Annotation sémantique à partir de textes : Cas des observations dans les Bulletins de Santé du végétal

Haïfa Zargayouna[**], Catherine Roussey[*], Synda Ouardani[**]

[*]Irstea UR TSCF, 9 avenue Blaise Pascal, CS 20085, Aubière, France
catherine.roussey@irstea.fr
[**]Laboratoire d'Informatique de Paris Nord (LIPN, UMR 7030), Université Paris 13
prenom.nom@lipn.univ-paris13.fr

Résumé : Dans cet article nous décrivons un schéma d'annotation pour annoter les observations extraites des bulletins agricoles disponibles sur le Web. Le but est de proposer un processus d'annotation automatique qui permet de générer des annotations complexes accessibles via un sparql endpoint. Nous partons d'annotations manuelles des portions de textes, ces annotations permettent d'identifier les entités sémantiques à repérer dans le texte. Nous nous appuyons, dans la mesure du possible, sur des schémas de données et des ontologies disponibles.
Mots-clés : annotation sémantique, ontologie, domaine agricole, bulletins de santé du végétal, schéma d'annotation

1 Introduction

Dans cet article, nous présentons la mise en place d'un schéma d'annotation pour guider le processus d'annotation à partir de plusieurs ontologies. Les annotations produites doivent être structurées pour être exploitables par un moteur d'interrogation tel que SPARQL (*SPARQL Protocol and RDF Query Language* (Segaran *et al.*, 2009). L'objectif de l'interrogation est d'effectuer des recherches au sein d'un corpus d'alerte agricole pour rendre visible une dynamique temporelle et spatiale de phénomènes agricoles. Ceci permettra aux agronomes de visualiser ces dynamiques et facilitera leurs interprétations et leurs prévisions de l'état sanitaire des cultures dans les régions françaises.

Nous nous inscrivons dans une optique d'ouverture de données de l'agriculture. Le domaine agricole est à l'intersection de plusieurs autres domaines : environnement, santé, alimentation etc. En effet les opérations agricoles sont impactées et impactent plusieurs facteurs tels que l'eau, le climat, la biodiversité, etc. L'ouverture des données agricoles pose débat. Le ministère a prévu la création d'une plateforme AgGate pour l'ouverture des données agricoles en France (Bournigal, 2016). Des plateformes existantes tels que API-AGRO (Plateforme de données et de services pour l'écosystème agricole) participent au partage de données avec néanmoins un contrôle en droits de lecture. Même si il y a actuellement de grandes bases de données d'observations, leur ouverture peut être problématique pour leurs auteurs. Les observations scientifiques faites sur les parcelles agricoles ont besoin d'être anonymisées et protégées pour que les agriculteurs puissent faire leur travail correctement.

2 Corpus et cas d'usage

Nous présentons dans ce qui suit les bulletins de santé du végétal et les besoins que nous prenons en compte dans ce travail.

2.1 Les Bulletins de Santé du Végétal

Le Grenelle de l'Environnement [1] et le plan Ecophyto [2] 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. 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, etc.) ou sur la réglementation peuvent figurer également dans les BSV. 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.

Une archive pérenne de ces bulletins agricoles a été constituée, afin d'en extraire un ensemble d'information sur les cultures. Cette archive est disponible comme jeux de données sur le Web de données (Roussey *et al.*, 2016). Une première série d'annotations de ces bulletins a été réalisé pour retrouver un bulletin au sein du corpus. Les annotations portent sur la grande catégorie de culture indiquées dans le titre du bulletin, la date de publication du bulletin et sa région de publication. Ces méta-données ont été extraites des sites Web de publication des BSV et non des contenus des BSV (Roussey & Bernard, 2015). La figure 1 présente un exemple de BSV. La figure 2 présente les annotations actuelles d'un fichier BSV.


BULLETIN DE SANTÉ DU VÉGÉTAL
BOURGOGNE
Grandes cultures n° 20 du 5 avril 2011
A RETENIR CETTE SEMAINE
- Grande majorité de parcelle avec des fleurs
- Avec le développement rapide des colzas le risque méligèthes s'estompe
- Signalement de quelques dégâts de charançons de la tige du colza
- Premiers charançons des siliques piégés mais sans incidence car stade de sensibilité des colzas pas atteint (à partir du stade G2)
Réseau 2010-2011
Cette semaine, 53 parcelles ont fait l'objet d'au moins une observation.
Stade des colzas
Rappel : un stade est atteint lorsque 50% des plantes sont à ce stade.
Les conditions climatiques estivales de la semaine dernière ont permis une accélération des stades et a fortiori dans les secteurs qui ont eu la chance de bénéficier des pluies.
- D1 boutons accolés encore cachés par les feuilles terminales : 2%
- D2 inflorescence principale dégagée, boutons accolés, inflorescences secondaires visibles : 8%
- E boutons séparés, les pédoncules floraux s'allongent en commençant par ceux de la périphérie : 66%
- F1 premières fleurs ouvertes : 20%
- F2 allongement de la hampe florale, nombreuses fleurs ouvertes : 4%
Méligèthes
Météo
Prévisions météorologiques du mercredi 06 au mardi 12 avril :
Temps très ensoleillé et sec pendant toute cette période. Les températures sont en hausse avec des minimales ne descendant pas en dessous de 9°C et des maximales atteignant 20 à 22°C (source météociel).

FIGURE 1 – Un exemple de BSV de la région bourgogne catégorie grande culture à la date du 5 avril 2011

1. Le Grenelle de l'Environnement a eu lieu en 2007 et avait pour but de proposer un ensemble d'actions afin de permettre la mise en place d'une nouvelle politique environnementale.

2. Le plan Ecophyto fut lancé en 2008 dans le but de réduire progressivement l'utilisation des pesticides en France tout en maintenant une agriculture économiquement performante.

2.2 Cas d'usage

Notre objectif est d'extraire du contenu des bulletins agricoles des informations précises. Les informations extraites du contenu textuel sont de plusieurs types : des observation sur les stades de développement des cultures, des observations de la présence des ravageurs et de maladies sur les cultures, à partir de ces deux types d'observations des estimations de risques sur les cultures sont présentées. Dans cet article, nous prenons comme exemple les observations sur les stades de développement des cultures. Le but à terme est de généraliser la proposition à tout type d'observation. Les acteurs sont de deux types :

— agriculteurs et conseillers techniques : ils recherchent l'état général des cultures par type de culture dans leur région ou les régions limitrophes. Leur objectif est de caractériser l'état de leurs cultures par rapport aux autres (retard dans le développement etc...).

— agronomes : ils sont intéressés pour faire le bilan d'une année culturale c'est à dire : savoir quelles cultures sont produites en France par région et leur rendement, et de suivre l'évolution temporelle des cultures pour identifier les causes des variations dans le rendement. Ils ont besoin de connaître précisément l'échantillon sur lequel portent les observations c'est-à-dire combien de parcelles culturales sont observées.

3 État de l'art

L'annotation sémantique est le processus qui consiste à établir des liens entre une ressource (le texte) et une autre ressource (ressource sémantique). L'annotation sémantique fait référence aussi au résultat de ce processus. Zargayouna *et al.* (2015) présentent les différentes étapes du processus d'annotation ainsi que les plates-formes de l'état de l'art. Nous nous intéressons spécifiquement aux schémas d'annotation. Un schéma d'annotation permet de guider le processus d'extraction en définissant les entités sémantiques pertinentes et les liens entre elles. Le schéma d'annotation peut être assez simple et générique tel que modèle appelé OADM (*Open Annotation Data Model*) proposé par le groupe de travail du W3C (2014) (*Web Annotation Working Group*). D'autres schémas correspondent au schéma de l'ontologie projeté comme dans (Abacha & Zweigenbaum, 2010). Nous proposons un schéma d'annotation complexe adossé à plusieurs ontologies et commençons donc l'élaboration de ce schéma par une analyse manuelle des différentes entités et relations pertinentes à annoter. Cette phase d'analyse est suivie par une phase de recherche de ressources et ontologies existantes. Le schéma d'annotation défini manuellement permet de lier ensemble les différentes entités repérés et de les intégrer dans une représentation commune.

4 Exemple d'annotation

Nous présentons dans ce qui suit un extrait de texte présentant une observation sur des colzas. Ce texte est issu du bulletin présenté dans la figure 1. Les entités sémantiques intéressantes sont mises en exergue.

Grandes cultures n° 20 du **5 avril 2011**

..

Cette semaine, **53 parcelles** ont fait l'objet d'au moins **une observation**. ..

Stade des colzas

Rappel : un stade est atteint lorsque 50% des plantes sont à ce stade. Les conditions climatiques estivales de la semaine dernière ont permis une accélération des stades et a fortiori dans les secteurs qui ont eu la chance de bénéficier des pluies.

- **D1** boutons accolés encore cachés par les feuilles terminales : **2%**
- ...

Plusieurs informations sont intéressantes dans ce texte. Tout d'abord nous savons que ce bulletin concerne uniquement des cultures catégorisées sous "Grande culture" et qu'il a été publié le 5 avril 2011. Ainsi toutes les observations détectées dans ce texte seront datées du 5 avril 2011. Nous savons ensuite que des parcelles de colza ont été observées pour déterminer le stade de développement de cette culture dans la région Bourgogne. À noter que le nom de la région (Bourgogne) n'apparaît pas dans le texte car le titre est une image intégrée dans le fichier pdf. En revanche, cette information peut être trouvée dans les annotations déjà disponibles sur les BSV extraites au moment de la constitution du jeu de données [3]. L'échantillon des parcelles de colza observées en Bourgogne se monte à 53 parcelles. Parmi ces 53 parcelles seules 2% d'entre elles ont atteint le stade D1, soit une parcelle. En conclusion dans ce texte nous pouvons déduire qu'il y a une observation du stade de développement des cultures de colza. Le résultat de cette observation est D1. L'échantillon est une parcelle de colza. Cette échantillon est lié à un échantillon globale de 53 parcelles de colza.

5 Schéma d'annotation pour les observations

Le but est d'enrichir le schéma d'annotation des BSV (voir figure 2) avec des annotations sur les observations du stade de développement des cultures. Le schéma d'annotation proposé par Roussey & Bernard (2015) s'appuie sur des concepts et instances publiés par l'IGN [4] ainsi que les concepts FOAF(Brickley & Miller, 2007), SKOS(Miles *et al.*, 2005) et les propriétés définies par Dublin Core.

3. Informations trouvées dans les url des pages aspirées.
4. L'Institut Géographique National publie ses données liées sur `http://data.ign.fr/`

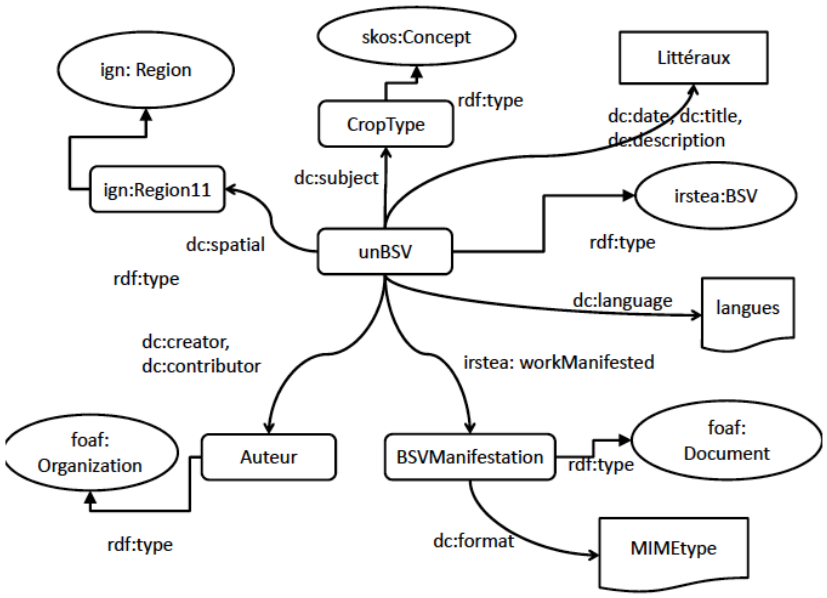

FIGURE 2 – Schéma d'annotation des BSV (Roussey & Bernard, 2015)

Peu de ressources sémantiques existent dans le domaine agricole, notamment pour le français. Il existe, néanmoins des ontologies et schémas de données qui peuvent être réutilisées. Nous nous appuyons également sur le thésaurus qui a été construit manuellement à partir du Larousse Agricole pour référencer les types de culture. Nous avons effectué une recherche sur le Web de données liées ainsi que sur des plates-formes dédiées telle AgroPortal [5]. Nous présentons dans ce qui suit l'ensemble des ressources que nous exploitons pour définir le schéma d'annotation sur les observations.

Thésaurus FrenchCropUsage est un thésaurus construit à l'Irstea. il référence les types de cultures organisés en fonction de leurs usages, intitulé FrenchCropUsage. Les classes d'usage de la production agricole sont destinées, à l'alimentation humaine, ou à l'alimentation animale ou à l'industrie. Les besoins de l'industrie sont la conception de textile (lin, chanvre) ou d'autres utilisations industrielles (huiles à usage particulier, biocarburants). Dans ce thésaurus, les liens hiérarchiques représentent des relations de généralisation/spécialisation entre cultures (céréale/blé). Ce vocabulaire de type de culture est disponible sur le Web de données liées sous format SKOS [6]. Il contient 272 concepts, la profondeur maximale de la hiérarchie est de 6 niveaux.

Ontologie Sosa ou Sensor Observation Sample Actuator (Haller *et al.*, 2017) est une ontologie du domaine des capteurs en cours de validation au W3C. Elle est une évolution de l'ontologie Semantic Sensor Network (SSN). Elle est liée à SSN grâce à une architecture de modularisation horizontale et verticale. Cette ontologie suit le patron de conception de l'ontologie SSN en ajoutant de nouvelles classes et propriétés pour les

5. AgroPortal accessible à `http://agroportal.lirmm.fr/`.
6. Accessible à partir d'agroportal `http://agroportal.lirmm.fr/ontologies/CROPUSAGE`

actionneurs et l'échantillonnage. Cette ontologie est en cours de validation et continue à être modifiée. Par exemple, la définition de la propriété sosa:hasFeatureOfInterest n'est pas encore stabilisée, car cette propriété peut avoir comme co-domaine une instance de sosa:FeatureOfInterest ou une instance de sosa:Sample.

QU QU (Lefort, 2011) repose partiellement sur l'ontologie des systèmes des quantités, des unités, des dimensions et des valeurs (QUDV). Elle est créée pour définir les unités et les quantités.

Prov L'ontologie PROV (PROV-O) (Timothy *et al.*, 2013) fournit un ensemble de classes et de propriétés pour représenter des relations de provenance entre différentes entités. PROV a la caractéristique d'être légère et facilement réutilisable pour à la fois être intégrée dans de nouvelles ontologies ou pour caractériser des jeux de données.

À l'aide de ces 4 ressources nous allons définir un nouveau schéma d'annotation des BSV. La figure 3 présente le schéma d'annotation que nous allons utiliser.

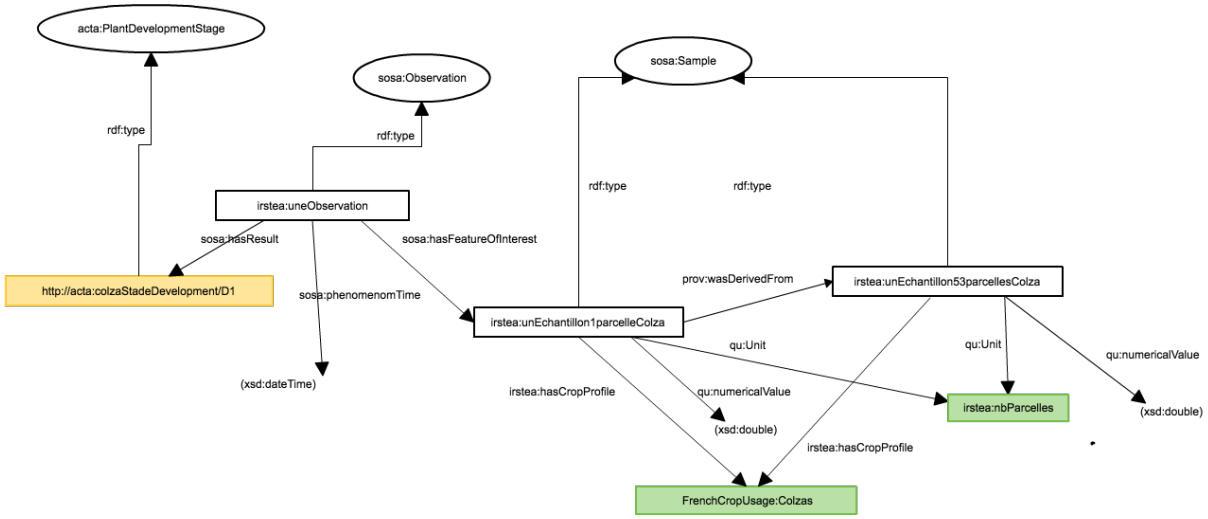

FIGURE 3 – Schéma d'annotation sur les observations. Les concepts sont représentés par des ovales, les instances par des rectangles. Les couleurs montrent les instances déjà définies dans des ressources (en vert dans la ressource locale irstea et en jaune les ressources extérieures).

Nous avons utilisé la classe sosa:Observation pour définir chaque observation. Une observation porte sur l'étude d'un phénomène. Dans notre cas ce phénomène est une parcelle de colza. Cette information est représenté par le lien sosa:hasFeatureOfInterest entre une instance de sosa:Observation et une instance de sosa:Sample. L'instance de Sample représente l'échantillon sur lequel porte cette observation. Cet échantillon est relié à un échantillon plus large représentant les 53 parcelles de colza observées dans la région Bourgogne. Nous utilisons la propriété prov:wasDerivedFrom pour relier l'instance de sosa:Sample représentant une parcelle de colza à l'instance de sosa:Sample représentant les 53 parcelles de colza. Le résultat de l'observation est le stade de développement D1 représenté par la ressource `http://acta:colzaStadeDevelopment/D1`. Ce code est issu d'une nomenclature des stades de déve-

loppement des cultures défini par l'acta [7].

Nous avons utilisé qu:unit pour représenter les unités associées aux échantillons. Cette propriété pointe sur une unité "nombre de parcelles". La propriété qu:numericalValue relie l'échantillon à une valeur xsd :double. Pour préciser la date des observations, nous avons employé la propriété sosa:phenomenonTime pour lier une observation à une date au format xsd:DateTime.

6 Annotations attendues

Nous présentons dans ce qui suit les annotations attendues à partir du texte présenté en exemple plus haut. Nous présentons ces annotations en précisant les calculs nécessaires pour l'automatisation de leur génération à partir de texte.

Repérage des instances d'entités sémantiques

Le repérage des instances d'entités revient à un problème classique de reconnaissance de termes et d'entités nommées. Ainsi par exemple, il faut reconnaître que « 53 parcelles » fait référence à une instance de sosa:Sample et donnera lieu à la création de <irstea:sample_ParcellesColza_Bourgogne_53>. Le nommage des URI nécessite une prise en compte du contexte ainsi qu'une structuration et typage de l'information (par exemple Région/Culture/unité/sample/nombre pour l'exemple précédent).

Voici les annotations qui devraient être obtenues :

```
 <irstea:sample_ParcellesColza_Bourgogne_53>
  rdf:type sosa:Sample.

 <irstea:obs_parcelles_colza_bourgogne_stadeDeveloppement_20110405>
     rdf:type sosa:Observation ;
     sosa:phenomenomTime "05/04/2011".

<irstea:sample_ParcellesColza_Bourgogne_1> rdf:type sosa:Sample.

<acta:colzaStadeDevelopment/D1> rdf:type <acta:PlantDevelopmentStage>.
```

Structuration des informations extraites

Certaines annotations nécessitent de structurer l'information et de reconnaître les types des entités extraites. Il est, par exemple, nécessaire de repérer les valeurs numériques et l'unité.

```
<irstea:sample_ParcellesColza_Bourgogne_53>
  hasCropProfile <FrenchCropUsage:Colzas> ;
  qu:Unit <irstea:nbParcelles> ;
  qu:numericalValue "53".
```

Le typage permet d'effectuer des calculs. Cela est le cas par exemple pour retrouver que 2% de 53 parcelles fait référence à une seule parcelle. Le repérage du pourcentage permet de déduire

7. Les instituts techniques agricoles `http://www.acta.asso.fr/`

le nombre de parcelles concernés par l'observation sans que cela soit mentionné explicitement dans le texte.

```
<irstea:sample_ParcellesColza_Bourgogne_1>
  hasCropProfile <FrenchCropUsage:Colzas> ;
  qu:Unit <irstea:nbParcelles> ;
  qu:numericalValue "1" ;
  prov:wasDerivedFrom <irstea:sample_ParcellesColza_Bourgogne_53>.
```

Détection de relations

Il s'agit de mettre en lien les entités annotées. Comme par exemple lier le stade de développement à l'échantillon concerné.

```
<irstea:obs_parcelles_colza_bourgogne_stadeDevelopement_20110405>
  sosa:hasResult <acta:colzaStadeDevelopment/D1> ;
  sosa:hasFeatureOfInterest <irstea:sample_ParcellesColza_Bourgogne_1>.
```

7 Conclusion

Nous avons présenté la méthodologie de construction d'un schéma d'annotation des observations dans les bulletins de santé du végétal et les annotations RDF qui en découlent. Nous allons mettre en place l'automatisation des annotations à partir de textes en définissant des heuristiques utiles pour l'extraction. L'automatisation des annotations nécessitent une segmentation préalable des BSV selon la culture et une définition de contextes pour l'extraction afin de réduire le bruit. Nous avons commencé à utiliser l'outil OMTAT, qui permet une annotation automatique ou manuelle de textes, et comptons l'enrichir par les heuristiques propres à notre domaine. Notre objectif est de généraliser l'annotation à partir de textes à d'autres types d'observation telles que l'observation des ravageurs et des maladies sur les cultures.

Remerciements

Ce travail a bénéficié partiellement d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083.

Références

ABACHA A. B. & ZWEIGENBAUM P. (2010). Annotation et interrogation sémantiques de textes médicaux. In *Atelier Web Sémantique Médical, IC*.

BOURNIGAL J.-M. (2016). Portail de données pour l'innovation en agriculture. http://agriculture.gouv.fr/un-portail-de-donnees-pour-linnovation-en-agriculture-la-synthese-du-rapport.

BRICKLEY D. & MILLER L. (2007). Foaf vocabulary specification 0.91.

HALLER A., JANOWICZ K., COX S., PHUOC D. L., TAYLOR K. & LEFRANÇOIS M. (2017). Semantic sensor network ontology. https ://www.w3.org/TR/vocab-ssn/.

LEFORT L. (2011). Ontology for quantity kinds and units : units and quantities definitions. https ://www.w3.org/2005/Incubator/ssn/ssnx/qu/qu-rec20.html.

MILES A., MATTHEWS B., WILSON M. & BRICKLEY D. (2005). Skos core : simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*, p. 3–10.

ROUSSEY C. & BERNARD S. (2015). Annotation des bulletins de santé du végétal. In *Actes de la 7ème édition de l'Atelier Recherche d'Information SEmantique (RISE)*, p. 12 pages.

ROUSSEY C., BERNARD S., PINET F., REBOUD X. & CELLIER V. (2016). Gestion sémantique des bulletins de santé du végétal dans le projet vespa. In *Actes de l'atelier IN-OVIVE @IC 2016*, p. 12 pages.

SEGARAN T., EVANS C., TAYLOR J., TOBY S., COLIN E. & JAMIE T. (2009). *Programming the semantic web*. O'Reilly Media, Inc.

TIMOTHY L., SATYA S. & DEBORAH M. (2013). Prov-o : The prov ontology. https ://www.w3.org/TR/prov-o/.

W3C (2014). Web annotation data model. http ://www.w3.org/TR/annotation-model/.

ZARGAYOUNA H., ROUSSEY C. & CHEVALLET J.-P. (2015). Recherche d'information sémantique : état des lieux. *Traitement Automatique des Langues*, **56**(3), 49–73.