

# Annotation sémantique pour une interrogation experte des Bulletins de Santé du Végétal

Catherine ROUSSEY<sup>1</sup>, Tayeb ABDERRAHMANI GHORFI<sup>1</sup>

UR TSCF Irstea, Aubière, France  
prenom.nom@irstea.fr

**Résumé** : Le corpus des Bulletins de Santé du Végétal est actuellement disponible sur le Web de données. Les annotations associées à chaque bulletin sont organisées par des propriétés issues du vocabulaire du Dublin Core. Ces annotations répondent à un besoin de recherche documentaire basé sur trois composants : régions françaises, période de publication et types de cultures. Nous avons appliqué une des méthodes de construction d'ontologies de Neon pour faire évoluer le modèle d'annotation des bulletins. Ce modèle contient une ontologie des observations des parcelles et un modèle d'annotation liant le contenu textuel à une entité définie dans l'ontologie précédente. Ces nouveaux modèles répondent à des besoins d'information exprimés par les agronomes.

**Mots-clés** : développement d'ontologies, annotation, Bulletin de Santé du Végétal, competency questions, SPARQL

## 1 Introduction

Lors du projet Vespa, un corpus de bulletins d'alerte agricole intitulé Bulletins de Santé du Végétal (BSV) a été publié sur le Web de données liées (Roussey *et al.*, 2017). Pour ce faire, un modèle d'annotation a été construit en réutilisant des propriétés du vocabulaire du Dublin Core. Ce modèle répond à un besoin de recherche documentaire classique. Un utilisateur va rechercher un ensemble de bulletins en spécifiant une ou plusieurs régions spatiales, une période de publication et un ou plusieurs types de culture.

- Dans ce contexte, une annotation spatiale établit un lien entre un bulletin et une entité géographique. L'entité géographique est définie dans un jeu de données représentant les régions de France avec leurs liens d'inclusion spatiale.
- Une annotation temporelle associe un bulletin à sa date de publication.
- Une annotation thématique établit un lien entre un bulletin et un concept issu d'un thésaurus organisant les types de culture de façon hiérarchique.

Dans le cadre de la recherche documentaire, que nous intitulerons recherche thématique, une annotation établit un lien entre une entité textuelle (le bulletin) et une entité sémantique issue d'une ressource sémantique (le concept d'un thésaurus). La ressource sémantique, d'où est extraite l'entité sémantique, ne possède qu'un seul type d'entité et les organise à l'aide d'une relation hiérarchique représentant soit une inclusion spatiale (entre *ancienne* et *nouvelle* région), soit une relation de généralité entre thèmes (par exemple une culture de céréales est plus générale qu'une culture de blé). Cette ressource est alors intitulée « système d'organisation des connaissances ». Un thésaurus est un bon exemple de ce type de ressource.

À la fin du projet Vespa, de nouveaux besoins ont été exprimés par des agronomes. Ils souhaitent être capables de retrouver les bulletins à partir des observations réalisées sur les parcelles cultivées. Ces observations sont référencées dans les bulletins. Ce nouveau type de recherche, que nous intitulerons recherche experte, implique :

- de proposer un modèle pour décrire ces observations et les parcelles associées,
- d'être capable d'extraire des bulletins les entités représentant ces observations,
- de proposer un modèle d'annotation pour lier le contenu du bulletin aux observations,
- de vérifier que ces nouvelles annotations ne sont pas en contradiction avec les annotations concernant la recherche documentaire thématique. Il serait en effet étonnant de retrouver des observations concernant des parcelles de blé dans un bulletin annoté précédemment avec le concept "vigne".

Ce besoin de recherche experte revient à développer une nouvelle ontologie décrivant les observations et les parcelles. Cette ontologie devra intégrer les systèmes d'organisation des connaissances (le thésaurus des cultures et le jeu de données des régions) utilisés pour la recherche thématique.

Pour ce faire, nous avons choisi d'utiliser une méthode de développement d'ontologie travaillant avec des « *competency questions* » (CQ). Les *competency questions* sont des questions en langue naturelle utilisées pour spécifier les besoins lors de la construction d'une ontologie à partir de zéro (Suárez-Figueroa *et al.*, 2009). Ces questions sont ensuite traduites en langage formel pour construire l'ontologie (Grüninger & Fox, 1995). Plusieurs méthodes utilisent les CQ pour construire une ontologie. Nous pouvons entre autres citer la méthode *Extrem Design*, utilisée pour construire des patrons de conception ontologique (Presutti *et al.*, 2009). Dans notre cas nous avons utilisé les CQ pour développer l'ontologie décrivant les observations et modifier le modèle d'annotation des BSV existant. Nous avons appliqué la méthode de conception d'ontologie proposée dans le scénario 3 de la méthodologie Neon (Suárez-Figueroa *et al.*, 2012).

Cet article est structuré de la manière suivante : La première section présente le corpus annoté et son modèle d'annotation existant. Ensuite, le nouveau besoin d'informations et les CQ associées sont décrits dans la section 3. La nouvelle ontologie et le modèle d'annotation associé sont ensuite décrits dans la section 4. La validation de l'ontologie est décrite dans la section 5. Puis nous concluons avec les perspectives qui émergent de nos travaux.

## 2 Contexte

Nous présentons dans cette section le corpus des Bulletins de Santé du Végétal ainsi que le modèle d'annotation existant pour la recherche de bulletins. Le corpus est actuellement publié sur le Web de données à l'adresse <http://ontology.irstea.fr/bsv/snorql/>

### 2.1 Bulletin de Santé du Végétal

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 dans l'ensemble des régions et départements d'outre-mer. Le Bulletin de Santé du Végétal (BSV) est un document d'information à la fois technique et réglementaire, rédigé sous la responsabilité d'un comité régional d'épidémiologie. Le BSV a pour objectif de réunir et présenter les actualités majeures concernant l'état sanitaire des cultures. Il repose d'un côté sur des analyses du risque phytosanitaire à venir et d'un autre sur la diffusion des informations à caractère réglementaire (arrêtés de lutte obligatoire, notes nationales, évolutions de la réglementation, ...) et non réglementaire (éléments de description de la biologie des bioagresseurs ou des méthodes prophylactiques telles que la gestion des intercultures, du travail du sol, du choix des variétés, ...). Afin de mieux distinguer l'expertise de la préconisation, il n'a pas vocation à faire des préconisations d'utilisation de produits phytosanitaires. La figure 1 présente un exemple de première page d'un BSV de la région Bourgogne.

Les BSV sont une synthèse interprétée des observations effectuées en amont sur les cultures par différents organismes collecteurs, des éléments issus des modèles épidémiologiques, de données météorologiques et parfois d'analyse biologique. 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 potentiellement importante et suffisamment représentatif pour être signalé. Étant donné que de nombreux problèmes sanitaires sont d'autant mieux 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 des jugements sur des observations.

Les BSV sont gratuitement accessibles au format pdf sur les sites internet des Chambres Régionales d'Agriculture et des Directions Régionales de l'Alimentation de l'Agriculture et



FIGURE 1 – Un exemple de BSV de la région Bourgogne, rubrique "grande culture", daté du 5 avril 2011

de la Forêt (DRAAF).

Les BSV sont tout d'abord lus par les conseillers agricoles des coopératives et les agriculteurs pour déterminer leurs futures actions sur leurs cultures ou évaluer l'état de leurs cultures par rapport à l'état des parcelles du réseau. Mais ce corpus intéresse aussi les chercheurs en agronomie. Il constitue une archive des événements sanitaires importants perçus sur les cultures au cours du temps.

## 2.2 Modèle d'annotation existant des BSV

Le modèle d'annotation utilisé lors du projet Vespa pour publier les BSV sur le Web est inspiré du modèle d'annotation de la Bibliothèque Nationale de France (BNF) (Lapôtre, 2017). Ce modèle réutilise le vocabulaire du Dublin Core (dcterms) (Weibel *et al.*, 1998). Le Dublin Core propose un ensemble de propriétés pour enregistrer les métadonnées d'une ressource (titre, auteur, format, etc.). Le modèle de la BNF dissocie l'œuvre (Les misérables, de Victor Hugo), de son expression (une édition en 10 volumes publiée en 1862), et de sa manifestation physique (le scan de cette édition disponible sur Gallica). Il nous a paru intéressant de conserver la distinction entre l'expression et la manifestation physique, car un même bulletin peut être disponible sur plusieurs sites Web. Pour ce faire, nous avons repris la notion d'objet d'informations proposé dans l'ontologie fondationnelle *Dolce Ultra Light* (DUL)<sup>1</sup>. Cette ontologie est souvent utilisée pour définir des patrons de conception ontologique.

Avant de présenter en détail le modèle d'annotation correspondant à une recherche thématique, nous allons tout d'abord présenter les deux systèmes d'organisation des connaissances que nous avons utilisés.

1. <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl>

- Un thésaurus intitulé FrenchCropUsage<sup>2</sup> a été défini pendant le projet Vespa pour organiser les cultures en fonction de leur destination (alimentation humaine directe, alimentation animale, industrie alimentaire, etc) et du type de système de culture (grande culture, maraîchage, etc...). Ce thésaurus représente le point de vue français. Il est basé sur les définitions du Larousse Agricole et du wikipedia agricole français. Il contient 272 concepts, la profondeur maximale de la hiérarchie est de 6 niveaux. Ce thésaurus est publié sur le Web de données à l'aide du modèle SKOS à l'adresse <http://ontology.irstea.fr/cropusage/>.
- Pour identifier les régions de France nous avons dû aussi créer un jeu de données propre au projet Vespa, car à la date du projet aucune source ne décrivait les anciennes et les nouvelles régions de France. Dans ce jeu de données, chaque région est une instance d'une classe `irstea:Region`. Ces instances sont liées à des instances similaires issues des jeux de l'IGN, de l'INSEE ou de DBpedia.

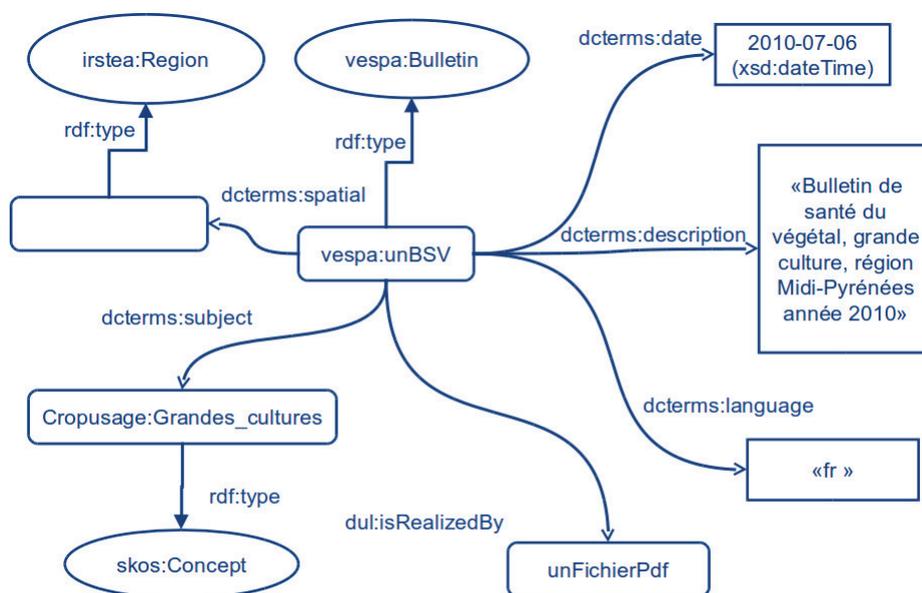


FIGURE 2 – une exemple d'utilisation du modèle d'annotation décrivant le BSV Grande Culture de la région Midi-Pyrénées du 6 juillet 2010

La figure 2 représente un exemple d'instanciation du modèle d'annotation thématique des BSV. Les entités que nous avons créées spécifiquement pour l'annotation des BSV sont préfixées par vespa.

Un bulletin est représenté par une instance de la classe `vespa:Bulletin`. Cette classe est une sous-classe de `dul:InformationObject`. Un objet d'information est une entité abstraite qui regroupe l'ensemble des informations relatives à un objet indépendamment de sa matérialisation. Par exemple, l'œuvre "les Misérables" de Victor Hugo est un objet d'information. Les informations de cet objet sont indépendantes d'un exemplaire en particulier. Nous retrouvons cette notion dans la classe `Oeuvre` du modèle d'annotation de la BNF. 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 `dul:isRealizedBy`. Les annotations sont portées par l'instance de

2. DOI : 10.25504/FAIRsharing.9228fv et accessible à partir d'agroportal <http://agroportal.lirmm.fr/ontologies/CROPUSAGE>

la classe `vespa:Bulletin`. Les propriétés utilisées pour décrire les BSV sont :

**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.

**dcterms:description** : contient une description textuelle du BSV. Cette description correspond aux informations que nous avons pu extraire des sites Web où le bulletin a été téléchargé : la région correspondant au site Web, l'année de publication et un type de culture qui correspond aux rubriques du site Web où le bulletin apparaît.

**dul: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. Dans l'exemple de la figure 2 ce nœud est intitulé `irstea:places/73` et représente l'ancienne région Midi-Pyrénées.

**dcterms:subject** est le lien vers le `skos:Concept` du thésaurus `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`).

### 3 Besoins d'informations

Le projet Vespa pour "Valeur et optimisation des dispositifs d'épidémiosurveillance dans une stratégie durable de protection des cultures" avait pour but d'étudier les différentes formes de contribution de l'épidémiosurveillance à la santé des cultures. Lors de ce projet, nous avons construit l'archive des BSV publiée sur le Web présentée dans la section précédente (Roussey *et al.*, 2017). La surveillance des cultures consiste à répertorier l'apparition de bioagresseurs sur les cultures au cours d'une année culturale (de septembre à août). Les bioagresseurs sont les ennemis des cultures, aussi appelés nuisibles des cultures. Ce sont des organismes vivants qui attaquent les plantes cultivées et sont susceptibles de causer des pertes économiques. Ils ont différentes formes :

- Les ravageurs des cultures sont des animaux qui attaquent les plantes cultivées ou les récoltes stockées, en causant un préjudice économique aux agriculteurs.
- Les maladies des plantes empêchent le développement correct des plantes. Elles ont plusieurs causes. Elles peuvent être dues à des champignons parasitaires microscopiques, des bactéries transmises par des insectes suceurs de sève ou par des nématodes dans le sol.
- Les adventices sont les "mauvaises" herbes qui concurrencent les plantes cultivées.

Évaluer le niveau de dangerosité d'un bioagresseur ne se limite pas à observer sa présence sur une parcelle cultivée. En effet, il faut que la plante cultivée ait atteint un certain stade de développement pour que le bioagresseur ait un impact sur la production finale de la parcelle. Il faut aussi parfois que les bioagresseurs soient suffisamment nombreux sur une parcelle pour que les plantes cultivées soient gênées dans leur développement. Lorsqu'un bioagresseur est observé sur une parcelle il faut donc aussi évaluer le niveau de sévérité (sa dangerosité) pour considérer sa présence comme étant une attaque.

Suite au développement de l'archive des BSV et des outils de recherche d'informations mis en place pour interroger ce corpus, les agronomes ont exprimé de nouveaux besoins plus évolués que ceux identifiés au départ du projet. À partir des multiples discussions avec des chercheurs en agronomie qui ont eu lieu durant le projet Vespa, nous avons construit, à la fin du projet, une série de CQ répondant à leurs nouveaux besoins sur le corpus des BSV.

1. Quelles sont les cultures observées en France ?
2. Quelles sont les cultures observées dans la région R (AURA par exemple) ?
3. Quel est l'échantillon de parcelles observées de la culture C dans la région R pendant la période P ?
4. Quels sont les bioagresseurs connus de la culture C (Maïs par exemple) ?

5. Quels sont les ravageurs connus de la culture C ?
6. Quels sont les maladies connues de la culture C ?
7. Quels sont les adventices connues de la culture C ?
8. Quels sont les bioagresseurs connus de la culture C dans la région R ?
9. Quelles sont les attaques du bioagresseur B survenues sur des parcelles de la culture C dans la région R pendant la période P ?
10. Quels sont les stades de développement atteints par la culture C dans les parcelles cultivées de la région R pendant la période P ?
11. Quelles sont les attaques du bioagresseur B survenues sur les parcelles de la culture C dans la région R qui ont atteint un niveau de sévérité S pendant la période P ?
12. Quelle est la chronologie et l'intensité des attaques du bioagresseur B sur les parcelles de culture C dans toutes les régions de France ? (carte de France)
13. Quelles sont les parties de texte dans un ensemble de bulletins qui portent sur des attaques du bioagresseur B sur la culture C ?
14. Quelles sont les parties de texte dans un ensemble de bulletins qui portent sur les stades de développement de la culture C ?
15. Quelles sont les parties de texte dans un ensemble de bulletins qui portent sur les échantillons des parcelles observées de la culture C ?

#### 4 Une ontologie d'observation des parcelles et le modèle d'annotation associé

À partir de ces CQ nous voulons modéliser l'ensemble des informations demandées. Il est clair que nous avons besoin d'une ontologie d'observation en environnement naturel. Il existe de notre point de vue deux grandes familles d'observations, les observations correspondant à des mesures automatiques de capteurs (comme les stations météo) et des observations humaines. Plusieurs ontologies correspondent à la première famille car c'est un sujet largement travaillé. Nous pouvons entre autres citer *Semantic Sensor Network (SSN)*(Compton *et al.*, 2012), *Smart Appliances REFERENCE For Environment (SAREF4ENVI)*(ETSI, 2017), *Observations and Measurements (OM)*(Cox, 2011) pour les plus connues. Concernant la seconde famille, nous ne connaissons que *Extensible Observation Ontology (OBOE)* (Madin *et al.*, 2007) dédiée aux observations scientifiques environnementales. À noter que les ontologies dédiées à la description des expérimentations ne font pas partie du périmètre de nos besoins.

Les observations des parcelles sont réalisées par des humains dans le cadre des BSV, mais il est tout à fait envisageable dans un futur proche d'imaginer que ces observations soient automatisées et puissent être réalisées par des équipements de mesures spécifiques. C'est pour cette raison que nous avons sélectionné l'ontologie SSN. Pour SSN, un capteur désigne toute entité capable de suivre une méthode d'observation, que ce soit une personne ou un équipement de mesure<sup>3</sup>.

SSN est développée sous l'égide du *World Wide Web Consortium (W3C)*. Pour compléter l'ontologie SSN, nous avons sélectionné les ontologies proposées par le W3C en préférant celles qui ont atteint le statut de recommandation. Notre objectif est donc de sélectionner un ensemble d'ontologies du W3C qui répondent aux besoins exprimés par les CQ. Le tableau 1 présente la liste des ontologies sélectionnées.

---

3. <https://www.w3.org/2005/Incubator/ssn/ssnx/ssn#Sensor>

| Nom  | Acronyme | Auteur       | Référence                        |
|--|----------|--------------|----------------------------------|
| Sensor Observation Sampler Actuator <sup>4</sup>   | ssn/sosa | W3C OGC      | (Armin <i>et al.</i> , 2018)     |
| Semantic Sensor Network <sup>5</sup>               | ssn      | W3C          | (Compton <i>et al.</i> , 2012)   |
| Time <sup>6</sup>                                  | time     | W3C          | (Hobbs & Pan, 2006)              |
| GeoSparql <sup>7</sup>                             | geo      | OGC          | (Battle & Kolas, 2012)           |
| Prov Ontology <sup>8</sup>                         | prov     | W3C          | (Lebo <i>et al.</i> , 2013)      |
| Event <sup>9</sup>                                 | event    | C4DM at QMUL | (Raimond & Abdallah, 2007)       |
| Web Annotation Data Model <sup>10</sup>            | oa       | W3C          | (Sanderson <i>et al.</i> , 2013) |
| Simple Knowledge Organisation System <sup>11</sup> | skos     | W3C          |                                  |

TABLE 1 – Liste des ontologies réutilisées

L'ontologie SSN a évolué (Compton *et al.*, 2012). En 2017, une nouvelle version de cette ontologie construite cette fois-ci sous l'égide du W3C et de l'*Open Geospatial Consortium* (OGC) a été acceptée comme recommandation (Armin *et al.*, 2018). Cette version intègre un nouveau patron de conception intitulé *Sensor Observation Sampler Actuator* (SOSA). SSN/SOSA recommande d'utiliser l'ontologie *GeoSparql* (Battle & Kolas, 2012) pour décrire les entités spatiales. A noter que pour le W3C, l'ontologie *Time* (Hobbs & Pan, 2006) est suffisante pour décrire un évènement (une entité localisée dans le temps). Dans le cas d'une alerte agricole, nous avons besoin d'un modèle simple pour décrire un évènement comme une entité spatialisée et temporalisée impliquant des agents. Nous avons sélectionné l'ontologie *Event* (Raimond & Abdallah, 2007) pour sa simplicité et sa couverture de l'ensemble de nos besoins. Mais d'autres ontologies étaient possibles comme l'ontologie *Linking Open Descriptions of Events* (LODE) (Shaw *et al.*, 2009).

Les sections suivantes présentent les modèles permettant de décrire les observations faites sur les cultures, les alertes agricoles et les liens vers les textes des BSV. A ce stade, ces modèles sont en cours de discussion et ne sont pas formalisés en RDF.

#### 4.1 Modèle d'observation des stades de développement des cultures

La figure 3 présente un exemple de description d'une observation du stade de développement atteint par un échantillon de parcelles à l'aide des ontologies SSN/SOSA, GeoSparql, Time, SKOS et QUDT. En plus des ces ontologies, nous réutilisons le thésaurus FrenchCropUsage qui définit des types de cultures à l'aide du modèle SKOS. Donc un type de culture est une instance de la classe `skos:Concept`.

Un nouveau système d'organisation des connaissances est nécessaire pour décrire les stades de développement des cultures. Nous avons sélectionné le référentiel BBCH qui est actuellement décrit au sein d'un jeu de données de l'ontologie CROP<sup>12</sup> disponible sur l'Agroportal du LIRMM. Dans notre exemple, un stade de développement est défini comme une instance de la classe `skos:Concept`.

Nous avons besoin aussi de décrire les unités. SSN préconise plusieurs jeux de données et leurs ontologies associées disponibles sur le Web de données : "*Quantities, Units, Dimensions and data Types*" (QUDT) (Hodgson *et al.*, 2014), "*Ontology of units of Measurements*"

4. <https://www.w3.org/TR/vocab-ssn/>

5. <https://www.w3.org/2005/Incubator/ssn/ssnx/ssn>

6. <https://www.w3.org/TR/owl-time/>

7. <http://www.opengeospatial.org/standards/geosparql>

8. <https://www.w3.org/TR/prov-o/>

9. <http://motools.sourceforge.net/event/event.html>

10. <https://www.w3.org/TR/annotation-model/>

11. <https://www.w3.org/TR/skos-reference/>

12. <http://www.cropontology.org/>

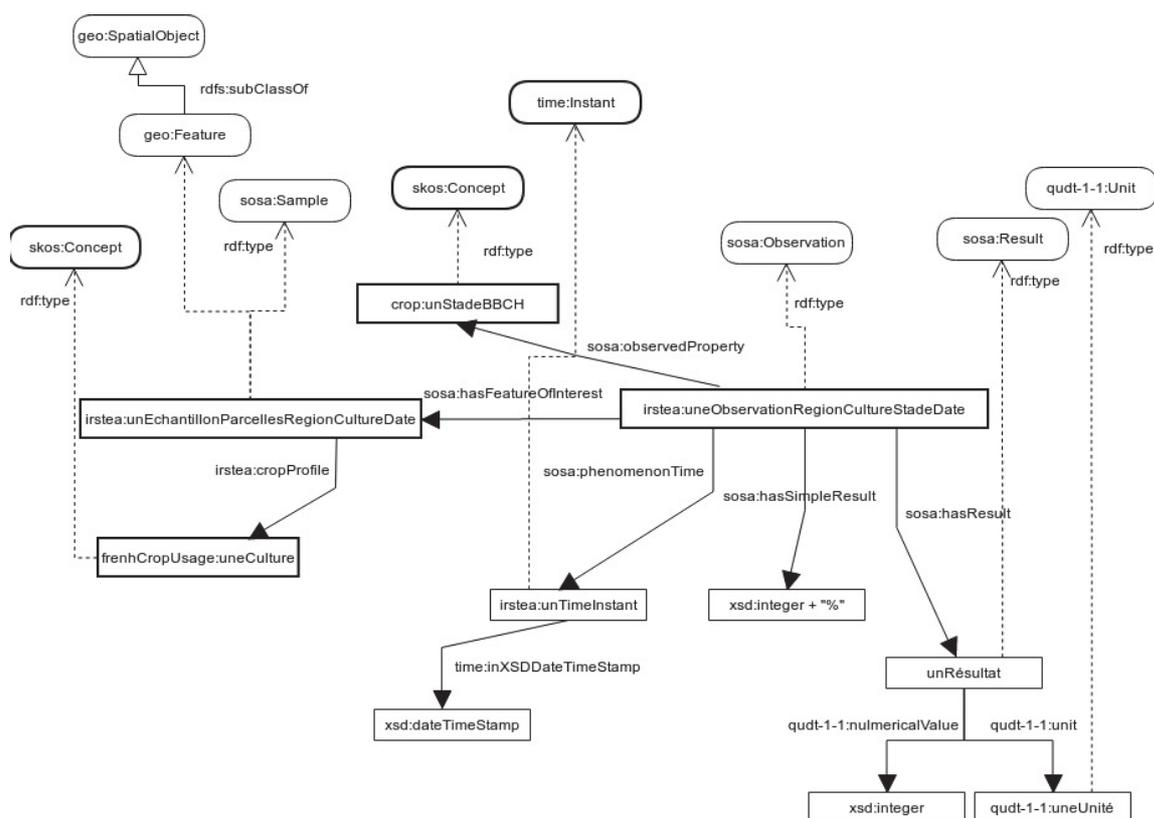


FIGURE 3 – Exemple d'observation de stades de développement

(OM) (Rijgersberg *et al.*, 2013) et "Unified Code for Units of Measure" (UCUM), (Lefrançois & Zimmermann, 2018). Dans notre exemple, nous utilisons la dernière version de l'ontologie QUDT. Ainsi, nous identifions une unité donnée comme une instance de la classe `qudt-1-1:Unit`.

Comme le montre la figure 3, une observation d'un stade de développement d'une culture est identifiée par une instance de la classe `sosa:Observation`. Cette observation porte sur un échantillon de parcelles. Un échantillon de parcelles est un objet géographique. Donc un échantillon est défini comme une instance de `sosa:Sample` et de `geo:Feature`.

Les propriétés utilisées pour décrire l'observation sont :

**sosa:hasFeatureOfInterest** lie une instance de `sosa:Observation` à l'instance représentant l'échantillon de parcelles observées.

**irstea:cropProfile** lie l'instance représentant l'échantillon de parcelles à une instance de `skos:Concept` représentant le type de culture cultivé sur ces parcelles extrait du thésaurus `FrenchCropUsage`.

**sosa:observedProperty** lie une instance de `sosa:Observation` à une instance de `skos:Concept` représentant le stade de développement observé, extrait du référentiel BBCH décrit dans l'ontologie CROP.

**sosa:phenomenonTime** lie une instance de `sosa:Observation` à une instance de `time:Instant` indiquant la date où cette observation a eu lieu. Dans notre cas il s'agit de la date de publication du BSV. Ainsi, la valeur stockée dans la propriété `time:inXSDDateTimeStamp` doit être la même que celle de `dcterms:date`.

**sosa:hasSimpleResult** est un attribut qui contient un pourcentage. Ce pourcentage indique le ratio entre le nombre de parcelles qui a atteint le stade de développement sur le nombre de parcelles totales composant l'échantillon.

**sosa:hasResult** lie une instance de `sosa:Observation` à une instance de `sosa:Result`. Cette instance possède deux attributs : l'un indique l'unité, l'autre la valeur.

#### 4.2 Modèle d'observation d'un bioagresseur dans une culture

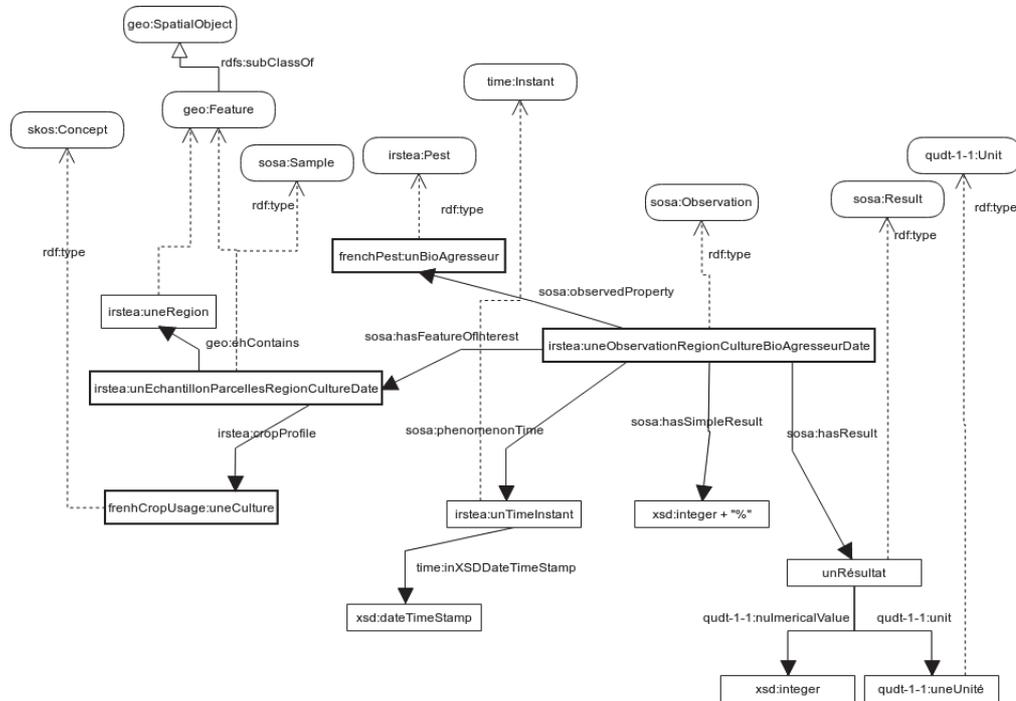


FIGURE 4 – Exemple d'observation de la présence d'un bioagresseur

La figure 4 présente une description de l'observation de la présence d'un bioagresseur sur un échantillon de parcelles à l'aide des ontologies : SSN/SOSA, Time, GeoSparql, SKOS et QUDT. Nous retrouvons dans cet exemple les références au thésaurus FrenchCropUsage et au jeu de données des unités.

Comme élément nouveau, nous avons indiqué une référence vers le jeu de données décrivant les régions de France. Une région est maintenant une instance de `geo:Feature`. Nous devons donc faire évoluer le jeu de données existant présenté dans la section précédente (cf figure 2) pour répondre à cette spécification.

Un nouveau jeu de données doit être référencé pour décrire les bioagresseurs des cultures. À ce stade nous n'avons pas encore trouvé de jeu de données préalablement publié sur le Web de données répondant à ce besoin. Il est possible que ce jeu puisse être modélisé en SKOS. Afin de ne pas contraindre la modélisation nous avons représenté un bioagresseur comme une instance de la classe `irstea:Pest` appartenant à un jeu de données intitulé FrenchPest.

Une observation de la présence d'un bioagresseur se représente comme une instance de la classe `sosa:Observation`. Les propriétés utilisées pour décrire cette observation sont :

**sosa:hasFeatureOfInterest** lie une instance de `sosa:Observation` à l'instance représentant l'échantillon de parcelles observées.

**irstea:cropProfile** lie l'instance représentant l'échantillon de parcelles à une instance de `skos:Concept` représentant le type de culture cultivé sur ces parcelles.

**geo:ehContains** lie l'instance représentant l'échantillon de parcelles à une instance de `geo:Feature` représentant la région.



**event :agent** lie une instance de `irstea:PestAlert` à une instance de `Pest` représentant le bioagresseur observé.

**event :time** lie une instance de `irstea:PestAlert` à une instance de `time:Instant` pour indiquer la date de l'observation. Cette instance est issue des modèles précédents sur les observations des parcelles.

**event :factor** est un attribut booléen qui indique si le seuil de nuisibilité est atteint.

**prov :wasDerivedFrom** lie une instance de `irstea:PestAlert` aux instances d'observation des stades de développement et de la présence des agresseurs qui ont permis de lancer cette alerte.

#### 4.4 Modèle d'annotation des bulletins

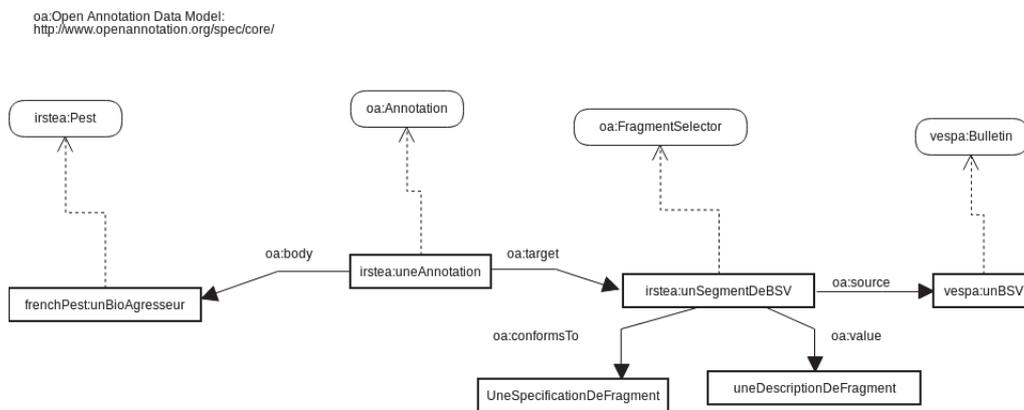


FIGURE 6 – Exemple d'annotation des BSV

Les figures précédentes décrivent les observations faites sur les parcelles cultivées. La figure 6 décrit le modèle permettant de lier les observations précédentes au texte des BSV. Pour ce faire nous réutilisons l'ontologie "Web Annotation Data Model" ou OA du W3C. Les propriétés utilisées pour décrire une instance de `oa:Annotation` sont :

**oa :body** lie une instance de `oa:Annotation` à une instance représentant le sens de l'annotation. Dans la figure 6 il s'agit d'une instance représentant un bioagresseur donné.

**oa :target** lie une instance de `oa:Annotation` à une instance de fragment de texte.

**oa :conformsTo** lie une instance de fragment de texte à un type de sélecteur, le sélecteur étant une fonction qui identifie un fragment de texte. Un fragment de texte peut être identifié par sa position, la chaîne de caractères, etc.

**oa :value** est un attribut qui contient les paramètres du sélecteur pour identifier le fragment.

**oa :source** lie une instance de fragment de texte à son document source. Dans notre cas il s'agit d'une instance de `vespa:Bulletin`.

## 5 Validation

Pour vérifier que nos modèles couvrent bien l'ensemble des besoins exprimés dans les CQ, nous avons demandé à ce qu'un expert en Web sémantique, qui n'a pas participé à la construction des modèles, traduise chacune des CQ en requêtes SPARQL. Cet expert joue le rôle de valideur.

L'ensemble des CQ ont été traduites en requêtes SPARQL. Ces requêtes sont visibles sur le site [ontology.irstea.fr](http://ontology.irstea.fr) à l'adresse <http://ontology.irstea.fr/pmwiki.php/Site/BSVCompetencyQuestions>.

Par exemple, la première CQ sur les cultures observées en France se traduit par la requête :

```
SELECT DISTINCT ?culture
WHERE {
  ?observation a sosa:Observation ;
              sosa:hasFeatureOfInterest ?echantillon .
  ?echantillon irstea:cropProfile ?culture.
}
```

La CQ numéro 9 sur les attaques des bioagresseurs se traduit par :

```
SELECT DISTINCT ?observation
WHERE { ?observation a sosa:Observation ;
                  sosa:hasFeatureOfInterest ?echantillon ;
                  sosa:observedProperty frenchPest:B;
                  sosa:phenomenonTime ?time.
  ?time time:inXSDDateTimeStamp ?stamp.
  ?echantillon geo:ehContains irstea:R ;
              irstea:cropProfile frenchCropUsage:C.
FILTER ((?stamp > "t1"^^xsd:dateTimeStamp) &&
(?stamp < "t2"^^xsd:dateTimeStamp))
}
```

Le fait d'avoir traduit l'intégralité des CQ laisse supposer que l'ensemble des besoins est couvert par le nouveau modèle d'annotation et les ontologies associées. Une deuxième phase de validation a été réalisée par le concepteur de l'ontologie. Quelques divergences sont apparues entre le concepteur et le valideur lors de ces phases de validation sur les CQ 13 et 14.

### 5.1 Problème de la CQ 13

La CQ numéro 13 porte sur les annotations des attaques des bioagresseurs et elle a été traduite en SPARQL ainsi :

```
SELECT DISTINCT ?text
WHERE { ?observation sosa:hasFeatureOfInterest ?echantillon ;
                  sosa:observedProperty frenchPest:B.
  ?echantillon irstea:cropProfile frenchCropUsage:C.
  ?annotation oa:body frenchPest:B;
              oa:target ?segment.
  ?segment oa:value ?text.
}
```

Le concepteur espérait que l'annotation lierait l'observation de la présence d'un bioagresseur au texte des BSV. Le valideur a lié le texte à l'instance de bioagresseur. Le concepteur et le valideur ont donc deux interprétations différentes de la même CQ. Chacune de ces interprétations va donner naissance à une nouvelle CQ et requête SPARQL associée. Ainsi, l'ontologie finale formalisera chacune de ces deux interprétations.

À noter que la différence d'interprétation de la CQ 13 entre le concepteur et le valideur nous montre les limites des modèles. Un modèle (définition de classes et de propriétés) n'est

pas suffisant pour décrire comment utiliser ces classes et ces propriétés. Chacun peut travailler avec les mêmes entités (classe et propriété) mais organisées (liées) de manière différente. Ce qui signifie qu'il faut maintenant clarifier, définir et documenter des "patrons d'usage" pour améliorer la cohérence de jeux de données instanciant les modèles, afin de toujours exprimer la même chose de la même manière.

## 5.2 Problème de la CQ 14

La CQ numéro 14 porte sur les annotations des stades de développement. La première requête SPARQL proposée était la suivante :

```
SELECT DISTINCT ?text
WHERE { ?observation sosa:hasFeatureOfInterest ?echantillon ;
        sosa:observedProperty ?stage.
        ?stage a skos:concept.
        ?echantillon irstea:cropProfile frenchCropUsage:C.
        ?annotation oa:body ?stage;
                   oa:target ?segment.
        ?segment oa:value ?text.
}
```

Malheureusement, rien dans cette requête n'indique que ce qui est lié au texte est un stade de développement. Il faudrait indiquer le source d'où est extrait le concept représentant le bioagresseur. Une autre solution proposée est de spécialiser la classe observation pour créer une classe "Observation de stade de développement" et une autre classe "Observation de la présence d'un agresseur".

## 6 Travaux connexes

Il existe des modèles pour stocker des observations agricoles. Nous pouvons entre autre citer le modèle *Agricultural Model Intercomparison and Improvement Project* (AGmip) (Porter *et al.*, 2014) basé sur le modèle americano-canadien ICASA. Ces modèles permettent de stocker et d'échanger les résultats des modèles de simulation du développement des cultures. Ce sont des modèles exprimés sous forme tabulaire. Plusieurs ontologies du domaine agricole sont disponibles sur le Web. La plus ancienne est agroRDF, développée par KTBL (Martini *et al.*, 2013). Cette ontologie est la traduction d'un schema XML agroXML. Elle est monolithique, elle n'intègre pas d'ontologies existantes et elle est faiblement documentée. La plus récente est l'ontologie FOODIE issue du projet européen du même nom (Palma *et al.*, 2016). Elle a pour but de faciliter l'intégration de données issues de fournisseurs différents dans le domaine de l'agriculture de précision. A noter que cette ontologie est en fait la traduction d'un modèle de base de données relationnelle. Par conséquent, les attributs de chaque classe donnent naissance a une nouvelle propriété. Elle ne répond donc pas au principe de modélisation RDFS, où une propriété est une "first class citizen". Autrement dit, une propriété n'appartient à aucune classe. Elle peut s'appliquer à plusieurs classes. Pour limiter la portée d'une propriété il faut définir une contrainte sur cette propriété.

Dans notre cas, nous avons favorisé la réutilisation d'ontologies reconnues sur le Web de données liées pour construire notre nouveau modèle d'annotation des BSV. Notre modèle est devenu un réseau d'ontologies incluant entre autre une ontologie pour les observations, une ontologie pour les événements et le modèle d'annotation du W3C. Un état de l'art de 2006 (Uren *et al.*, 2006) décrivant les fonctionnalités des outils de gestion de connaissances indiquait que peu d'outils d'annotation sont capables de faire évoluer leur modèle d'annotation. A notre connaissance ce constat est toujours d'actualité car faire évoluer le modèle d'annotation implique qu'il faut être capable de faire évoluer les annotations associées. L'évolution des annotations constitue un axe de recherche à part entière (Cardoso *et al.*, 2017).

Les outils d'annotation font l'hypothèse principale qu'une seule ontologie est utilisée pour structurer les annotations. De plus la plupart de ces outils répondent à un besoin de recherche documentaire thématique. Donc les entités sémantiques utilisées dans les annotations sont organisées dans un système d'organisation des connaissances (un thésaurus) et non dans un jeu de données structuré par une autre ontologie. L'évolution des outils d'annotation est de travailler avec plusieurs systèmes d'organisation des connaissances : un par type d'entités de la recherche thématique. Nous pouvons par exemple citer la plateforme KIM (Popov *et al.*, 2004) qui permet de mettre en place plusieurs chaînes de traitements GATE pour reconnaître des entités sémantiques issues de lexiques différents. Cette plateforme utilise une seule ontologie de haut niveau intitulé KIM. Le projet Parménides a mis en place une autre chaîne de traitements GATE pour extraire des événements structurés par une ontologie dédiée et donc lier des entités sémantiques issues de lexiques différents (Hogenboom *et al.*, 2010).

Dans nos travaux, nous allons développer et instancier un nouveau modèle d'annotation distinct du modèle existant car répondant à des besoins différents. Donc le future modèle ne remplacera pas le modèle existant d'annotation. Par compte, nous devons vérifier la cohérence des annotations instanciant les deux modèles.

## 7 Conclusion et Perspectives

Le corpus des Bulletins de Santé du Végétal est actuellement disponible sur le Web de données à partir d'un modèle d'annotations basé sur le vocabulaire dublin core. Le modèle existant répond à un besoin de recherche documentaire basé sur trois composants : région spatiale, date de publication et type de cultures.

Nous avons appliqué une méthode de construction d'ontologies de Neon pour faire évoluer le modèle d'annotation des bulletins et définir un modèle d'observation des parcelles. Cette méthode exprime les besoins à partir de competency questions et réutilise des ontologies existantes. Ce nouveau réseau d'ontologies modélise des informations détaillées demandées par les agronomes à la fin du projet Vespa.

Nos travaux futurs porteront tout d'abord sur l'implémentation des modèles présentés dans cet article. Puis, nous peuplerons ces modèles à l'aide des sorties d'une chaîne d'outils de traitement de la langue appliquée sur les BSV. Une fois validés, les résultats devront être publiés sur le Web des données pour compléter la description des BSV existants. Nous devons aussi travailler sur la cohérence entre les différents modèles et jeux de données associés (l'existant et le futur). Par exemple, nous pourrions vérifier que les échantillons de parcelles culturelles explicités dans le modèle d'annotation des observations sont bien inclus dans la région explicitée dans le modèle d'annotation documentaire existant. D'autres types de vérification plus complexes pourront être exploités.

## Références

- ARMIN H., KRZYSZTOF J., COX S., LE PHUOC D., TAYLOR K. & LEFRANÇOIS M. (2018). Semantic Sensor Network Ontology.
- BATTLE R. & KOLAS D. (2012). Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web*, 3(4), 355–370.
- CARDOSO S. D., REYNAUD-DELAÎTRE C., DA SILVEIRA M. & PRUSKI C. (2017). Combining rules, background knowledge and change patterns to maintain semantic annotations. In *AMIA 2017*.
- COMPTON M., BARNAGHI P., BERMUDEZ L., GARCÍA-CASTRO R., CORCHO O., COX S., GRAYBEAL J., HAUSWIRTH M., HENSON C., HERZOG A. *et al.* (2012). The ssn ontology of the w3c semantic sensor network incubator group. *Web semantics : science, services and agents on the World Wide Web*, 17, 25–32.
- COX S. (2011). Observations and measurements-xml implementation.
- ETSI (2017). *ETSI TS 103 410-2 - v1.1.1. SmartM2M; Smart Appliances Extension to SAREF; Part2 : Environment Domain*. Rapport interne, ETSI.
- GRÜNINGER M. & FOX M. S. (1995). Methodology for the design and evaluation of ontologies.
- HOBBS J. R. & PAN F. (2006). Time ontology in owl. *W3C working draft*, 27, 133.
- HODGSON R., KELLER P. J., HODGES J. & SPIVAK J. (2014). Qudt-quantities, units, dimensions and data types ontologies. *USA*, Available from : <http://qudt.org> [March 2014].
- HOGENBOOM F., HOGENBOOM A., FRASINCAR F., KAYMAK U., VAN DER MEER O., SCHOUTEN K. & VANDIC D. (2010). SPEED : A Semantics-Based Pipeline for Economic Event Detection. In J. PARSONS, M. SAEKI, P. SHOVAL, C. WOO & Y. WAND, Eds., *Conceptual Modeling – ER 2010*, p. 452–457, Berlin, Heidelberg : Springer Berlin Heidelberg.
- LAPÔTRE R. (2017). Library metadata on the web : the example of data. bnf. fr. *JLIS. it*, 8(3), 58.
- LEBO T., SAHOO S., MCGUINNESS D., BELHAJJAME K., CHENEY J., CORSAR D., GARIJO D., SOILAND-REYES S., ZEDNIK S. & ZHAO J. (2013). Prov-o : The prov ontology. *W3C recommendation*, 30.
- LEFRANÇOIS M. & ZIMMERMANN A. (2018). The Unified Code for Units of Measure in RDF : cdt :ucum and other UCUM Datatypes. *ESWC 2018*.
- MADIN J., BOWERS S., SCHILDHAUER M., KRIVOV S., PENNINGTON D. & VILLA F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3), 279–296.
- MARTINI D., SCHMITZ M. & MIETZSCH E. (2013). agrordf as a semantic overlay to agroxml : a general model for enhancing interoperability in agrifood data standards. In *CIGR Conference on Sustainable Agriculture through ICT Innovation*.
- PALMA R., REZNIK T., ESBRÍ M., CHARVAT K. & MAZUREK C. (2016). An INSPIRE-Based Vocabulary for the Publication of Agricultural Linked Data. In *Ontology Engineering*, volume 9557, p. 124–133. Cham : Springer International Publishing.
- POPOV B., KIRYAKOV A., OGNJANOFF D., MANOV D. & KIRILOV A. (2004). KIM – a semantic platform for information extraction and retrieval. *Natural Language Engineering*, 10(3-4), 375–392.
- PORTER C. H., VILLALOBOS C., HOLZWORTH D., NELSON R., WHITE J. W., ATHANASIADIS I. N., JANSSEN S., RIPOCHE D., CUFI J., RAES D., ZHANG M., KNAPEN R., SAHAJPAL R., BOOTE K. & JONES J. W. (2014). Harmonization and translation of crop modeling data to ensure interoperability. *Environmental Modelling & Software*, 62, 495–508.
- PRESUTTI V., DAGA E., GANGEMI A. & BLOMQUIST E. (2009). extreme design with content ontology design patterns. In *Proc. Workshop on Ontology Patterns*.
- RAIMOND Y. & ABDALLAH S. (2007). *The event ontology*. Rapport interne, Citeseer.
- RIJGERSBERG H., VAN ASSEM M. & TOP J. (2013). Ontology of units of measure and related concepts. *Semantic Web*, 4(1), 3–13.
- ROUSSEY C., BERNARD S., PINET F., REBOUD X., CELLIER V., SIVADON I., SIMONNEAU D. & BOURIGAUT A.-L. (2017). A methodology for the publication of agricultural alert bulletins as LOD. *Computers and Electronics in Agriculture*, 142, 632 – 650.
- SANDERSON R., CICCARESE P., VAN DE SOMPEL H., BRADSHAW S., BRICKLEY D., CASTRO L. J. G., CLARK T., COLE T., DESENNE P., GERBER A. *et al.* (2013). Open annotation data model. *W3C community draft*, 8.
- SHAW R., TRONCY R. & HARDMAN L. (2009). Lode : Linking open descriptions of events. In *Asian Semantic Web Conference*, p. 153–167 : Springer.
- SUÁREZ-FIGUEROA M. C., GÓMEZ-PÉREZ A., MOTTA E. & GANGEMI A. (2012). *Ontology engineering in a networked world*. Springer.

- SUÁREZ-FIGUEROA M. C., GÓMEZ-PÉREZ A. & VILLAZÓN-TERRAZAS B. (2009). How to write and use the ontology requirements specification document. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, p. 966–982 : Springer.
- UREN V., CIMIANO P., IRIA J., HANDSCHUH S., VARGAS-VERA M., MOTTA E. & CIRAVEGNA F. (2006). Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Web Semantics : Science, Services and Agents on the World Wide Web*, **4**(1), 14 – 28.
- WEIBEL S., KUNZE J., LAGOZE C. & WOLF M. (1998). *Dublin core metadata for resource discovery*. Rapport interne.