



**LABORATOIRE DE RECHERCHE EN INFORMATIQUE**  
**Unité mixte de recherche C.N.R.S.**  
**UNIVERSITE D'ORSAY**

---

**Chantal Reynaud, Professeur**

Equipe LaHDAK (Large scale Heterogeneous DATA and Knowledge)

Tel : 33 1 69 15 68 37

Fax : 33 1 69 15 65 79

e-mail : [chantal.reynaud@lri.fr](mailto:chantal.reynaud@lri.fr)

## **Rapport sur le mémoire de thèse de Ghislain Auguste Ateazing intitulé Publishing and Consuming Government Linked Data on the Semantic Web**

**17 mars 2015**

Le travail de thèse de Ghislain Auguste Ateazing concerne le Web des données. Il s'est intéressé à plusieurs problèmes : (1) la publication de jeux de données dans le nuage de façon à les rendre le plus possible interopérables, (2) la visualisation des données liées, (3) les standards de publication des données sur le Web des données. Les contributions sont très nombreuses et précisées au fur et à mesure de ce rapport. Elles sont à la fois théoriques et pratiques. Ainsi, 8 jeux de données ont été publiés dans le nuage français (correspondant à environ 10 % en volume), plusieurs outils et applications exploitant des jeux de données réelles ont été développés. Ces travaux ont été effectués dans le cadre du projet ANR DataLift et des activités de standardisation du W3C.

### **Contenu du mémoire :**

Cette thèse comporte 8 chapitres dont une introduction et une conclusion, organisés en 3 parties composées chacune de 2 chapitres, l'ensemble étant complété par 2 annexes (32 pages).

L'introduction présente le web des données et l'information géographique dans le nuage des données puis les problèmes de recherche qui se posent aujourd'hui liés à l'intégration des données géographiques provenant de jeux de données différents. Ghislain Auguste Ateazing présente ensuite l'ensemble de ses contributions en les structurant en trois parties : (1) la modélisation, la publication et le requêtage de données géographiques, (2) la visualisation de données liées gouvernementales, (3) les activités de standardisation des données liées gouvernementales. L'introduction se termine en introduisant le contenu de chacun des chapitres du manuscrit.

La 1ère partie traite de la modélisation, publication et requêtage de données géographiques.

Le chapitre 2 porte sur les données géo-spatiales du web. Les différentes approches utilisées pour modéliser des données géographiques sur le web sont présentées (entités géographiques et géométries associées). Cette étude met en évidence certaines limites, notamment la nécessité de disposer d'URI relativement intuitives, déréférencées et facilement interrogeables, pour tous les systèmes de référence de coordonnées (CRS) utilisés et des noms d'entités administratives françaises décrites plus nombreuses. C'est à ce besoin que Ghislain Auguste Ateazing répond alors en proposant des vocabulaires pour représenter des objets géométriques, des systèmes de référence de coordonnées (CRS) et des entités géographiques, son objectif étant d'accroître l'interopérabilité entre les jeux de données géographiques françaises et d'autres jeux de données du nuage. Les vocabulaires proposés sont des extensions de vocabulaires pré-existants avec en plus, un usage explicite de CRS identifiés par des URIs pour la géométrie et une représentation structurée d'objets géométriques en RDF. *Ce chapitre fait état de données très précises. Le fait que les propositions faites tirent partie des technologies actuelles mises en oeuvre dans le cadre du web des données tout en s'appuyant sur*

*l'existant (extension de vocabulaires pré-existants), dénote une connaissance très fine des points forts et des limites du domaine d'étude, le Web des données et également son application au domaine des données géographiques, notamment françaises. Par ailleurs, le fait que des éléments issus de ces propositions fassent partie de propositions de standardisation du W3C est une preuve de leur grande pertinence.*

Le chapitre 3 porte sur la publication, l'interconnexion et l'interrogation des données géographiques. Ghislain Auguste Ateazing montre d'abord, sur un exemple réel, les différentes représentations associées à des vues différentes d'une même entité géographique selon le jeu de données. Quatre outils de conversion de données géo-spatiales en RDF sont ensuite présentés et leurs limites mises en évidence. GeomRDF, outil de la plate-forme Datalift, se distingue des autres outils par le fait qu'il est conforme à GeoSPARQL tout en fournissant des représentations d'objets géométriques structurées. Ghislain Auguste Ateazing passe ensuite à l'étude de l'étape d'interconnexion, énonce les critères utiles pour réaliser l'interconnexion, des mesures de distance adaptées à la comparaison d'objets géométriques, et présente un scénario d'alignement de l'ontologie GeOnto avec 6 autres jeux de données. La section suivante est une étude sur le mode de stockage des données RDF via des *triple stores* suivie de la description de la boîte à outils Geoknow Stack et de DataLift. Enfin, le chapitre se termine par une présentation de différentes contributions en matière de publication de jeux de données et d'interconnexion dans le cadre du projet DataLift. Huit jeux de données relevant de quatre domaines différents ont été ainsi publiés, soit environ 10 % des données du nuage français (340 millions de triples). L'apport de ces jeux de données est illustré. *La façon dont ce chapitre est rédigé ne se comprend qu'après la lecture de la dernière partie décrivant les contributions effectuées dans le cadre de la thèse. Cela rend la lecture difficile, le lecteur s'interrogeant constamment sur le pourquoi des descriptions faites. Pourquoi décrire la boîte à outils Geoknow Stack ? Pourquoi décrire DataLift ? Par ailleurs, les recommandations orientées stockage à adopter lors de la publication des données ne sont pas explicites alors qu'il s'agit d'une contribution rappelée dans la synthèse faite en fin de chapitre. Ceci mis à part, les contributions en matière de données françaises publiées dans le nuage sont très significatives.*

La partie 2 traite de la génération de représentations visuelles pour les données liées.

Le chapitre 4 concerne l'analyse et la description d'outils de visualisation et d'applications innovantes exploitant des données liées. Deux catégories d'outils de visualisation sont étudiées, les outils portant sur des données structurées autres que des données RDF et ceux opérant sur des données RDF. A partir de cette étude, Ghislain Auguste Ateazing conclut sur un certain nombre de limites et énumère les critères d'évaluation de tels outils. Ce chapitre se poursuit par une description d'applications qui permettent d'exploiter facilement des données liées grâce à des outils de visualisation appropriés. Enfin, un vocabulaire est proposé pour faciliter la recherche d'applications et d'outils de visualisation.

*Bien que le contenu de chaque sous-section soit intéressant en soi, le lecteur a du mal à comprendre l'objet du chapitre pris dans sa globalité. Il manque des transitions entre les sous-sections montrant le lien entre chacune d'elles. Le vocabulaire DVIA proposé est précisément décrit mais rien n'est dit de très précis sur son utilisation. On imagine que la réponse se trouve dans le chapitre suivant. Il aurait été souhaitable de davantage justifier l'ensemble des éléments présentés en indiquant beaucoup plus explicitement en tout début de la partie le résultat visé par le travail présenté.*

Le chapitre 5 traite de la création et de la génération d'applications de visualisation. Un outil d'aide à la visualisation de jeux de données du LOD (LDVizWiz) est présenté. L'objectif est de permettre à un utilisateur inexpérimenté de comprendre rapidement le contenu de données via des visualisations appropriées en lui évitant d'avoir à écrire lui-même des requêtes SPARQL. Dans un premier temps, il s'agit d'identifier la présence de données appartenant à l'une des grandes catégories de données préalablement présentées, puis de compléter les propriétés de ces entités avec celles d'autres jeux de données. L'outil recommande ensuite des modes de visualisation adaptés aux catégories de données concernées. Une implémentation réalisée en exploitant des catégories accessibles à partir de 444 end-points valide la proposition. La possibilité de restreindre la visualisation aux propriétés les plus importantes via une modélisation utilisant le vocabulaire du W3C Fresnel a également été étudiée. L'étude s'appuie sur une enquête auprès d'utilisateurs et sur l'examen des propriétés retenues dans l'expérimentation GKP (Google Knowledge Panel) de Google. Ce chapitre comprend aussi la description très précise de deux applications de visualisation de données, l'une concernant l'organisation d'une conférence, exploitant les données du LOD mais aussi des données sociales, la seconde concernant des statistiques dans le domaine de l'éducation. Une dernière section porte sur

l'amélioration de la découverte d'applications exploitant les données du LOD en décrivant ces applications comme des données liées, via un vocabulaire spécifique et un plugin d'annotation. *Le contenu de ce chapitre est très intéressant. Il est le résultat d'un travail conséquent. Les propositions répondent parfaitement aux besoins d'aujourd'hui. Elles sont pertinentes et ont donné lieu au développement d'outils et d'applications implémentant des scénarios très actuels, ce qui est très convaincant. En revanche, il manque l'aspect validation par les utilisateurs, un point important puisque les représentations visuelles leur sont destinées.*

La partie 3 traite de l'utilisation de vocabulaires standards et du contrôle de compatibilité de licences.

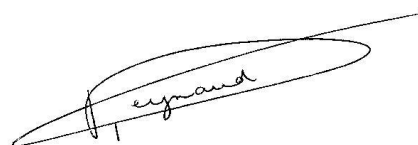
Le chapitre 6 traite de contributions concernant les vocabulaires pour décrire les données publiques. Ghislain Auguste Ateazing montre que la création de ces vocabulaires respecte des principes de la méthodologie Neon appliquée aux ontologies. Il propose une technique d'harmonisation des préfixes permettant d'aligner des vocabulaires, puis une métrique, basée sur la notion de contenu de l'information, pour évaluer les vocabulaires et aider à les sélectionner. Enfin, un module de DataLift aidant à réutiliser des vocabulaires de données publiques de façon à améliorer l'interconnexion de données publiées et favoriser leur exploitation, est présenté. *Ce chapitre montre que les travaux réalisés sont en accord avec les travaux actuels de standardisation du W3C, les propositions faites portant sur l'utilisation de vocabulaires standards.*

Le chapitre 7 aborde le problème de compatibilité des licences de vocabulaires et propose un framework en ligne, LIVE, basé sur une approche existante de composition de licences, qui vérifie automatiquement la compatibilité de licences hétérogènes en raisonnant sur les données portant sur ces licences. Un état de l'art permet de positionner ce travail par rapport à l'existant et montre en quoi il est original. Des indications de performance attestent de la faisabilité dans des temps acceptables. *Ce travail est présenté comme une première piste pour traiter du problème de compatibilité de licences. L'idée développée est pertinente. La partie concernant le raisonnement automatique réalisé aurait toutefois mérité davantage d'explications.*

Enfin, le chapitre 8 conclut en rappelant toutes les contributions (16 points sont listés) et énonce des perspectives.

En conclusion, il s'agit d'un énorme travail, avec de nombreuses propositions de solutions intéressantes et innovantes, portant sur la publication, l'exploitation et la visualisation de données dans le Web des données, ces réalisations étant accompagnées d'un investissement dans le cadre d'activités de standardisation du W3C. Le nombre important de contributions donne parfois une impression de travail tout azimut, toutefois chaque thème fait l'objet d'un travail approfondi, débutant par une analyse d'un problème, menant à des propositions mises en oeuvre dans le cadre d'outils ou d'applications sur des jeux de données réelles appliqués à de vrais cas d'utilisation. Ainsi travail théorique et travail appliqué sont étroitement imbriqués, montrant que Ghislain Auguste Ateazing maîtrise parfaitement le domaine du Web des données, tant sur le plan des idées que des techniques mises en oeuvre. Ces travaux ont été reconnus par la communauté internationale au travers de nombreuses publications. Au vu de tous ces éléments, je donne un avis très favorable pour la soutenance de Ghislain Auguste Ateazing en vue de l'obtention du Doctorat en Informatique de Telecom ParisTech.

Fait à Orsay, le 17 mars 2015



Pr. Chantal Reynaud  
LRI – Université Paris-Sud