ecoQuery: a Semantic Module to Query Biodiversity Data on the Web

Julie CHABALIER, Olivier COULLET, Amandine SAHL and Olivier ROVELLOTTI
NATURAL SOLUTIONS, 68 rue Sainte, 13001 Marseille, France
{julie_chabalier, olivier_coullet, amandine_sahl, olivier_rovellotti}@natural-solutions.eu

Abstract *The rapid growth of collected biodiversity data, their heterogeneity and their dissemination can be the bottleneck of safeguarding our environment. Semantic Web technologies aim to address these issues. Based on these technologies, we extended the ecoRelevé platform which is dedicated to biodiversity data management with the ecoQuery module. This module is able to query the Web of data in order to move up scale and provide an answer to complex issues.*

Keywords Biodiversity, linked data, Web of data, semantic Web.

ecoQuery : un Module Sémantique pour Interroger les Données de Biodiversité sur le Web

Résumé *L'accroissement rapide du nombre de données de biodiversité collectées, leur hétérogénéité et leur dispersion sur le Web peuvent être un frein pour la préservation du monde qui nous entoure. Les technologies du Web sémantique ont pour objectifs de pallier ces problèmes. Basé sur ces technologies, nous avons étendu la plateforme ecoRelevé dédiée à la gestion des données de biodiversité avec le module ecoQuery. Ce module est capable d'interroger le Web de données pour apporter un changement d'échelle et une réponse à des questions complexes.*

Mots-clés Biodiversité, linked data, Web de données, Web sémantique.

1 Introduction

L'érosion accélérée de la biodiversité, diagnostiquée lors de l'Evaluation des écosystèmes du Millénaire [1], a été confirmée par les résultats de l'étude des coûts de l'inaction en matière de biodiversité en 2008 [2]. Ces études ont eu pour conséquence d'accentuer, ces dernières années, la prise de conscience internationale sur le rôle vital de la biodiversité.

Cette prise de conscience s'est naturellement accompagnée d'un besoin généralisé d'approfondir les connaissances du monde qui nous entoure dans l'objectif de mieux le préserver. Ce besoin se traduit par une augmentation de la collecte des données de biodiversité. Leur analyse vise à proposer des zones de protection spéciale ou de conservation i.e. dans le cadre du réseau Natura 2000 [3], faire des études d'évaluation du risque pour une espèce ou un écosystème, quantifier le nombre d'individus d'une espèce dans une zone géographique donnée, etc. Répondre à ces objectifs nécessite d'intégrer un grand nombre de données de biodiversité. La nature complexe, l'hétérogénéité et la dispersion sur le Web de ces données rendent leur intégration et, en conséquence, leur analyse longues et fastidieuses.

Le Web sémantique désigne un ensemble de technologies dont l'objectif est de favoriser l'interopérabilité des données en vue notamment de leur intégration. Utilisant la famille de langages développés par le W3C, le contenu des ressources du Web devient accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles [4]. Dans ce cadre, le « Linked Data » désigne un ensemble de bonnes pratiques pour publier et connecter les données structurées sur le Web [5]. Ces données constituent le Web de données. Elles peuvent être découvertes et collectées automatiquement par des machines sans intervention humaine. L'utilité du « Linked Data » pour représenter la sémantique des observations relatives à la biodiversité et favoriser ainsi l'interopérabilité des données a été décrite dans [6]. Les auteurs soulignent également l'importance de construire des outils offrant aux utilisateurs la possibilité de fouiller le Web de données et facilitant l'intégration automatique des données.

Récemment, nos travaux se sont concentrés sur la construction de la plateforme open source ecoRelevé[1] dédiée à la gestion des données de biodiversité [7]. Cette plateforme a été initialement conçue pour stocker et visualiser les données d'observation. Le module ecoQuery, présenté ici, est une extension de cette plateforme, permettant de consommer les données publiées au sein du Web de données rendant ainsi transparentes leur recherche et leur intégration. Agrémentée du module ecoQuery, la plateforme ecoRelevé favorise des études de données à plus large échelle et a pour objectif de répondre à des questions plus complexes.

Ce papier s'organise de la manière suivante : la section 2 présente le cas d'utilisation sur lequel nous avons basé la construction du module ecoQuery, la section 3 traite du Web de données et de son application à la biodiversité, la section 4 présente l'extension de la plateforme ecoRelevé avec le module ecoQuery.

2 Définition d'un Cas d'Utilisation

Pour spécifier les exigences fonctionnelles du module ecoQuery, nous avons défini un cas d'utilisation concret dans le domaine de la biodiversité. Ce cas d'utilisation devait répondre à un problème courant et transposable et nécessiter l'intégration de données pour être résolu.

Suivant ces critères, nous nous sommes intéressés à la réalisation d'une étude environnementale dans le cadre du projet de restructuration des dispositifs de protection contre les crues entre les Communes de Sénas et de Cheval-Blanc dans le département du Vaucluse. Plusieurs études ont en effet démontré qu'il existe un risque majeur de débordement de la Durance lors de crues pouvant conduire à l'inondation des communes situées sur la rive droite de la rivière. Les communes concernées ont décidé d'établir une ligne de protection empêchant tout risque de débordement (digue insubmersible, élévation du remblai de la voie ferrée, etc.). La construction de ce type d'ouvrages nécessite la réalisation d'un inventaire de la zone. Nous avons restreint ce cas d'utilisation à l'inventaire de l'avifaune.

Pour chaque inventaire naturaliste, la première étape consiste généralement à établir une liste de taxons déjà observés sur le site. Ces informations sont, tout d'abord, recherchées sur le Web et utilisées, dans un deuxième temps sur le terrain, en tant que référence. Les naturalistes notent cependant avec attention les nouveaux taxons observés. Différentes informations peuvent être ensuite collectées de manière à évaluer l'incidence de la construction d'ouvrages sur ces taxons. Dans le cadre des oiseaux, il est, par exemple, important de distinguer pour le site donné, les oiseaux nicheurs des oiseaux de passage et des oiseaux hivernants. Le statut de protection des espèces est également important lors des études d'impacts.

Répondre à cette étude nécessite donc d'intégrer un ensemble de données. A minima, ces données doivent être décrites par trois attributs spécifiques, i.e. le nom scientifique du taxon, ses coordonnées géographiques et la date d'observation. La section suivante présente les données disponibles dans le Web de données et comment y accéder.

3 Le Web de Données et la Biodiversité

Le « Linked Open Data Cloud » désigne l'ensemble des données en accès libre publiées suivant les principes du « Linked Data ». En septembre 2010, 203 jeux de données dont 40 en sciences de la vie et 15 en géographie[2] étaient présents dans le Web de données libres. Parmi ces jeux de données, nous avons recensé six jeux impliquant des données ouvertes relatives à la biodiversité (voir table 1).

Les données minimales nécessaires à la mise en place de notre cas d'utilisation concernent essentiellement des données d'observations (taxons, coordonnées, dates d'observation). Seul le jeu de données « TaxonConcept » propose des données d'observations. Cependant, bien qu'il y ait actuellement pratiquement 100000 concepts relatifs aux taxons, ce jeu de données ne contient qu'un nombre relativement faible d'observations. Ce nombre est amené à évoluer, de nombreux travaux vont dans ce sens (notamment dans le cadre du Global Biodiversity Information Facility, GBIF [8]).

En attendant l'enrichissement du Web de données, nous avons choisi d'appliquer les principes du « Linked Data » sur une source qui, nativement, ne les respecte pas. L'agrégateur de données aviaires le plus

[1] https://code.google.com/p/ecoreleve/ (stable release)
[2] http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22_colored.html

important est l'« Avian Knowledge Network (AKN) » [9]. L'AKN contient actuellement environ 89 millions d'observations sur plus de 9500 taxons. Nous avons extrait de l'AKN les données correspondant aux observations d'oiseaux en France et transformé ces données en RDF (Resource Description Framework) en utilisant le standard de données de biodiversité Darwin Core[1]. Les données obtenues ont été stockées dans un entrepôt RDF, la version open source de Virtuoso (VOS). Elles sont accessibles par le langage de requête SPARQL, spécifique au données RDF, via un « endpoint[2] ». La section suivante présente les outils mis en place pour consommer ces données de manière à répondre aux besoins du cas d'utilisation.

Jeu de données	Description
TaxonConcept	Fournit des URI informatives afin d'améliorer la qualité et la stabilité des liens entre une espèce et les données relatives à cette espèce.
Geospecies	Information sur les ordres, familles, espèce.
European Nature Information System	Informations sur les espèces et sur les sites européens où elles sont attendues.
DPpedia	Extraction d'informations structurées de Wikipedia. Les données sont liées à un ensemble de données présentes sur le Web.
Fishes of Texas	Standardisation et géoréférencement des informations scientifiques connues sur les poissons d'eau douce du Texas.
Geonames	Fournit des données géographiques à partir de différentes sources telles que les noms de lieux, l'altitude, la population, etc. Les coordonnées (latitude et longitude) sont en WGS84 (World Geodetic System 1984).

Table 1. Jeux de données du Web de données relatif à la biodiversité.

4 Les Requêtes Sémantiques et l'Intégration de Données

Cette section décrit nos travaux en cours, à savoir la plateforme ecoRelevé, un système de gestion de données de biodiversité, et son nouveau module ecoQuery interrogeant, de manière transparente pour l'utilisateur, le Web de données.

4.1 La Plateforme ecoRelevé

La plateforme ecoRelevé comprend actuellement trois modules : 1) ecoRelevé Core dédié au stockage des données, 2) ecoRelevé Data pour l'import des données de terrain à partir du logiciel mobile pocket eRelevé [10] et 3) ecoRelevé Explorer permettant de visualiser les données sur une carte. Cette application internet riche (RIA) est basée sur Adobe Flex/AIR et OpenScales pour la couche de présentation (module Explorer et Data), JAVA/Hibernate pour la manipulation/persistance des objets biologiques (i.e. relevés, taxons), geoserver pour la manipulation des objets cartographiques, et PosgreSQL/PosGIS pour le stockage des données (module Core).

Le module Explorer propose, via un filtre, une visualisation cartographique des données suivant trois dimensions : la taxonomie, le temps et la localisation géographique. Le filtre traduit les contraintes sur ces dimensions en un service Web interrogeant la base de données sous-jacente au module ecoRelevé Core. Les filtres ainsi créés sont gérés au sein d'un gestionnaire. Afin d'obtenir de la plateforme ecoRelevé qu'elle réponde aux exigences fonctionnelles spécifiées lors de l'élaboration du cas d'utilisation, il s'agissait d'étendre l'application d'un filtre au Web de données, d'intégrer les données obtenues avec celles stockées dans le module Core et de visualiser les résultats. La section suivante présente le module ecoQuery répondant à ces exigences.

[1] http://rs.tdwg.org/dwc/
[2] http://natural01.gn-noc.com:8890/sparql (URI du graphe : urn:rdf.Occurences_AKN_dwc)

4.2 Le Module ecoQuery

Le module ecoQuery transforme à la volée les contraintes sur les trois dimensions de chaque filtre établi par l'utilisateur dans le langage SPARQL. La requête ainsi obtenue s'exécute sur le « endpoint » spécifique que nous avons mis en place pour pallier le manque actuel de données d'observations au sein du Web de données. Le XML résultant de la requête est ensuite traité afin de visualiser les résultats au sein du module Explorer. Le gestionnaire a également été modifié de manière à étendre un filtre à une autre source de données sans modifier la requête sous-jacente.

La Figure 1 présente la requête SPARQL sur laquelle sont automatiquement appliquées les contraintes définies par l'utilisateur dans le module Explorer (sous forme de filtres). La construction de la requête est transparente pour l'utilisateur. A l'instar du service Web interrogeant la base de données sous-jacente au module ecoRelevé Core, la requête SPARQL est construite automatiquement à partir de l'interface cartographique, de la liste de taxons et du diagramme de temps présents au sein de l'interface graphique du module Explorer.

Actuellement, le standard Darwin Core autorise seulement des requêtes basiques, d'expressivité limitée, nous verrons dans la discussion les efforts effectués pour pallier ce problème.

```
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT distinct ?scientificName ?lat ?long ?date
FROM <urn:rdf:Occurences_AvianKnowledgeNetwork>
WHERE{
?occurrence rdf:type dwc:Occurrence;
dwc:Identification ?id;
geo:lat ?lat;
geo:long ?long;
dcterms:date ?date.
?id dwc:scientificName ?scientificName.
}
```

Figure 1. Requête SPARQL générique.

Associé à la plateforme ecoRelevé[1], le module ecoQuery permet désormais de réaliser une requête sémantique sur le Web de données et d'intégrer les résultats avec les données stockées dans le module Core de la plateforme (voir Figure 2 pour l'architecture du système).

[1] https://code.google.com/p/ecoreleve/ (latest release)

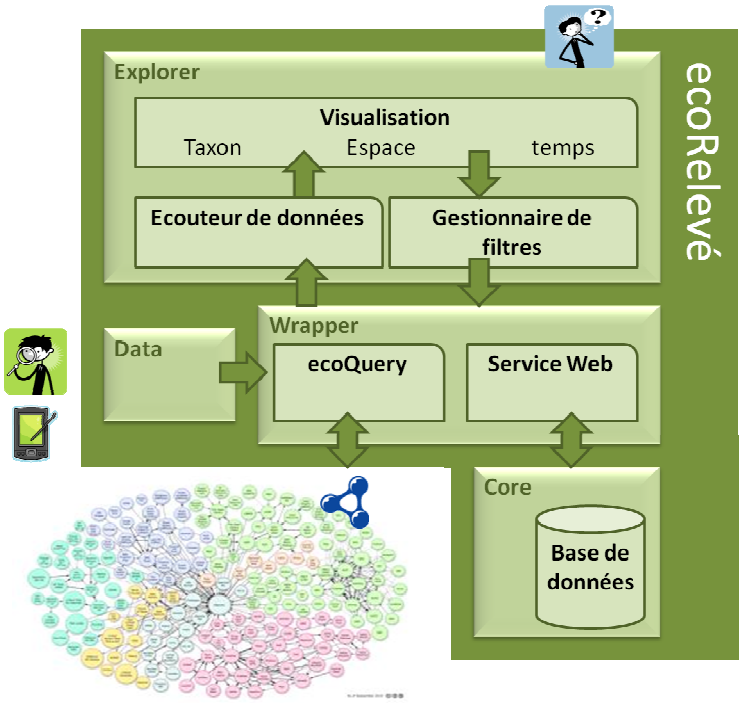

Figure 2. Architecture de la plateforme ecoRelevé agrémentée du module ecoQuery.

4.3 L'Inventaire de l'Avifaune

Afin de valider l'utilisation de la plateforme ecoRelevé agrémentée du module ecoQuery, nous avons réalisé l'étude définie par le cas d'utilisation présenté dans la section 2.

L'utilisation du module ecoRelevé Explorer débute par la sélection d'une emprise géographique. Cette sélection peut se faire à partir de la carte ou via une recherche textuelle utilisant le service Web associé à l'ontologie Geonames [11]. Dans ce dernier cas, l'échelle de la portion de carte affichée dépend de l'unité administrative associée à la localisation (Pays, Région, Département, Ville, etc.). Nous avons ensuite établi un filtre en fixant comme contraintes la commune de Cheval-Blanc, la classe des oiseaux (taxon Aves) et une plage de dates (2000 – 2011). Le module ecoQuery permet d'exécuter le filtre sur le jeu de données présent dans le VOS, à savoir les données de l'AKN.

Seules deux stations (association des coordonnées géographiques, d'une date, d'un observateur) sont affichées sur la carte, couvrant six taxons : *Columba livia* (pigeon biset), *Garrulus glandarius* (geai des chênes), *Phoenicurus ochruros* (Rougequeue noir), *Erithacus rubecula* (Rouge-gorge familier), *Sitta europaea* (Sittelle torchepot) et *Fringilla coelebs* (Pinson des arbres). Avec deux stations, l'effort de prospection est trop faible pour tirer des conclusions sur ces données. Cette première utilisation de la plateforme enrichie du module ecoQuery permet cependant de dégager une liste de taxons déjà observés sur le site.

Pour pallier le manque de données, nous avons effectué une mission d'inventaire de l'avifaune muni du logiciel mobile pocket eRelevé. La liste de taxons précédente a été utilisée comme référence pour cette étude de terrain. Après deux jours de prospection sur la commune de Cheval-Blanc, tous les oiseaux rencontrés ont été identifiés et les données correspondantes ont été enregistrées sur l'appareil mobile, i.e. la date, le nom scientifique du taxon et les coordonnées GPS. De retour de mission, les données ont été importées à l'aide du module ecoRelevé Data au sein de la base de données du module Core. Les données ainsi stockées dans le Core, le gestionnaire de filtres est utilisé pour les intégrer avec les données du Web. L'ensemble des stations est ensuite visualisée sur une carte. En supplément des six taxons de la liste de référence, 17 taxons sont désormais affichés sur la carte. Parmi ces taxons, certains, tels que *Coracias garrulus* (Rollier d'Europe), sont quasi-menacés (statut de conservation de l'IUCN).

Pour finir, les données de l'AKN ont permis d'étendre cette étude à tout le département du Vaucluse sans nouvel effort de prospection. Nous obtenons 49 taxons dont 34 nouveaux par rapport au jeu de données

précédent (voir Figure 3). Ces nouveaux taxons pourraient éventuellement être associés au rapport d'inventaire en tant que taxons potentiels.

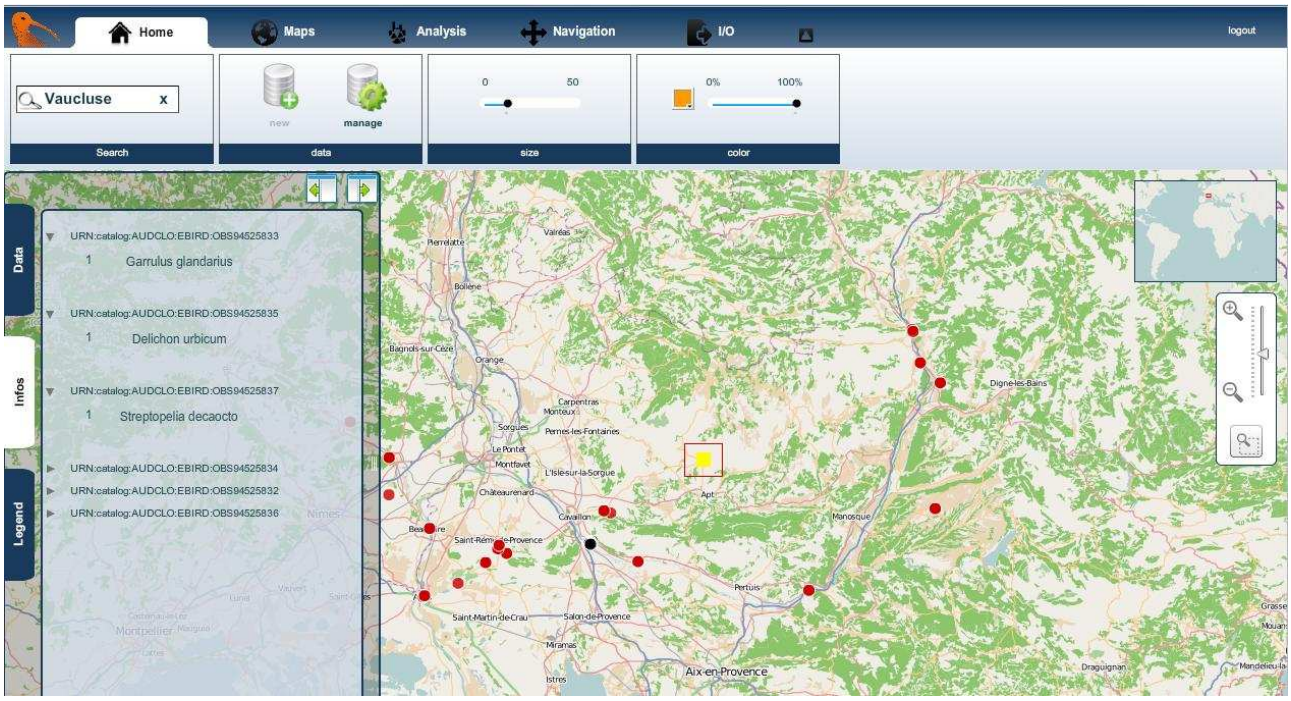

Figure 3. Présentation des stations sur l'emprise du Vaucluse.

5 Discussion

ecoQuery est un module sémantique permettant d'interroger le Web de données de manière transparente et de visualiser les résultats intégrés avec des données locales au sein de la plateforme ecoRelevé, une plateforme dédiée à la gestion des données de biodiversité. Bien qu'actuellement les observations naturalistes soient peu présentes dans le Web de données, ce premier prototype démontre la faisabilité de l'approche sur un jeu de données aviaires transformé en RDF et stocké dans un entrepôt accessible via un « SPARQL endpoint ».

Dans cette approche, l'avantage d'utiliser le RDF réside essentiellement dans l'interopérabilité qu'il procure. Contrairement aux données contenues dans des bases de données classiques, difficilement accessibles et généralement non compatibles entre elles, les jeux de données, décrits en RDF et présents dans le Web de données sont désormais accessibles, liés et exploitables par tous. L'application sur un cas d'utilisation concret a démontré que l'accès aux données de biodiversité via le Web de données est un gain de temps pour le naturaliste par rapport au travail long et fastidieux de recherche de données, notamment bibliographiques, en amont des études environnementales. En effet, la plateforme ecoRelevé agrémentée du module ecoQuery propose un accès centralisé et transparent à des données unifiées.

Récemment, un jeu de données d'observation a été mis à disposition au sein du « endpoint » du créateur de TaxonConcept[1]. Il s'agit des données obtenues dans le cadre d'un bioblitz[2] organisé lors de la conférence TDWG 2010 à Woods Hole, Massachussetts[3]. ecoQuery a donc été utilisé pour interroger ce nouveau jeu de données et obtenir une visualisation intégrée avec les données de l'AKN concernant le Massachussetts dans la plateforme ecoRelevé. Cette dernière expérimentation a permis de valider notre étude sur un jeu de données distant, RDF-natif et a démontré que notre système était techniquement prêt à accueillir les données d'observation qui viendront prochainement enrichir le « Linked Open Data Cloud ».

[1] http://lsd.taxonconcept.org/sparql
[2] Bioblitz : inventaire biologique intensif sur une portion bien précise de terrain.
[3] http://bioblitz.tdwg.org/

Actuellement, la version RDF du standard Darwin Core utilisé pour structurer les données de biodiversité n'exploite pas pleinement les capacités du Web sémantique pour définir des données. Bien qu'offrant un vocabulaire commun pour partager les données, ce vocabulaire ne permet pas de définir formellement les termes utilisés ni de raisonner sur les données de manière à produire automatiquement de nouvelles connaissances. Le niveau d'expressivité des requêtes qui en découle reste donc assez bas. Pour pallier ce problème, nous travaillons sur l'élaboration d'une ontologie de la biodiversité dans le cadre du projet collaboratif « ecoOnto » [12]. Basée sur une extension de l'ontologie OBOE [13], les classes de cette ontologie seront mises en relation avec les standards actuels de la biodiversité, notamment Darwin Core, de manière à obtenir une meilleure structuration des données d'observation et ainsi résoudre des problèmes complexes de biodiversité via les mécanismes de raisonnement associés aux logiques de description. Cette ontologie sera notamment utilisée dans la plateforme ecoRelevé pour élaborer des requêtes complexes sur le Web de données et fournir ainsi une aide à la décision pour l'utilisateur. Dans cet objectif, la plateforme ecoRelevé sera prochainement étendue de manière à accueillir des données de biodiversité autres que les observations, e.g. les statuts de protection des espèces, les habitats, etc.

Remerciements

Ce travail est soutenu par OSEO INNOVATION.

Références

[1] Millennium Ecosystem Assessment (MEA). Current state and trends assessment, *Washington D.C., Island Press*, 2005.

[2] P. Sukhdev, The Economics of Ecosystems and Biodiversity (TEEB) interim Report, 2008.

[3] O. P. Ostermann, The need for management of nature conservation sites designated under Natura 2000. *Journal of Applied Ecology*, 35, 968-97, 1998.

[4] T. Berners-Lee, J. Hendler and O. Lassila. *The semantic web. Scientific American*, p. 29-37, 2001.

[5] C. Bizer, T. Heath and T. Berners-Lee, Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5 (3). pp. 1-22. 2009.

[6] O. J. Reichman, M. B. Jones and M. P. Schilldhauer, Challenges and Opportunities of Open Data in Ecology. *Science*, 331,6018:703-705, 2011.

[7] O. Coullet, A. Sahl, J. Chabalier and O. Rovellotti, ecoRelevé: an open source response to the biodiversity crisis. *Proceedings of the conference for Open Source Geospatial Software (FOSS4G)*, 2010.

[8] T. Catapano, D. Hobern, H. Lapp, R.a. Morris, N. Morrison, N. Noy, M. Schildauer and D. Thau, Recommendations for the Use of Knowledge Organisation System by GBIF, *GBIF white paper*, February 2011.

[9] M. Iliff, L. Salas, E. Ruelas Inzunza, G. Ballard, D. Lepage and S. Kelling, The avian knowledge network: a partnership to organize, analyse, and visualize bird observation data for education, conservation, research, and land management. *Proceedings of the fourth Inrternational Partners in Filght Conference: Tundra to Tropics*, pp. 365-376.

[10] R. Kipré, O. Coullet, A. Sahl, J. Chabalier, C. Duval, O. Assunçao, O. Rovellotti, Pocket eRelevé : nouvelle approche de collecte de données sur le terrain. *Géomatique Expert*, 69, 24 – 27, 2009.

[11] GeoNames ontology. http://www.geonames.org/ontology/ontology_v2.2.1.rdf.

[12] J. Chabalier, EcoOnto : une ontologie pour la biodiversité. *Acte du colloque national d'écologie scientifique (Ecologie 2010)*, 2010.

[13] J. Madin, S. Bowers, M. Shildhauer, S. Krivov, D. Pennington and F. Villa. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2: 279-296, 2007.