Le Web de données pour faciliter l'exploitation des Bulletins de santé du végétal

Pour réduire l'usage agricole des pesticides, plusieurs actions ont été mises en œuvre en France, dont la diffusion des Bulletins de santé du végétal. Depuis 2012, plus de trois mille bulletins sont publiés chaque année. Répartis sur différents sites, le contenu et la présentation de ces bulletins ne sont pas uniformes, ce qui rend leur exploitation difficile. En mobilisant le Web de données et les technologies du Web sémantique, les chercheurs d'Irstea ont développé un système d'interrogation qui facilite la recherche et l'extraction d'informations utiles aux acteurs économiques et scientifiques.

Dans cet article, nous présentons le système que nous avons développé afin de faciliter la recherche d'information dans un corpus de bulletins agricoles intitulés les Bulletins de santé du végétal (BSV). Ce système s'adresse *a priori* aux différents acteurs des filières végétales, mais il peut aussi faciliter des objectifs de recherche agronomique. Les BSV publiés ont été collectés sur différents sites. Le contenu de chaque BSV a été décrit par des annotations utilisant un thésaurus français des cultures. Une architecture innovante est proposée pour publier ces annotations sur le Web de données. Le Web de données (*Linked Data* en anglais) et les technologies du Web sémantique, en reliant les données à des référentiels (ontologies et thésaurus), permettent non seulement une meilleure visibilité des données sur le Web, mais aussi une organisation des contenus dans un réseau global d'information afin d'offrir une large palette de moyens pour leur utilisation.

Cet article s'organise de la manière suivante : tout d'abord, nous présentons le Web de données et ses avantages. Les Bulletins de santé du végétal sont ensuite détaillés ainsi que les problématiques associées. La méthode de publication d'une archive de bulletins est quant à elle décrite en commençant par le thésaurus construit pour harmoniser la description des bulletins. Avant de conclure, nous présentons quelques analyses de ce corpus montrant l'intérêt d'avoir harmonisé la description des bulletins.

Présentation du Web de données

Cette section reprend les informations du site internet de la Bibliothèque nationale de France sur le Web de données et les technologies du Web sémantique [BNF].

Les technologies du Web sémantique s'adressent aux acteurs intéressés par la publication de données structurées sur le Web, c'est-à-dire aux institutions et organisations fournisseuses de données qui souhaitent que les développeurs informatiques accèdent plus facilement à leurs bases de données pour développer de nouvelles applications. Ces fournisseurs mettent à jour régulièrement leurs bases et publient sur le Web de données des bases de qualité, mises à jour régulièrement et qui leur apportent une reconnaissance et une visibilité. En effet, le Web de données représente une opportunité nouvelle de diffuser et d'encourager la réutilisation des données, en favorisant :

- la visibilité et le référencement des données sur le Web, en facilitant l'accès aux données aux moteurs de recherche et aux internautes spécialistes de ces technologies ;
- l'interopérabilité, en permettant la fédération de silos de données de nature, de provenance et de structure différentes ;
- la fiabilité, en traçant l'origine de la ressource, grâce aux identifiants Web URI, permettant ainsi aux organismes utilisateurs de données de se lier à des données fiables tout en se positionnant comme tiers de confiance ;

• la souplesse de réutilisation par des tiers, en permettant de récupérer et de retraiter l'ensemble des données nécessaires, de les croiser avec des jeux de données extérieures ou des données locales.

Les technologies du Web sémantique sont un ensemble de standards et de technologies développé par le W3C (l'un des principaux organismes de normalisation du Web) visant à faciliter l'exploitation des données structurées, notamment en permettant leur interprétation par des machines. Il s'appuie sur les concepts suivants :

• URI : ce sont les identifiants uniques des ressources sur le Web. Pour être valides ils doivent suivre un certain nombre de règles (*URI dereferenceables*) ;

• RDF : c'est un modèle de description des données dans lequel toute ressource est identifiée par une URI et les données sont décrites sous la forme de triplets sujet/prédicat/objet. Dans ce triplet, le sujet et le prédicat sont toujours exprimés par des URI. L'objet peut être exprimé sous la forme d'une URI ou d'une chaîne de caractères (littérale). Un ensemble de triplets RDF qui décrit une ressource compose un graphe ;

• SPARQL : c'est un langage de requêtes sur les données en RDF. Les requêtes SPARQL permettent d'interroger dynamiquement les données en RDF, sans télécharger l'ensemble des données ;

• RDFS et OWL : quand on exprime des données en RDF, il est préférable d'utiliser autant que possible des classes et des propriétés déjà définies dans des vocabulaires existants. En RDF, les ressources appartiennent généralement à des classes qui les regroupent par types (documents, concepts, personnes, etc.). Elles sont qualifiées grâce à des propriétés (prédicats) qui définissent un aspect, une caractéristique, un attribut, ou une relation spécifique de ces ressources. Les classes et les propriétés sont décrites dans des vocabulaires RDF qui permettent aux machines de les comprendre et de les exploiter. Ces vocabulaires affectent des URI à chaque classe et chaque propriété qu'ils définissent. RDFS et OWL sont deux vocabulaires RDF qui permettent de définir d'autres vocabulaires, c'est-à-dire de définir des classes, des propriétés, des hiérarchies de classes et de propriétés et leurs comportements. OWL permet en outre d'associer à chaque classe des contraintes facilitant l'exploitation des données par des machines : on parle alors d'ontologies.

Les Bulletins de santé du végétal

Dans un premier temps, nous présentons les Bulletins de santé du végétal (BSV), et nous détaillons ensuite les difficultés rencontrées pour interroger ce corpus de bulletins agricoles.

Présentation des bulletins

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 BSV sont une des modalités mises en place par ces réseaux de surveillance dans l'ensemble des régions et départements d'outre-mer. Le BSV est un document d'information à la fois technique et réglementaire, rédigé sous la responsabilité d'un comité régional d'épidémiosurveillance.

❶ **Première page d'un Bulletin de santé du végétal (BSV) de la région Midi-Pyrénées, catégorie « grandes cultures ».**


BULLETIN DE SANTE DU VEGETAL
MIDI-PYRENEES
Grandes Cultures n°28
A retenir
Maïs
• Stades phénologiques et état des cultures
• Sésamie
• Pyrale

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 comme 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 ❶ présente un exemple de première page d'un BSV de la région Midi-Pyrénées.

Problèmes liés aux BSV

Depuis le début de leur parution, les BSV sont gratuits et accessibles sur les sites Web des chambres régionales d'agriculture et des directions régionales de l'alimentation, de l'agriculture et de la forêt (DRAAF). Ils sont donc répartis sur différents sites Web (un site par région).

Les BSV mis en ligne sur les sites Web ne sont pas toujours pérennes ; ainsi les BSV des années antérieures ne sont souvent plus accessibles sur leurs sites.

Les BSV sont téléchargeables au format pdf. Ce format est très pratique pour rendre accessible le contenu aux humains. Malheureusement, ce format rend difficilement accessible le contenu aux traitements informatiques.

Les comités de rédacteurs des BSV varient en fonction de la région et de la filière végétale. Leur contenu et leur présentation ne sont donc pas uniformes.

Les typologies des cultures utilisées pour organiser les BSV sur chaque site Web varient aussi en fonction des régions. Par exemple, le site Web de la région Midi-

Pyrénées découpe ses bulletins viticoles en plusieurs rubriques « Viticulture – Cahors, Lot », « Viticulture – Fronton », etc. Le site Web de la région Île-de-France a défini une rubrique qui lui est propre « Grandes cultures – Pommes de Terre – Légumes industriels ».

La construction d'une archive pérenne des BSV organisée suivant une typologie des cultures unique pour faciliter leur interrogation est donc nécessaire pour disposer d'un historique de l'observation des cultures en France sur plusieurs années.

Publication d'une archive des BSV sur le Web de données

Dans cette section, nous détaillons comment nous avons construit l'archive des BSV. Après avoir construit un thésaurus français des cultures intitulé *FrenchCropUsage*, nous avons collecté et stocké les BSV. Les BSV et le thésaurus ont été publiés sur le Web de données. Ainsi des URI ont été créées pour représenter chaque BSV et chaque élément du thésaurus. Des annotations ont été créées pour associer à l'URI du bulletin une URI du thésaurus représentant une culture observée dans le bulletin. L'ensemble des annotations a été rendu accessible sur un serveur capable de répondre à des requêtes SPARQL (un *SPARQL end-point*).

Thésaurus FrenchCropUsage

Un thésaurus est une liste organisée de termes normalisés et validés. Les termes sont reliés entre eux par des relations d'équivalence, de hiérarchie ou d'association. Ces termes représentent les concepts d'un domaine explicités par des définitions en langue naturelle. Ils constituent un langage contrôlé pour l'indexation de documents et facilitent la recherche d'information. Certains de ces termes sont préconisés pour décrire le contenu des documents (les descripteurs), d'autres sont rejetés et pointent vers le descripteur à utiliser. Le thésaurus a pour objectif de diminuer les ambiguïtés en rejetant les termes polysémiques et d'éviter la dispersion en préconisant l'usage d'un terme unique pour représenter la même notion.

Méthodologie de construction

À notre connaissance il n'existe pas de ressource structurée française permettant d'organiser les cultures en fonction de leur destination (alimentation humaine directe, alimentation animale, industrie alimentaire, etc.) et du type de système de culture (grandes cultures, maraîchage, etc.). Le thésaurus AGROVOC de la FAO [1] couvre plusieurs domaines allant de l'agriculture à la foresterie. Malheureusement, son organisation des noms de culture ne correspond pas à nos besoins.

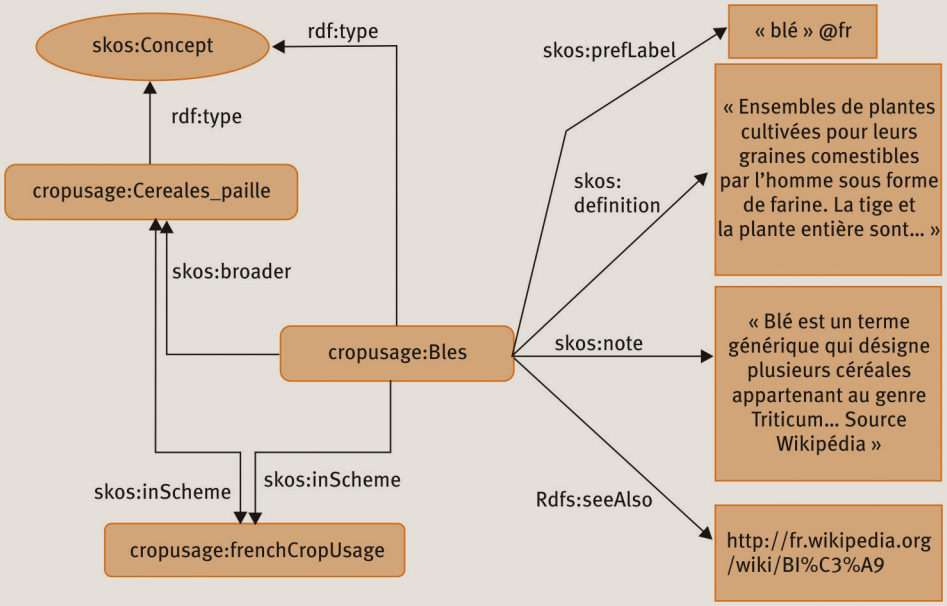
Pour construire manuellement le thésaurus FrenchCropUsage, nous avons utilisé des documents disponibles comme les statistiques agricoles annuelles, le site Wikipédia France et le Larousse agricole.

L'ensemble de la hiérarchie a été modélisé à l'aide du vocabulaire SKOS proposé par le W3C. SKOS est un vocabulaire RDF permettant de décrire des référentiels de type thésaurus. Il permet de décrire des concepts représentés par des termes et d'exprimer les relations entre ces concepts. Ce vocabulaire est disponible sur le Web de données.

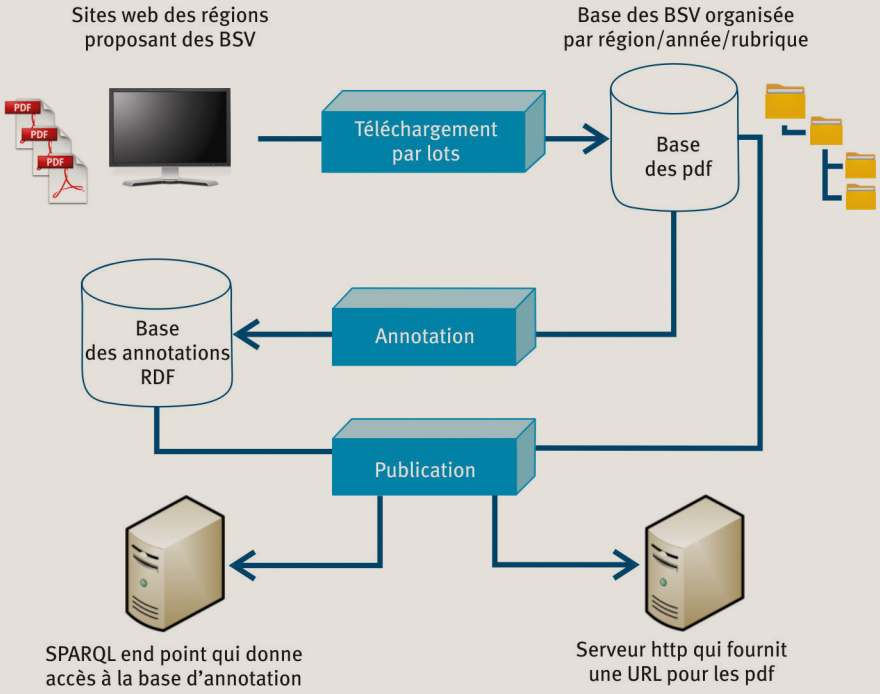
Le thésaurus FrenchCropUsage [2] contient 272 concepts. La profondeur maximale de sa hiérarchie est de six. La figure ❷ présente le concept représentant la culture du blé.

1. Organisation des Nations unies pour l'alimentation et l'agriculture (en anglais *Food and Agriculture Organization of the United Nations*).
2. Plus d'information sur FrenchCropUsage est disponible sur : **http://ontology.irstea.fr/pmwiki.php/Site/FrenchCropUsage**

❷ **Exemple de concept extrait du thésaurus FrenchCropUsage représentant la culture du blé.**



❸ **Processus mis en œuvre pour publier les annotations des Bulletins de santé du végétal (BSV).**



Publication sur le Web de données

Dans notre approche, les annotations sont construites à partir des données issues des sites Web donnant accès aux BSV. Chacun de ces sites propose un classement des BSV au minimum par année et par type de culture. Le type de culture est généralement repris dans le titre du bulletin. Comme le montre la figure ❸, ces informations constituent la base du processus d'annotation des BSV. Une fois les annotations RDF construites, elles sont publiées sur le Web à l'aide d'un SPARQL end point. Un serveur apache fournit une URI pour identifier chaque BSV. Notre archive contient à l'heure actuelle 19 448 BSV.

Nous déterminons la région à partir de la localisation du site Web qui met les BSV à disposition. Par exemple, les BSV Auvergne se téléchargent sur le site Web de la chambre régionale d'agriculture d'Auvergne. Le nom du site Web est donc associé à la région concernée et tous les bulletins extraits du même site sont annotés par la même région.

Comme vu précédemment, chaque site Web organise l'accès à ses BSV de manière différente, mais au moins une des rubriques porte sur la culture. Pour chacune des rubriques en lien avec le type de culture, nous leur avons donc attribué manuellement un ensemble de concepts de notre thésaurus FrenchCropUsage.

Exemples d'analyse

Nous décrivons ici trois analyses que nous avons pu réaliser grâce à notre archive publiée sur le Web de données. Notre système permet d'interroger la description des BSV par le biais de requêtes SPARQL.

Les filières végétales les plus représentées dans l'archive

Le système a permis d'identifier quelles étaient les filières végétales les plus représentées au sein du corpus, c'est-à-dire les cultures les plus suivies. L'analyse s'est limitée aux six grandes filières définies à la racine du thésaurus FrenchCropUsage : « cultures légumières », « cultures fruitières », « cultures fourragères », « grandes cultures », « horticulture ornementale » et « zones non agricoles ». Les résultats sont présentés dans le tableau ❶.

Comme l'indique le tableau ❶, les cultures les plus représentées dans l'archive sont les grandes cultures, suivies par les cultures légumières et les cultures fruitières. Les cultures les moins suivies par les BSV sont les cultures fourragères. Ces résultats sont conformes au tableau

❶ **Répartition des Bulletins de santé du végétal (BSV) par filière de 2009 à 2016.**

Grande catégorie de culture	Nombre de BSV	Pourcentage	Nombre de régions
Culture fourragère	175	0,89 %	6
Culture fruitière	3 118	16 %	21
Culture légumière	4 916	25 %	21
Grandes cultures	6 279	32 %	20
Horticulture ornementale	1 085	6 %	14
Zone non agricole	1 143	6 %	18

publié sur Alim'Agri qui présente pour chaque région les différentes éditions de BSV. La quatrième colonne du tableau ❶ indique le nombre de régions qui ont une édition en lien avec l'une des six filières. La plupart des régions, sauf les régions d'outre-mer, ont une édition en lien avec les grandes cultures, les cultures fruitières et les cultures légumières. Par contre, uniquement six régions publient des BSV dans la filière « cultures fourragères ». En cultures fourragères, seul le campagnol est suivi. Un BSV est réalisé uniquement dans les régions où le campagnol a le plus d'impact. Pour les autres bioagresseurs, la prairie est peu suivie du fait qu'il y a très peu de produits phytosanitaires utilisés sur ces cultures. Il est à noter que le maïs à destination fourragère est intégré dans les grandes cultures.

Les régions les plus impliquées dans le suivi des cultures

Les résultats précédents laissent penser qu'il existe une disparité géographique dans le suivi des cultures. Cette disparité géographique a été étudiée plus en détail dans notre archive. Pour ce faire, nous avons calculé le nombre de BSV publiés par région (tableau ❷).

❷ Nombre de Bulletins de santé du végétal (BSV) par région.

Région	Nombre de BSV
Alsace	773
Aquitaine	1 284
Auvergne	803
Basse-Normandie	920
Bourgogne	695
Bretagne	812
Centre-Val de Loire	1 735
Champagne-Ardenne	506
Corse	348
Franche-Comté	476
Guadeloupe	122
Guyane	14
Haute-Normandie	1 511
Ile-de-France	832
La Réunion	125
Languedoc-Roussillon	158
Limousin	566
Lorraine	824
Martinique	45
Mayotte	1
Midi-Pyrénées	1 293
Nord-Pas-de-Calais	1 019
Pays de la Loire	1 008
Picardie	1 135
Poitou-Charentes	987
Provence-Alpes-Côte d'Azur	725
Rhône-Alpes	731

On remarque une réelle disparité entre les régions dans notre archive. Le Centre-Val de Loire est la région qui publie le plus de bulletins, suivie par la région Haute-Normandie. En effet, les régions n'ont pas tous la même politique de publication des BSV. Par exemple, les régions Centre-Val de Loire et Normandie ont plusieurs éditions pour les grandes cultures : « oléagineux », « céréales à paille », « protéagineux », « maïs », « betterave sucrière », etc., alors que dans la majeure partie des régions, il n'y a qu'une seule édition en grandes cultures. Autre exemple, la région Midi-Pyrénées publie plusieurs éditions de BSV en viticulture, une par zone géographique de production, soit sept éditions en viticulture. Ces éditions supplémentaires induisent une augmentation du nombre de bulletins publiés.

Les départements d'outremer publient moins de BSV que les régions métropolitaines. Les publications sont généralement plus espacées avec peu de filières suivies.

La publication des BSV est donc fortement dépendante de la région et des productions présentes.

Répartition spatiale du suivi des cultures

Pour compléter l'étude de la disparité géographique du suivi des cultures, nous avons calculé le nombre de BSV par filière par région.

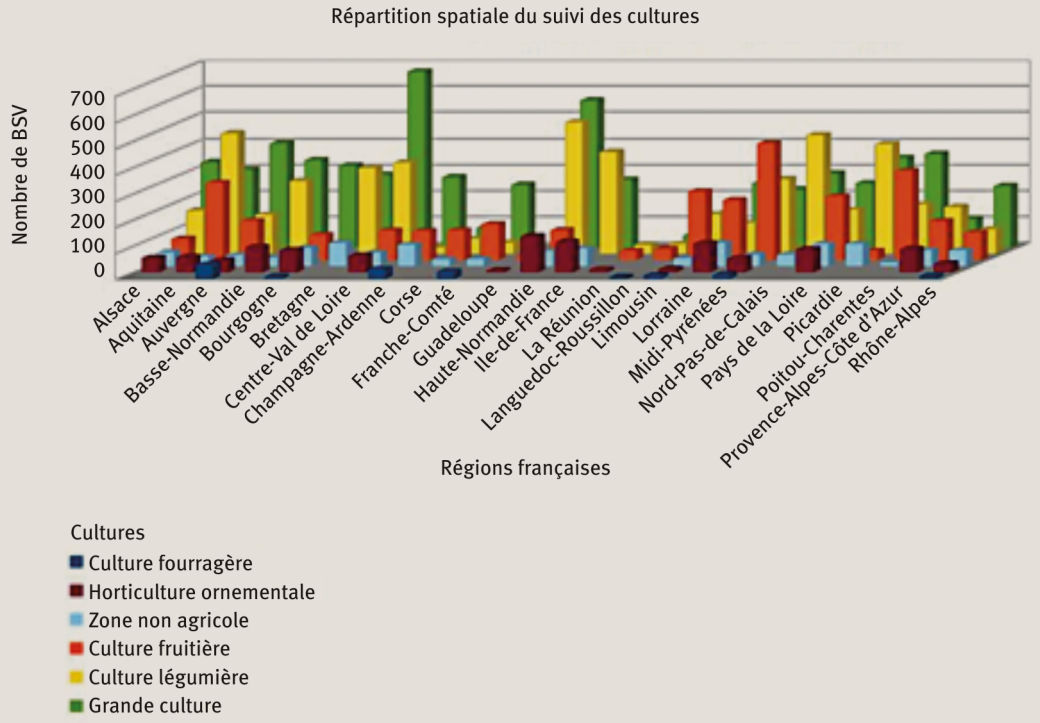
Le graphique de la figure ❹ montre l'importance du suivi de certaines cultures dans chaque région. Les régions Centre-Val de Loire et Haute-Normandie sont très orientées sur le suivi des grandes cultures. Les légumes sont très observés en Aquitaine, Haute-Normandie, Ile-de-France, Nord Pas-de-Calais et Picardie. Les cultures fruitières sont très suivies en Aquitaine, Midi-Pyrénées et Poitou-Charentes.

Conclusion

Cet article présente l'archive des BSV que nous avons publiée sur le Web de données. Pour faciliter l'interrogation de cette archive, chaque bulletin a été annoté avec des concepts du thésaurus FrenchCropUsage. Ce nouveau thésaurus francophone organise les noms de culture en fonction de leur usage. Notre archive et son système d'interrogation visent *a priori* les différents acteurs des filières végétales et ils répondent aussi à des objectifs de recherche agronomique sur le suivi des cultures.

Ces annotations sont publiées sur le Web de données afin dêtre reprises et connectées à d'autres. Ainsi, grâce à notre système d'interrogation, il est par exemple possible de rechercher des BSV de régions différentes portant sur la même culture et la même période. On pourrait aussi compléter les annotations des BSV en les liant à d'autres sources, telles les bulletins météorologiques. Un tel lien aurait du sens dans la mesure ou beaucoup de processus épidémiques de maladies ou de ravageurs des cultures sont très dépendants des conditions météorologiques. Comme les BSV sont disponibles sur le Web de données, tout organisme peut rajouter ses propres annotations pour compléter la description des BSV. ■

❹ **Graphique représentant les résultats de la requête sur le nombre de Bulletins de santé du végétal (BSV) par région et par culture.**



Les auteurs

Catherine ROUSSEY et Stephan BERNARD
Université Clermont Auvergne, Irstea, UR TSCF,
Centre de Clermont-Ferrand,
9 avenue Blaise Pascal, CS 20085,
F-63178 Aubière, France.
catherine.roussey@irstea.fr
stephan.bernard@irstea.fr

En savoir plus...

BNF, *Web sémantique, Web de données : définitions*, Site Web de la BNF,
disponible sur : **http://www.bnf.fr/fr/professionnels/anx_web_donnees/a.web_donnees_definitions.html**

ROUSSEY, C., BERNARD, S., PINET, F., REBOUD, X., CELLIER, C., SIVADON, I., SIMONNEAU, D., BOURIGAULT, A., 2017,
A methodology for the publication of agricultural alert bulletins as LOD, *Computers and Electronics in Agriculture*, volume 142,
partie B, novembre 2017, p. 632-650,
disponible sur : **http://www.sciencedirect.com/science/article/pii/S0168169917306361**

ROUSSEY, C., BERNARD, S., PINET, F., REBOUD, X., CELLIER, C., 2016, Gestion Sémantique des Bulletins de Santé du Végétal dans le projet Vespa, *in : Actes de l'atelier IN-OVIVE adossé à la conférence IC*, 7 juin 2016, Montpellier, France.

IRSTEA BSV, Le site Web présentant notre archive des Bulletins de santé du végétal (BSV),
disponible sur : **http://ontology.irstea.fr/pmwiki.php/Site/BSV**

IRSTEA FRENCHCROPUSAGE, Le site Web présentant le thésaurus FrenchCropUsage,
disponible sur : **http://ontology.irstea.fr/pmwiki.php/Site/FrenchCropUsage**