Annotation des Bulletins de Santé du Végétal

Catherine Roussey[1], Stephan Bernard[1]

UR TSCF, Irstea, 9 av. Blaise Pascal CS 20085, 63172 Aubière,France
prénom.nom@irstea.fr

Résumé : Dans cet article nous décrivons les différents schémas d'annotation envisagés pour annoter des bulletins agricoles disponibles sur le web. Notre but est de publier aussi sur le web de données les annotations manuelles permettant le catalogage des bulletins mais aussi les index utilisables par un système de recherche d'information sémantique.

Mots-clés : Annotations sémantiques, annotations spatio-temporelles, recherche d'information sémantique, bulletins agricoles.

1 Introduction

Pour être plus respectueuse de l'environnement, l'agriculture doit modifier ses pratiques, notamment au niveau de l'usage des produits phytosanitaires. Pour ce faire, le plan Ecophyto s'appuie entre autres sur le système de surveillance des pratiques agricoles, dont les Bulletins de Santé du Végétal (BSV) sont un des moyens de communication. Ce corpus du domaine agricole contient des informations sur les attaques des bio-agresseurs des cultures, région par région (par exemple : la DRAAF de la région PACA signale une explosion des attaques de la rouille du blé sur les cultures de blé dur en vallée du Rhône, dans son bulletin du 23 mai 2011).

Nous souhaitons mettre en place plusieurs processus d'annotation spatio-temporelle afin de permettre à des acteurs du domaine agricole de retrouver les BSV répondant à un besoin.

Cet article présente les schémas d'annotation que nous avons définis pour faciliter la recherche des bulletins agricoles sur le web de données. Tout d'abord nous présenterons le corpus des BSV. La section suivante présente un état de l'art des vocabulaires et ontologies que nous avons étudiés pour définir nos schémas d'annotation. Ensuite nous présentons les deux schémas d'annotation que nous proposons pour publier les BSV sur le web de données. Nous décrivons brièvement le premier processus d'annotation manuelle que nous sommes en train de mettre en place.

2 Le corpus des Bulletins de Santé du Végétal

Le Grenelle de l'environnement et le plan Ecophyto ont renforcé les réseaux 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 que la Chambre Régionale d'Agriculture ou encore la Direction Régionale de l'Alimentation, de l'Agriculture et de la Forêt (DRAAF). La figure 1 présente un exemple de BSV de la région Midi-Pyrénées. Ce représentant doit mettre ses bulletins à disposition du public sur son site internet. La conséquence est que les BSV sont

répartis sur différents sites web (un par région). À notre connaissance, il n'existe pas encore de système donnant un accès uniforme à l'ensemble des BSV.

Les BSV sont rédigés en collaboration avec de nombreux partenaires impliqués dans la protection des cultures. 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 stratégies de lutte (notes nationales, ...) ou sur la réglementation peuvent figurer également 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 une synthèse des observations effectuées sur les cultures. Il existe des bases de données d'observations mais la rédaction des BSV oblige leurs auteurs à décider si une observation est un phénomène unique non représentatif ou un phénomène important représentatif d'une réalité. Les BSV ne sont pas une agrégation automatique de données mesurées mais bien une synthèse humaine des jugements sur des observations.

Nous avons récupéré les BSV publiés entre 2009 et 2014 dans 24 régions, soit un peu plus de 15500 bulletins. En moyenne, une région publie plus d'une centaine de BSV par an. Notre but est de constituer une archive pérenne de ces bulletins agricoles afin d'en extraire un ensemble d'information sur les cultures et les niveaux d'attaques de ces cultures au cours du temps. Cette archive sera disponible comme jeux de données sur le web de données. Cette tâche d'archivage fait partie du projet Vespa "Valeur et optimisation des dispositifs d'épidémiosurveillance dans une stratégie durable de protection des cultures", dirigé par l'INRA.

3 Etat de l'art sur les vocabulaires RDF et ontologies utilisés pour l'annotation

Plusieurs vocabulaires RDF et structures de données du web sémantique (ou ontologie) sont proposés pour stocker des schémas d'annotations. Nous présentons dans la section suivante ceux qui ont servi de base à nos schémas d'annotation. L'annotation dans le monde des bibliothèques consiste à associer des données aux documents pour permettre leur catalogage et faciliter leur accès ; on parle alors de métadonnées. L'annotation sur le web consiste à associer à une ressource web une autre ressource (un tag, une note, un autre document).

3.1 DC : Dublin Core

Le Dublin Core est un vocabulaire RDF utilisé dans le monde des bibliothèques pour déclarer les métadonnnées des documents. Il est décrit dans DCMI Usage Board (2012). Ce vocabulaire définit une série de propriétés ("rdf :property") qui, en l'absence de déclarations plus précises, sont interprétées comme des "annotation properties" sur le web de données. La figure 2 présente une partie des propriétés du Dublin Core.

3.2 FOAF : Friend Of A Friend

FOAF, Brickley & Miller (2014), est un vocabulaire RDF définissant les relations (principalement professionnelles) entre personnes. Ce vocabulaire est basé sur un petit ensemble de classes : *Agent*, *Project*, *Organization*, *Document*, *Group*, etc...


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 SANTE DU VEGETAL – 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

Une personne se définit par un ensemble de "data type properties" : *name*, *age*, etc... Les relations entre personnes sont définies par l'"object property" *knows*, qui peut se spécialiser en fonction des besoins (par exemple, deux personnes créatrices d'un même document sont des *co-authors*). Cette information de création de documents est stockée par le biais de l'"object property" *maker*, entre une personne et le document qu'elle a créé, comme le montre la figure 3. FOAF a aussi été étendu pour stocker des données issues du web social.

3.3 SKOS : Simple Knowledge Organization Schema

SKOS ou Simple Knowledge Organization System (système simple d'organisation des connaissances) est un vocabulaire RDF proposé par le W3C pour représenter les thésaurus, les classifications et d'autres types de vocabulaires contrôlés ou de langages documentaires, W3C (2009).

SKOS permet de stocker les réseaux terminologiques constituant les vocabulaires contrôlés, utilisés entre autres par les documentalistes et les bibliothécaires. La figure 4 est un exemple

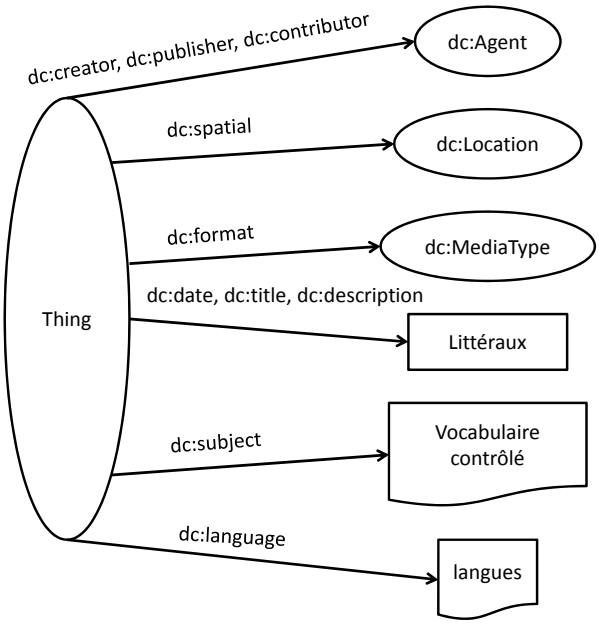

FIGURE 2 – sous ensemble des propriétés définies par le Dublin Core

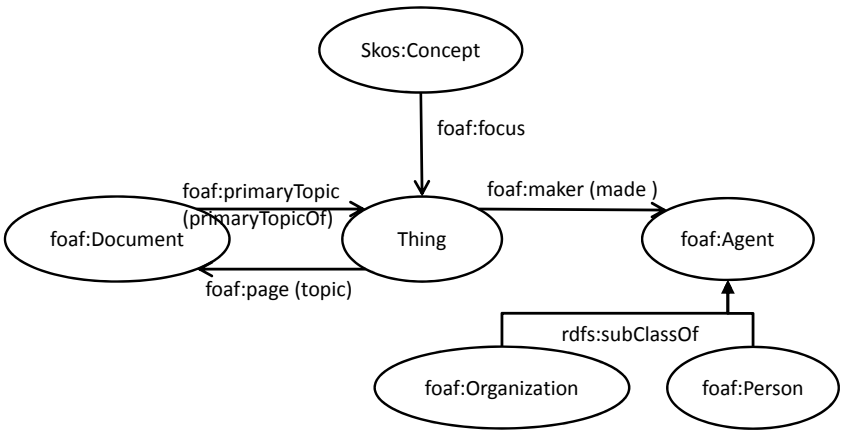

FIGURE 3 – extrait du vocabulaire foaf

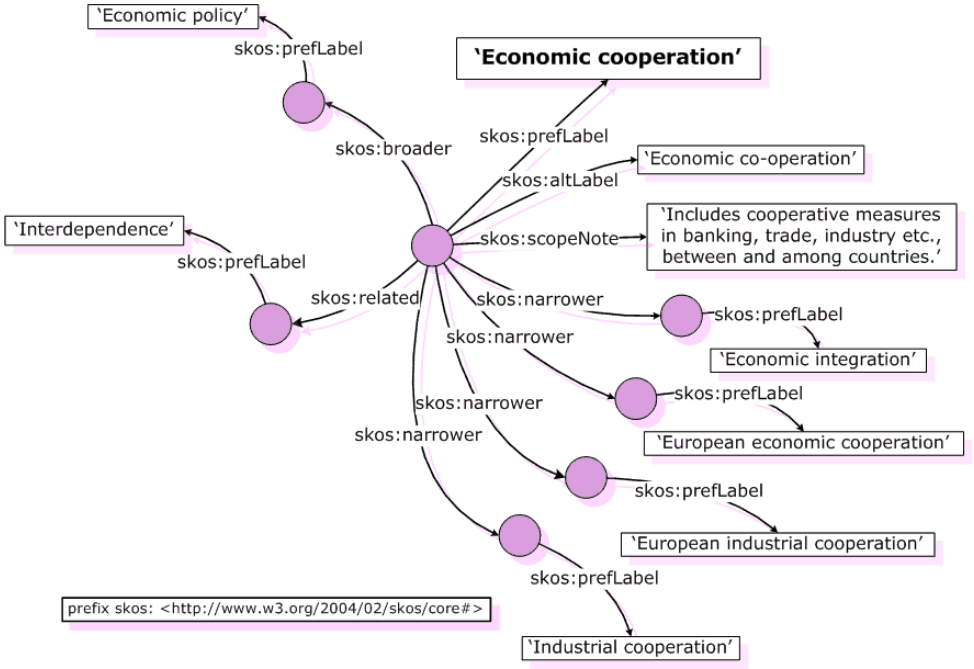

FIGURE 4 – graphe RDF utilisant le vocabulaire SKOS présentant les différents termes liés à "Economic Coopération"

de réseau terminologique issu de W3C (2009). Chaque nœud est un concept SKOS auquel sont rattachés des termes.

3.4 data.bnf.fr

Le schéma d'annotation de la BNF, Bibliothèque Nationale de France (2015), est fondé sur le schéma FRBR (Functional requirements for Bibliographic Records) élaboré par l'IFLA. Comme présenté dans la figure 5, ce schéma comprend trois groupes d'entités liées par des relations :

— les informations sur les documents sont déclarées avec le vocabulaire du Dublin Core,
— les informations sur les personnes physiques ou morales sont déclarées avec le vocabulaire FOAF,
— les informations sur les thèmes sont déclarées avec le vocabulaire SKOS.

Le groupe d'entités qui représente les documents décrit les différents aspects d'une production intellectuelle ou artistique à travers 4 niveaux : l'œuvre, l'expression, la manifestation et l'item.

— Le niveau de l'œuvre est celui de la création intellectuelle ou artistique. Un exemple est l'oeuvre intitulée les Misérables créée par Victor Hugo,
— le niveau de l'expression est caractérisé par la langue, le type de document et les liens de contributions (préfacier, illustrateur, traducteurs...),
— le niveau de la manifestation est celui de la matérialisation d'une expression. Un exemple de manifestation est une édition des Misérables « Nouvelle impression illustrée. 1879-1882. Paris. E. Hugues »,

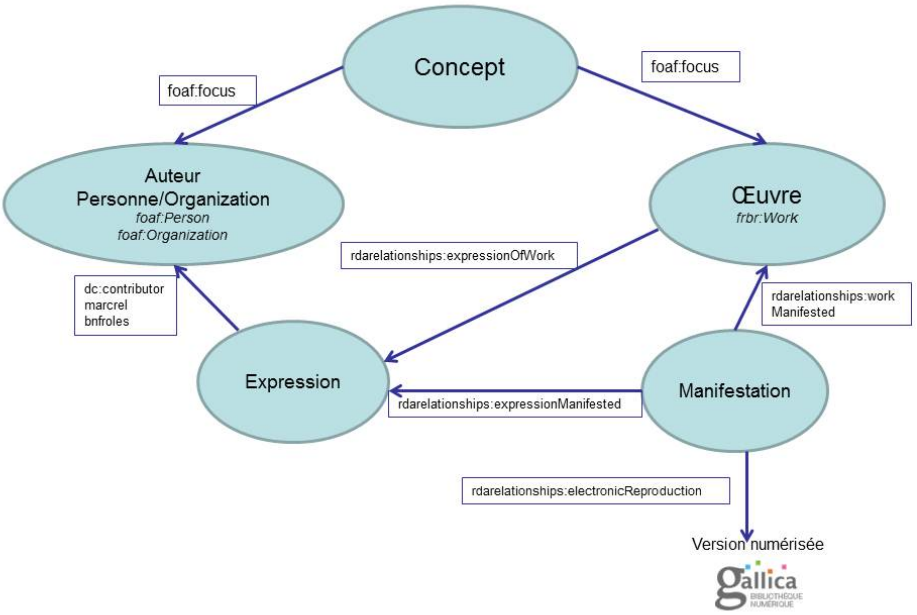

FIGURE 5 – le modèle RDF data.bnf.fr BNF(2015)

— le niveau de l'item est celui de l'exemplaire physique.

Une personne peut être auteur d'une œuvre ou contributeur d'une expression (préfacier, traducteur, librettiste...).

3.5 AO : Annotation Ontology

Cette ontologie est l'un des résultats du projet wf4ever visant à la préservation des résultats expérimentaux. Elle a ensuite donné naissance au projet researchObject pour la publication des ressources scientifiques (article, code, experimentation, etc...) sur le web de données.

Cette ontologie permet d'annoter les documents scientifiques disponibles sur le web à l'aide d'autres ressources, qui peuvent être des mot-clés issus d'un vocabulaire contrôlé (SKOS) ou d'une ontologie du domaine (OWL).

AO permet de préciser si le concept SKOS associé à un mot-clé représente exactement ou approche le contenu de l'annotation, à l'aide des relations *skos :broader* (sens plus générique) ou *narrower* (sens plus spécifique).

Les mot-clés peuvent aussi être une chaîne de caractères proposée par un humain sans contrôle. L'annotation ne se limite pas au «tagage» de document. Dans le contexte du projet wf4ever elle peut aller jusqu'à la prise de note voire la correction collaborative d'un document.

Cette ontologie a été mise en œuvre dans le domaine biomédical et les sciences du vivant (voir Ciccarese *et al.* (2011)). Elle a été utilisée en collaboration avec d'autres ontologies comme PAV qui est une spécialisation de l'ontologie de provenance du W3C pour l'annotation.

3.6 OA : Open Annotation Data Model

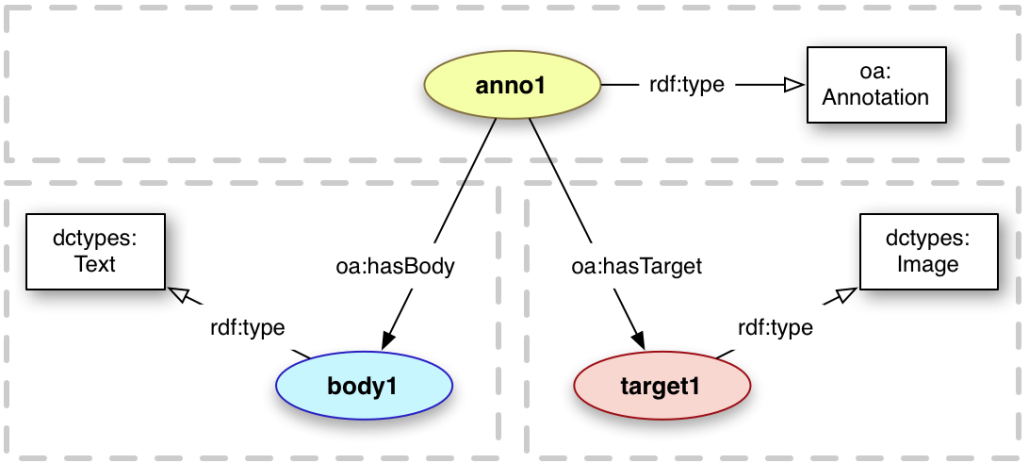

FIGURE 6 – le modèle RDF de base de Open Annotation Core W3C (2013)

Cette série d'ontologies est en cours de développement par un groupe du W3C Haslhofer *et al.* (2014). Les auteurs de AO participent aussi à ce groupe de travail.

L'ontologie Open Annotation Core vise à identifier et décrire les ressources liées à une annotation et à fournir des informations sur la création et l'intention associée à cette annotation ; W3C (2013).

Open Annotation peut être utilisé pour annoter des pages web, éditer collaborativement un document etc... On peut voir OA comme une généralisation et une simplification de AO. Par exemple, OA permet d'exprimer que le contenu de l'annotation est un graphe, sans ajouter plus de détail. Toutefois OA ne donne pas d'indication aussi spécifique que AO sur l'annotation sémantique d'un document web avec une ontologie ou un concept SKOS.

3.7 Synthèse

Nos objectifs sont multiples. Nous voulons tout d'abord proposer un schéma d'annotation permettant le catalogage des BSV afin de faciliter leur recherche. Pour ce faire nous avons choisi de travailler avec des schémas d'annotation standards mis en œuvre par de grandes institutions (BNF).

Nous souhaitons aussi que ces BSV soient utilisés par différents systèmes de Recherche d'Information Sémantique (RIS) et comparer les performances de ces systèmes. Le W3C développe un schéma d'annotation type qui deviendra, s'il est utilisé, un standard. Dans un système de RIS le contenu des documents est représenté par des vecteurs pondérés de concepts, un concept pouvant être soit un concept SKOS issu d'un vocabulaire contrôlé, soit un individu ou la classe d'une ontologie de domaine OWL. Ces vecteurs pondérés sont le résultat d'un processus d'indexation et sont donc appelés index.

Même si AO approche ce besoin, aucune de ces ontologies ne donne de solution pour stocker sur le web de données les vecteurs pondérés de concepts. Nous pouvons noter les travaux de Nešić [Nešić *et al.* (2010)] qui proposent de pondérer les termes utilisés pour l'annotation de document.

Le corpus des BSV sera indexé par différents processus d'indexation issus de plusieurs systèmes de RIS, et nous voulons pouvoir stocker, combiner et comparer ces différents index. Les résultats de plusieurs expériences d'indexation seront disponibles sur le web de données avec le corpus associé. Nous pourrons aussi simuler les résultats d'un système de recherche d'information à l'aide d'un moteur SPARQL en ordonnant les résultats d'une requête.

4 Nos schémas d'annotation

Nous allons proposer deux schémas d'annotation pour les BSVs. Le premier sera un schéma d'annotation pour stocker les métadonnées des BSV comme le ferait un documentaliste, le but étant d'indiquer la date de publication, la région et le type de culture associés à chacun des BSV. Le second sera utilisé pour stocker les index pondérés utilisables par un système de RIS, en étendant l'Open Annotation data model.

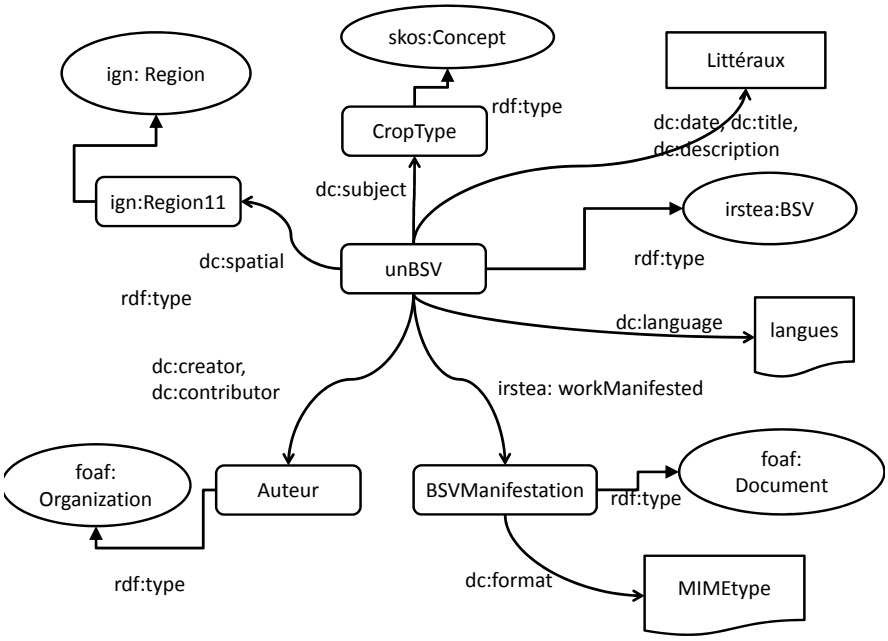

FIGURE 7 – schema d'annotation des BSV

Le premier schéma d'annotation, présenté dans la figure 7 est proche de celui utilisé par la BNF. Nous avons différencié l'entité représentant le BSV comme expression de la création intellectuelle de l'entité représentant sa manifestation physique. Il est en effet possible que différentes copies d'un même BSV soient accessibles sur le web de données avec des formats distincts.

Un bulletin agricole se caractérise par :

1. une métadonnée spatiale correspondant à sa région de publication, indiquée par la propriété *dc :spatial*. Cette propriété lie un bulletin à au moins une région définie dans le jeu de données RDF de l'IGN (`http://data.ign.fr/endpoint.html`).

2. une métadonnée temporelle correspondant à sa date de publication, indiquée par la propriété *dc :date*.
3. une métadonnée thématique correspondant aux types de culture abordées dans le bulletin agricole, indiquée par la propriété *dc :subject*. Cette propriété lie un bulletin à au moins un concept SKOS du thésaurus d'usage des cultures en France que nous avons défini.

L'ensemble de ces données deva être accessible sur le web de données et être utilisable par des moteurs d'inférence. Ce qui signifie que ces données ne doivent pas être enregistrées sous forme d'"annotation properties". Nous devrons définir des "data type properties" et des "object properties" similaires aux "annotation properties" du Dublin Core.

Le second schéma d'annotation a pour objectif de stocker les index produits par différents systèmes de RIS. La figure 8 présente la manière dont OA permet de stocker les informations relatives à la provenance d'une annotation. Ainsi nous pourrons indiquer à quelle date et par qui ont été produits les index, mais aussi quand et par qui ils ont été sauvegardés.

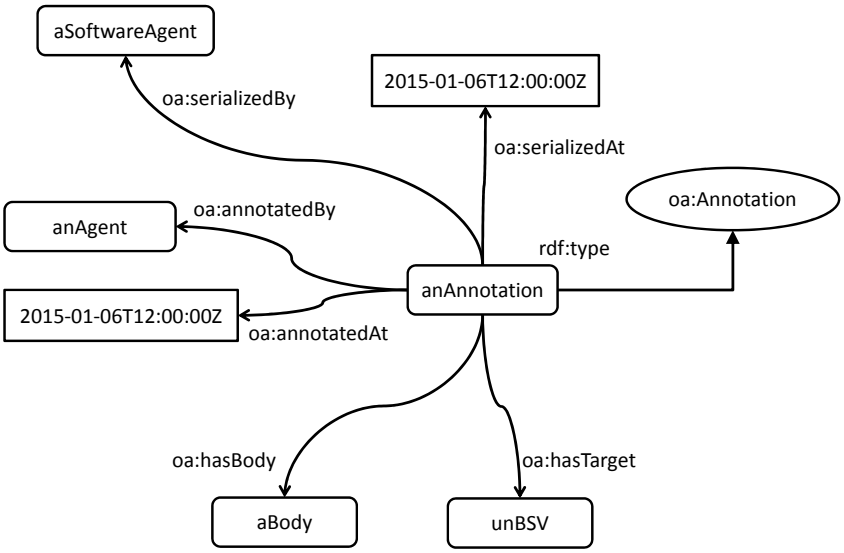

FIGURE 8 – sous-partie de Open Annotation Data Model décrivant la provenance d'une annotation

Nous proposons d'étendre Open Annotation Core pour stocker les index des systèmes de RIS. Cette extension porte le nom de Open Annotation for Indexing (OAI). La figure 9 présente en pointillés les éléments ajoutés à OA et définis par OAI.

Nous définissons d'abord un nouvel objectif d'annotation *oai :indexing*. OA permettant de définir des annotations composites, nous allons définir un nouveau type de tag pondéré représenté par la classe *oai :WeightedTag*. Les tags pondérés sont des éléments d'un individu de type *Composite*. Nous pourrons par exemple utiliser ce schéma d'annotation pour associer non

seulement une région mais aussi les départements de cette région à un BSV. La région et les départements seront les éléments d'une même annotation composite. Le poids affecté à la région et au départements dependra des algorithmes d'indexation.

Concernant les types de culture, nous pourrons, pour chaque type de culture identifié lors du catalogage d'un BSV, associer un sous-ensemble de types de cultures voisines dans le thésaurus des types de cultures. De la même manière que pour les régions, les poids associés aux types de cultures voisines dépendront de l'algorithme d'indexation sémantique.

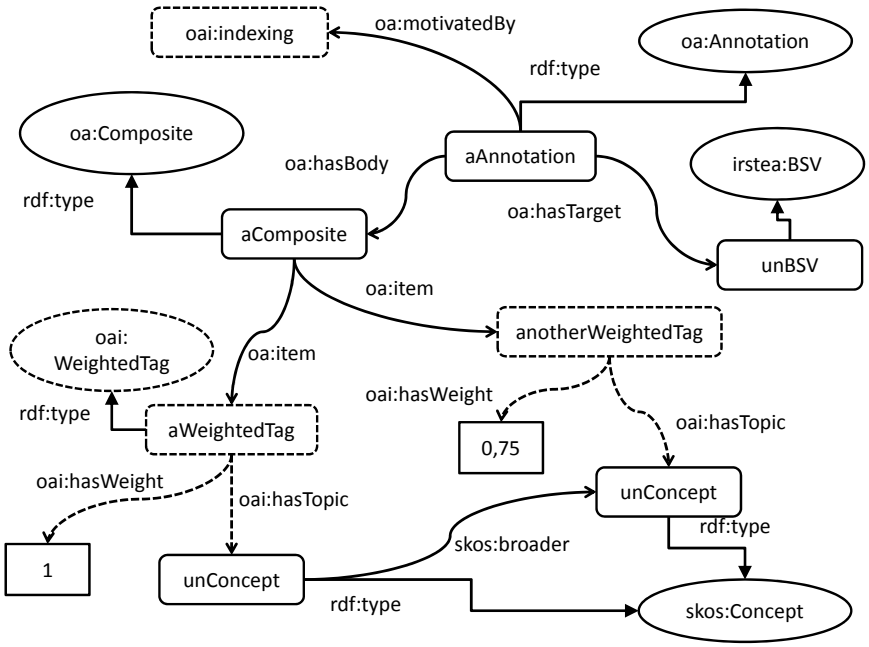

FIGURE 9 – extension de open annotation pour l'indexation

5 Les processus d'annotation

La première méthode d'annotation vise à caractériser tout BSV par au moins sa date de publication, sa région et les types de culture concernés, en utilisant le premier schéma d'annotation.

La description des BSV et leur mise à disposition sur un site web est faite manuellement. Le but est d'extraire semi-automatiquement ces informations, à partir des sites web et des noms de fichiers pdf, pour générer notre premier jeu de données d'annotation. Ces annotations dites "manuelles" sont le résultat de notre travail de moissonnage des BSVs sur le web avec des processus automatiques ou semi-automatiques.

La région est celle de l'administration qui donne accès sur son site web aux BSV. Les sites web que nous avons moissonnés sont en nombre limité. Il nous a donc été possible de récupérer facilement la région associée à un site lors du moissonnage des BSV. Il arrive parfois que les BSV soient le fruit d'une collaboration entre les organismes de deux régions ; deux manifestations distinctes du même BSV existent alors sur les sites web des administrations concernées.

Cette indexation spatiale est faite automatiquement lors de la génération des URI des BSV téléchargés et ne nécessite pas d'intervention humaine.

En ce qui concerne les types de cultures, chaque région publie différentes sortes de bulletins. On trouvera par exemple des BSV sur le colza dans certaines régions, sur les oléagineux dans d'autres, et des BSV sur les grandes cultures dans la plupart des régions de France. Les noms des catégories de BSV ne sont pas normalisés et dépendent des productions principales des régions. En effet, une région peut avoir une catégorie intitulée "petits fruits" alors qu'une autre région l'intitulera "fraises et framboises". L'annotation du type de culture reviendra à associer la catégorie du BSV indiquée sur le site web à au moins une entrée du thésaurus des types de cultures que nous avons défini. Cette indexation thématique est automatisée et se fait à partir d'un patron de transformation construit à la main, qui traduit le nom de catégorie locale en un ensemble de concepts SKOS issus de notre thésaurus.

Obtenir la date de publication n'est pas aussi aisé qu'il n'y parait. Nous avons développé trois processus d'extraction des dates (présentés par ordre de priorité) :

— La date est souvent présente dans le nom du fichier pdf téléchargé. Un premier processus d'extraction à partir des noms de fichiers est réalisé à l'aide de patrons d'extraction de dates typiques.
— La date de création du fichier pdf est aussi présente dans les méta-données du fichier.
— Enfin, nous avons utilisé un processus d'extraction des dates à partir du contenu du fichier pour extraire la date la plus fréquemment rencontrée.

Aucun de ces processus ne permet d'obtenir avec certitude la date de publication du bulletin. Par exemple, certains noms de fichiers ne contiennent pas de date, ou au contraire contiennent une série de chiffres interprétés à tort comme étant une date. Les métadonnées sont parfois illisibles, et il arrive trop fréquemment que le fichier n'ait pas été créé le jour de la publication du BSV (il a été crée la veille, ou corrigé pour être re-créé à une date ultérieure, pas toujours proche). Enfin, le bulletin lui-même contient de nombreuses dates, comme par exemple des dates de relevés ou de mesures, et il est difficile d'identifier avec certitude laquelle correspond à la publication du BSV.

Ces trois processus sont automatiques et nous permettent de sélectionner la date de publication la plus probable selon un algorithme simple : si deux ou trois processus renvoient la même date, c'est celle qui est choisie (73% des cas, soit 11332 BSV sur 15569). Sinon le choix se fera dans l'ordre de priorité décrit ci-dessus (chacun des trois processus pouvant ne retourner aucune date, on sélectionne le premier processus ayant abouti). 0,2% des BSV (c'est-à-dire 37) n'ont pas de date identifiée par ce processus.

L'ordre de priorité a été défini par des statistiques sur les cas où deux dates sur trois sont identiques et par une validation manuelle sur un echantillon de BSV, qu'il conviendra d'étendre pour fiabiliser l'ensemble du processus.

Notre méthode d'annotation dite manuelle effectue une extraction automatique d'informations relatives aux BSV qui ont été publiées par les éditeurs des sites web. Cette méthode va nous permettre de renseigner en partie le schéma d'annotation de catalogage.

Ensuite nous allons pouvoir développer d'autres méthodes d'indexation en utilisant le schéma d'annotation oai. Une méthode serait de transformer et d'enrichir automatiquement les données du schéma de catalogage pour produire des index sémantiques.

Nous espérons par la suite développer une méthode d'indexation capable d'extraire automatiquement des index à partir du contenu des BSV. Cette méthode permettrait de proposer

un second jeux d'index, en particulier pour identifier les agresseurs des cultures et les niveaux de risque. Pour ce faire, nous espérons pouvoir utiliser les sorties du système Vespa Mining Turenne *et al.* (2015).

6 Conclusion

Cet article présente deux schémas d'annotation construits à partir de vocabulaires RDF et d'ontologies. Ces schémas d'annotation ont pour but de faciliter la recherche dans un corpus de bulletins agricoles intitulés Bulletins de Santé du Végétal. Le premier schéma est proche du schéma d'annotation utilisé par la BNF pour le catalogage des documents. Le second schéma basé sur les ontologies Open Annotation Data Model a pour but de stocker les index utilisables par des systèmes de recherche d'information sémantique. Dans des travaux futurs nous devrons valider la mise en œuvre de ces schémas sur une sous-partie du corpus des BSV. Par la suite, nous souhaitons pouvoir combiner et comparer les résultats de différents méthodes d'annotation et d'indexation.

Références

BIBLIOTHÈQUE NATIONALE DE FRANCE (2015). Web sémantique et modèle de données.

BRICKLEY D. & MILLER L. (2014). Foaf vocabulary specification 0.99.

CICCARESE P., OCANA M., GARCIA CASTRO L., DAS S. & CLARK T. (2011). An open annotation ontology for science on web 3.0. *Journal of Biomedical Semantics*, **2**(2).

DCMI USAGE BOARD (2012). DCMI Metadata Terms.

HASLHOFER B., SANDERSON R., SIMON R. & VAN DE SOMPEL H. (2014). Open annotations on multimedia web resources. *Multimedia Tools and Applications*, **70**(2), 847–867.

NEŠIĆ S., CRESTANI F., JAZAYERI M. & GAŠEVIĆ D. (2010). Concept-based semantic annotation, indexing and retrieval of office-like document units. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, p. 134–135 : Centre de hautes études internationales d'Informatique Documentaire (C.I.D).

TURENNE N., ANDRO M., ROSELYNE CORBIÈRE R. & PHAN T. (2015). Open data platform for knowledge access in plant health domain : Vespa mining.

W3C (2009). Skos simple knowledge organization system reference.

W3C (2013). Open annotation data model : Open annotation core.