experts showed a promising result when comparing with approaches without explanation dialogues. In the evaluation we observed an increase in the acquired knowledge (45%) and a considerable reduction of inconsistencies (24%).

Références

CALÌ A., GOTTLOB G. & LUKASIEWICZ T. (2012). A general datalog-based framework for tractable query answering over ontologies. *Web Semantics : Science, Services and Agents on the World Wide Web*, **14**, 57–83.

DUNG P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-persons games. *Artificial Intelligence*, **77**(2), 321–357.

LEMBO D., LENZERINI M., ROSATI R., RUZZI M. & SAVO D. F. (2010). Inconsistency-tolerant semantics for description logics. In *Proceedings of RR'10*, p. 103–117 : Springer-Verlag.

MCBURNEY P. & PARSONS S. (2002). Games that agents play : A formal framework for dialogues between autonomous agents. *Journal of Logic, Language and Information*, **11**(3), 315–334.

WALTON D. (2011). A dialogue system specification for explanation. *Synthese*, **182**(3), 349–374.

3. Can be found here : `http://www.lirmm.fr/~arioua/dkb/#rulesdalek`

Prise de décision à partir de données environnementales imparfaites

André Miralles[1], Franck Ravat[2] et Thérèse Libourel[3]

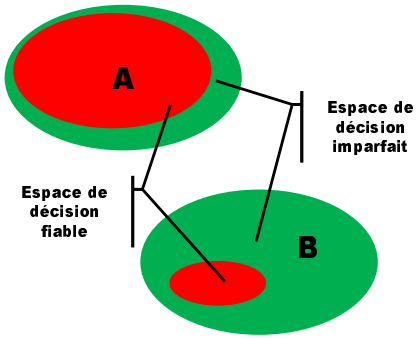
[1] Irstea – UMR Tetis - 34093 Montpellier Cedex 05 - andre.miralles@teledetection.fr
[2] Université Toulouse I Capitole - IRIT - 31062 Toulouse cedex 09 - Franck.Ravat@irit.fr
[3] Université de Montpellier - UMR Espace-Dev - 34093 Montpellier Cedex 05 - therese.libourel@umontpellier.fr

La vision standard des Systèmes d'Information d'Aide à la Décision (SIAD) repose sur d'une part, des sources transactionnelles classiques ou des répertoires de fichiers et, d'autre part, sur un ou plusieurs entrepôts de données à partir desquels sont effectuées des analyses décisionnelles reposant sur le principe OLAP (On-Line Analytical Processing), analyses restituant les indicateurs d'intérêt du domaine. Les entrepôts sont mis à jour régulièrement en extrayant les données des systèmes d'information transactionnels.

Un entrepôt de données étant matériellement indépendant du système d'information transactionnel, l'information y est structurée dans une vision d'analyse dans des cubes multidimensionnels dont les axes correspondent aux axes d'analyse des acteurs. Dans ces cubes multidimensionnels, l'information y est stockée avec comme objectif de réduire les temps de réponse afin qu'ils soient compatibles avec la prise de décision et de plus selon la granularité d'analyse souhaitée par les décideurs. Par exemple, sur la dimension spatiale, l'information peut être restituée à la parcelle, à la commune, au département, etc. Ces granularités sont en fait les échelles spatiales de décision pour de nombreux décideurs.

L'expérience montre que, dans beaucoup de domaines (pesticides, inondations, avalanches, etc.), les données environnementales stockées dans les systèmes transactionnels contiennent une part importante d'imperfection (Poujol et al., 2011; Vernier et al., 2013) et ce pour diverses raisons : changements de classe de précision des instruments de mesure de télédétection ou des capteurs, mise en place ou évolution des protocoles d'acquisition de données, amélioration de la précision de la géométrie des éléments du paysage, coût d'acquisition onéreux (satellites, enquêtes de terrain détaillées...), disponibilité hétérogènes de données cadastrales. De plus, les contraintes réglementaires comme la mise en place de la Directive européenne INSPIRE (http://www.developpement-durable.gouv.fr/La-directive-europeenne-Inspire-de.html) vont induire des problématiques d'imperfection au vu du nombre et de la disparité des secteurs concernés.

Dans le contexte environnemental, l'ensemble de l'information dont dispose le décideur relève de deux sous-ensembles que nous nommerons Espace de décision fiable et Espace de décision imparfait (cf. schéma). Lors d'une inondation, dans un contexte de gestion de crise, l'espace de décision fiable pourrait par exemple mobiliser le niveau du cours d'eau enrichi de la précision de la mesure mais cela n'est pas toujours possible. Dans le cas de données anciennes, de données obtenues via des techniques de crowdsourcing, etc., la mesure du niveau d'eau pourrait être disponible mais pas la précision. Cette infprmation appartient à l'espace de décision imparfait. Bien que non conforme aux données de l'Espace de décision fiable, cette donnée a une valeur intrinsèque surtout si elle est rare (crue centenaire par exemple). Classiquement dans ce contexte, cette information incomplète ne serait pas intégrée au SIAD car elle est imparfaite.



Tant que l'espace de décision fiable représente une part importante (cf. situation A), l'impact sur la décision finale reste acceptable. Il en est tout autrement lorsque l'espace de décision fiable est beaucoup plus petit que l'espace de décision imparfait (cf. situation B). En outre, l'expérience montre que le périmètre entre les deux sous-ensembles n'est pas constant et qu'il peut évoluer dans le temps dans un sens ou dans l'autre.

Classiquement, l'intégration de données sources dans les entrepôts de données vise à estimer les données manquantes et à éliminer ou à corriger les données imparfaites[1] afin d'obtenir un Système d'Information d'Aide à la Décision ne présentant a priori aucune imperfection. De ce fait, une grande quantité de données disponibles, dont certaines peuvent s'avérer stratégiques, est systématiquement exclue du processus décisionnel. Or dans de nombreux domaines, il peut s'avérer important, voire indispensable, d'intégrer des données imparfaites, parcellaires et d'actualité pour la prise de décision. De plus, la correction des données, lors du processus d'intégration des données décisionnelles, peut s'avérer fastidieuse, incomplète (spécifique à un nombre limité d'erreurs) et longue à mettre en œuvre.

L'approche que nous proposons est novatrice pour les Systèmes d'Information d'Aide à la Décision car nous ne souhaitons pas supprimer les données imparfaites mais les intégrer. Cette intégration au sein du processus d'entreposage des Systèmes d'Information d'Aide à la Décision soulève plusieurs questions de recherche :

- Comment enrichir la donnée par une sorte de «métadonnées» permettant d'expliciter la nature et la forme de l'imperfection ?
- Comment prendre en compte l'imperfection dès la conception des entrepôts de données mais probablement des systèmes d'information classiques ?
- Comment rendre plus flexible l'intégration des données imparfaites (donnée + métadonnée) ?
- Comment réaliser les agrégations aux différents niveaux de granularité sur des données imparfaites (donnée + métadonnée) ?
- Comment les imperfections de données influencent les outils OLAP l'exploration multidimensionnelle ?
- Comment rendre plus explicite la qualité des indicateurs présentés au décideur final ?

L'un des points clés reste la représentation et l'explicitation de la donnée imparfaite. Pour ce faire, il faut s'approprier les connaissances des acteurs (qui elles aussi peuvent être imparfaites) mais aussi leur mode de raisonnement afin de mieux représenter ces données mais aussi la manière dont les acteurs les manipulent.

Cela peut conduire (i) à améliorer les techniques d'extraction d'informations à partir des données structurées, peu structurées ou pas structurées, (ii) à mobiliser des techniques d'évaluation et qualification des sources d'informations mais aussi des indicateurs calculés à partir de ces données, (iii) à prendre en compte des biais cognitifs dans les bases de connaissances imparfaites, etc.

1. **VERNIER, F., MIRALLES, A., PINET, F., CARLUER, N., GOUY, V., MOLLA, G. & PETIT, K.** (2013). EIS Pesticides: An environmental information system to characterize agricultural activities and calculate agro-environmental indicators at embedded watershed scales. *Agricultural Systems, Vol. 122*, pp. 11-21.
2. **POUJOL, G. & LABBE, S.** (2011, 07 juillet). *Fusion de données pour l'évaluation nationale du risque sismique : une approche préventive et générale des vulnérabilités.* SAGEO 2011 - International Conference on Spatial Analysis and GEOmatics, Paris, 8 + 1 poster.

[1] Il existe de nombreux travaux dont il serait difficile d'en faire état dans un résumé de deux pages.

Système de veille sanitaire pour analyser l'émergence et la propagation de maladies animales

Sylvain Falala[1], Jocelyn De Goër[2], Elena Arsevska[1], Mathieu Roche[3,4], Julien Rabatel[4], David Chavernac[1], Pascal Hendrikx[5], Barbara Dufour[6], Renaud Lancelot[1], Thierry Lefrancois[1]

[1] CIRAD & INRA, UMR CMAEE, Montpellier

[2] INRA, UR EPIA, CLERMONT-FERRAND

[3] CIRAD, UMR TETIS, Montpellier

[4] LABEX NUMEV, LIRMM, Montpellier

[5] ANSES, UCAS, Maisons-Alfort

[6] ENVA, EpiMAI, Maisons-Alfort

Résumé : La veille en santé animale, et notamment la détection précoce d'émergences au niveau mondial d'agents pathogènes, est l'un des moyens permettant de prévenir l'introduction en France de dangers sanitaires (Paquet *et al.*, 2006). Dans ce contexte, cet article présente une plateforme dédiée à la veille automatique allant du recueil des données textuelles (dépêches) jusqu'à la restitution synthétique des informations extraites.
Mots-clés : Veille sanitaire, Recherche d'Information, Extraction d'informations

1 Contexte

Dans le cadre de la thématique "Veille sanitaire internationale" de la Plateforme nationale d'épidémiosurveillance en santé animale (Plateforme ESA), le Cirad, l'ANSES et la Direction générale de l'alimentation (DGAl) développent depuis 2013 un système de veille automatique du Web qui effectue : *(1)* le recueil quotidien de dépêches épidémiologiques provenant de sources non officielles, incluant les médias électroniques, *(2)* l'extraction automatique d'informations (nom de maladie ou symptômes, lieu, date et espèce touchée) issues de ces dépêches et *(3)* une restitution synthétique et agrégée de l'information : cartes, séries spatiotemporelles.

Les maladies actuellement surveillées sont la peste porcine africaine, l'Influenza aviaire, la fièvre catarrhale ovine, la fièvre aphteuse et la maladie de Schmallenberg. L'outil est développé de façon générique et permet la surveillance d'autres maladies. Ce système sera utilisé par la Plateforme ESA pour la France et par le réseau de vétérinaires CaribVet situé dans les Caraibes.

2 Approche mise en œuvre

Le but de notre système de veille est de disposer d'un outil très réactif qui se veut complémentaire aux sources officielles comme l'Organisation mondiale de la santé animale (OIE) ou l'Organisation des Nations unies pour l'alimentation et l'agriculture (FAO).

Le recueil des dépêches s'appuie sur des requêtes constituées de mots-clés de maladies, d'hôtes et de symptômes pour collecter des articles issus de Google News. Ces mots-clés ont été définis par des experts et/ou par des méthodes de fouille de textes (Arsevska *et al.*, 2016).

Chaque article est prétraité et normalisé (suppression de balises HTML, reconnaissance de la langue, etc.) avant d'être stocké dans une base de données MySQL. Une interface Web permet de paramétrer le processus de recueil et de consulter les articles collectés (cf. Figure 1).

L'extraction d'information dans les dépêches collectées identifie les éléments clés (noms de maladies, lieux, dates, nombres et espèces d'animaux touchées). Elle repose sur des dictionnaires dédiés et des règles préalablement construites par un processus de fouille de données. Les premiers résultats sur un corpus de 357 dépêches montrent des scores d'exactitude d'environ 70% pour les informations spatiales et d'au moins 80% pour les autres types d'informations.

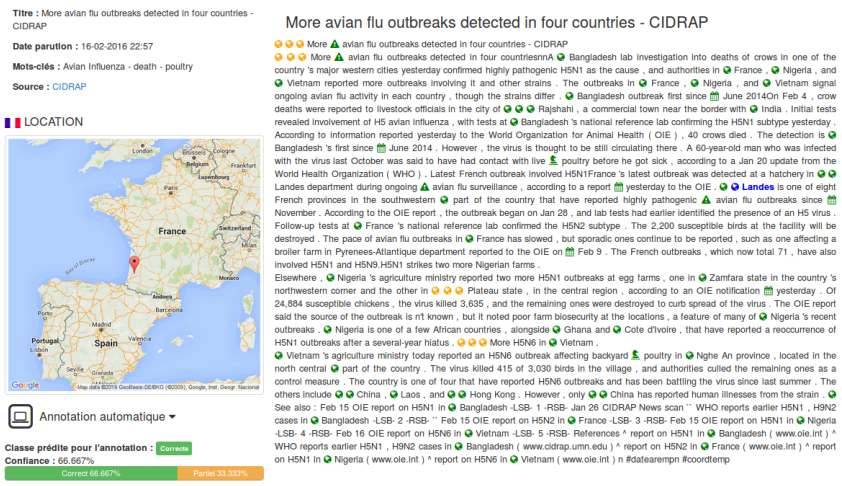


FIGURE 1 – Interface de consultation des dépêches.

3 Conclusion et perspectives

Cet article résume les travaux liés au développement d'une plateforme dédiée à la veille automatique allant du recueil des données textuelles (dépêches) jusqu'à la restitution synthétique des informations extraites dans les textes. Les informations actuellement extraites à partir des dépêches seront prochainement comparées aux informations issues des données officielles (OIE) afin de mettre en relief la découverte de l'émergence de maladies animales.

Remerciements : Les auteurs remercient les étudiants ayant participé au développement de l'outil : Max Devaud, Thomas Filiol, Baptiste Belot et Clément Hemeury. Ce travail est en partie financé par la DGAl et le Labex Numev (convention ANR-10-LABX-20).

Références

ARSEVSKA E., ROCHE M., HENDRIKX P., CHAVERNAC D., FALALA S., LANCELOT R. & DUFOUR B. (2016). Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, **123**, 104 – 115.

PAQUET C., COULOMBIER D., KAISER R. & CIOTTI M. (2006). Epidemic intelligence : a new framework for strengthening disease surveillance in europe. *Euro surveillance*, **11**(12), 212–214.