Gestion sémantique des bulletins de santé du végétal dans le projet Vespa

Catherine Roussey, Stéphan Bernard, François Pinet, Xavier Reboud, Vincent Cellier

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

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 paniculle, 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.

BULLETIN DE SANTÉ DU VÉGÉTAL – GRANDES CULTURES N° 28 du 08 JUILLET 2010 – Page 1/2

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.

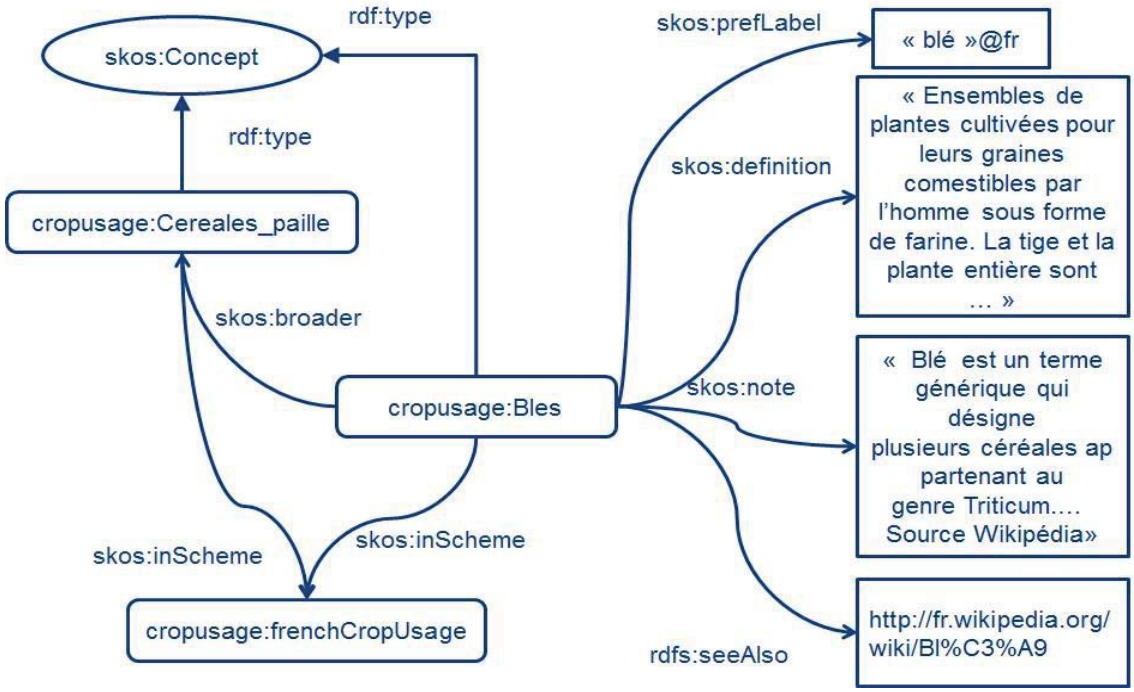

Figure 2 : Un exemple du contenu RDF de FrenchCropUsage

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.*

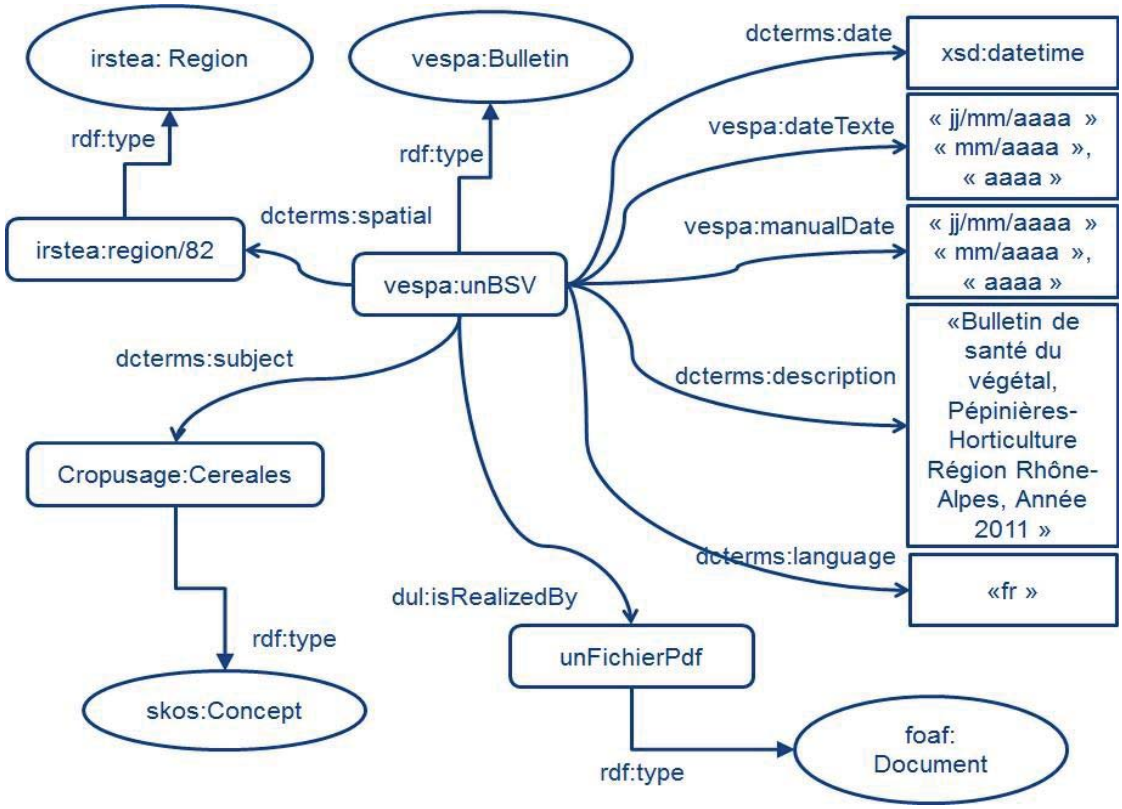

Figure 3 : Le schéma d'annotation général des BSV

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.

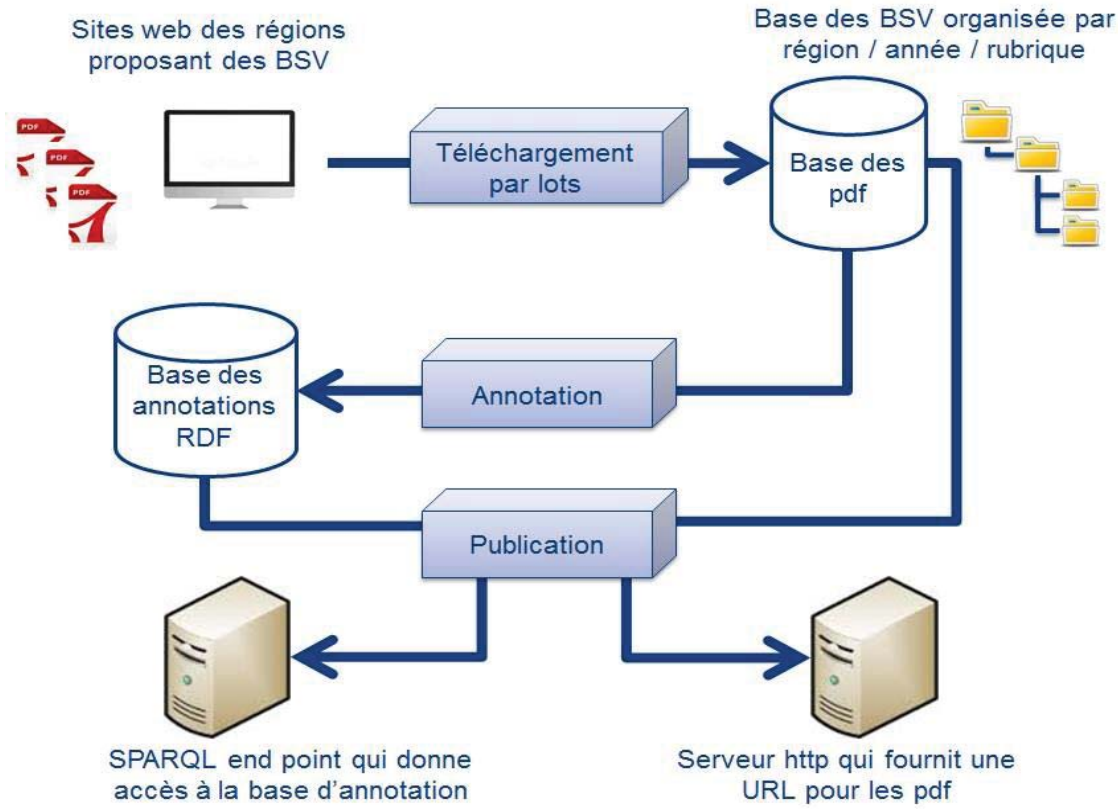

Figure 4 : La combinaison des processus utilisés pour publier les annotations des BSV sur le Web de données liées.

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