How to quickly detect the overlap and the consistency between LADM with LandInfra and LandXML: Application of schema matching techniques

Jacynthe POULIOT, Canada, Claire MONNEY, Switzerland, Jens INGENSAND, Switzerland, Suzie LARRIVÉE, Canada

Key words: LADM, LandInfra (InfraGML), LandXML, schema matching, syntactic-semantic comparison

SUMMARY

In this paper, we explore the schema matching techniques to compare the content of three geospatial standards which are LADM, LandInfra (InfraGML) and LandXML. Those standards all refer to the concept of "land" and we will try to quantify the similarity of them based on syntax and semantic comparison of the class names exposed in their respective schema. Consequently, we will demonstrate the applicability, the accuracy and the usefulness (rapidity and automation) of schema matching techniques for comparing the content of standards. The comparison is performed with XSD (XML Schema Definition) files that describe the schema in English. The results show that syntactic match rate between LADM-LandInfra (54%) is higher than LADM-LandXML (10%). In adding the semantic information extracted from Wordnet, the match rate between LADM-LandInfra goes to 84% and 59% for LADM-LandXML. In comparing our matching results with two independent sources of information that already and manually compared these three standards, we obtained distinctive results. The correctness of LADM-LandInfra is 60%, while the correctness of LADM-LandXML is only 20%. The applicability of schema matching is positively demonstrated while the usefulness and the accuracy still need further improvements in order to make any statement.

RÉSUMÉ

Dans ce papier, nous explorons les techniques d'appariement de schémas pour comparer le contenu de trois standards géospatiaux soit LADM, LandInfra (InfraGML) and LandXML. Ces trois standards réfèrent au concept de « land » et nous allons tenter de quantifier leur similitude en tenant compte de la syntaxe et de la sémantique des noms de classe contenu dans leur schéma respectif. Nous allons ainsi démontrer l'applicabilité, l'exactitude et la facilité (rapidité et automation) des techniques d'appariement de schémas. La comparaison est effectuée à partir des fichiers XSD (XML Schema Definition), qui présentent la modélisation en anglais des standards. Les résultats montrent que, lorsqu'uniquement la syntaxe est prise en compte, le taux d'appariement de LADM-LandInfra (54%) est plus élevé que celui de LADM-LandXML (10%). En tenant compte des relations sémantiques possibles extraites de Wordnet, le taux d'appariement de LADM-LandInfra grimpe à 84%, alors qu'il revient à 59% pour LADM-LandXML. En comparant nos résultats avec des sources d'information externes qui ont déjà comparées manuellement ces trois standards, nous obtenons un taux d'exactitude de 60% pour LADM-LandInfra et de 20% pour LADM-LandXML. L'applicabilité des techniques d'appariement de schémas est démontrée par nos expérimentations tandis que l'exactitude et la facilité montrent des résultats divergeant qui nécessiteront d'autres tests et analyses.

How to quickly detect the overlap and the consistency between LADM with LandInfra and LandXML: Application of schema matching techniques

Jacynthe POULIOT, Canada, Claire MONNEY, Switzerland, Jens INGENSAND, Switzerland, Suzie LARRIVÉE, Canada

1. INTRODUCTION

Standards, as proposed by the International Standard Organization (ISO), are beneficial to ensure reliable and good quality products for the consumers. It exists a large diversity of standards in the field of geospatial data and systems addressing various purposes (conceptual modelling, data modelling for specific feature, data exchange and interoperability, etc). For instance, ISO-19152 Land Administration Domain Model (LADM) offers a specific arrangement of common aspects related to land administration that include elements above and below the surface of the earth (ISO 19152-LADM).

Let's imagine an organization interested in cadastre and land administration systems looking to identify the most suitable and existing standards related to "land" concepts. At this preliminary phase, the organization is maybe not interested in getting full detail information about the standards but only needs to get an overall appreciation of overlap and consistency between standards. What would be the stratagem of the organization to answer this matter? They will certainly look at the geospatial standards that refer to the concept of "land" designed by known authorities as OGC[1], ISO[2], Inspire[3], etc. If we type "land+standards+geospatial" on a Web browser, the organization will probably and easily find LADM. Furthermore and because they contain the term "land" in the title, the organization will certainly find the standards LandInfra and LandXML. LandInfra (OGC 15-111rl), referring to the contraction of the terms Land and Infrastructure, proposes the conceptual modelling of objects as civil engineering infrastructure facilities and land (as road, railway, land division and condominiums, facilities). LandXML (LandXML 2.0 2016), created by the LandXML organization, is a XML file format commonly used to interchange land survey and civil engineering data.

The information about similarity levels between geospatial standards is relevant since it may help professionals and stakeholders who are interested in standardization for their own needs to better understand the content of existing standards and thus to clarify the subsequent selection processes. This information about similarity can also provide valuable material in the phase of design or standard alignments for organizations or people that develop and promote standards such as the ISO, the OGC or in Canada, the SCC (Standards Council of Canada). It may even provide a better understanding of the level of interoperability between standards in highlighting the matched concepts and even schema structure.

Consequently, comparing the content (concepts and relationships) of existing standards is of custom usage. For instance, it is interesting to note in the annex D of the official document of the Landinfra (OGC 15-111rl), a comparison with other standards such as LADM and

[1] OGC-Open Geospatial Consortium (https://www.opengeospatial.org/standards).
[2] ISO-International Standardization Organization (https://www.iso.org/home.html).
[3] Inspire – Infrastructure for spatial information in Europe (https://inspire.ec.europa.eu/).

LandXML. The tables in the annex D present a list of corresponding content between standards; it was identified by experts after reviewing in detail the content of all standards. Stubkjaer in 2015 also discussed and compared land-related models that are LADM, LandInfra and LandXML. They manually compared, class per class, the content of all the standards and draw some interesting tables and conclusions. More recently, Kumar et al., (2019) manually and precisely compare IFC, CityGML and InfraGML.

2. OBJECTIVES

Accordingly, unless an explicit comparison already exist, there is very few approaches to automatically and rapidly compare geospatial standard (Pouliot et al., 2018). This idea is exactly the starting point of this early research project: How an organization can quickly compare a limited number of geospatial standards to understand their similarity? In using the term "quickly" in the statement, we are referring to not spending hours in the learning and comparison process but being able, alone, with limited numbers of actions and resources, to get an overall and a systematic view of all available geospatial standards and progress in the understanding and selection process. Obviously, knowing in detail a standard is required if people wants to implement it (this action is not targeted by our experiment).

As hypothesized by Pouliot et al. 2018, we believe that schema matching techniques are valuable methods to rapidly and automatically compare geospatial standards, which are by definition normalized and well accepted by the communities. But this hypothesis still has to be demonstrated and this paper is a step forward in this direction. In this paper, and even though we understand that these standards are not at the same level and not design for the same purpose, we selected for the experiment LADM (ISO 19152), LandInfra-InfraGML (OGC InfraGML 2017) and LandXML (LandXML 2.0). First, it is obvious to observe that they all refer to the term LAND in their respective title; we may then guess that the standards refer to the same concept. In our comparison, we will first try to answer this simple question by comparing their respective content based on the use of core terms, such as "land". Besides, previous and independent comparisons exist between these three standards (OGC 15-111rl; Stubkjaer 2015) and we will use them as control reference. In this manner, we will be in a better position to demonstrate our hypothesis.

In summary, and based on the XML schema (XSD) comparison and considering both syntactic and semantic points of view, the paper tries to answer the following questions:

1. How applicable are schema matching techniques to compare geospatial standards?
2. What is the usefulness (rapidity and automation) of schema matching techniques?
3. What is the accuracy of schema matching techniques?
4. How do we define similarity levels between geospatial standards?
5. Does LADM propose similar contents with LandInfra and LandXML (what concepts and quantity)?

3. APPROACH

3.1 Literature review on schema matching

Schema matching techniques consist in comparing two schemas in order to identify the similarities (Rahm and Bernstein 2001). Figure 1 illustrates two very simple schemas to compare. We can first notice that no class between both schemas have the same syntax. Maybe a number of concepts have some sort of similarity as being synonyms (like *Spatial Unit* and *Land Division* might be) or have similar sense (like *Building* and *Condominium* might be).

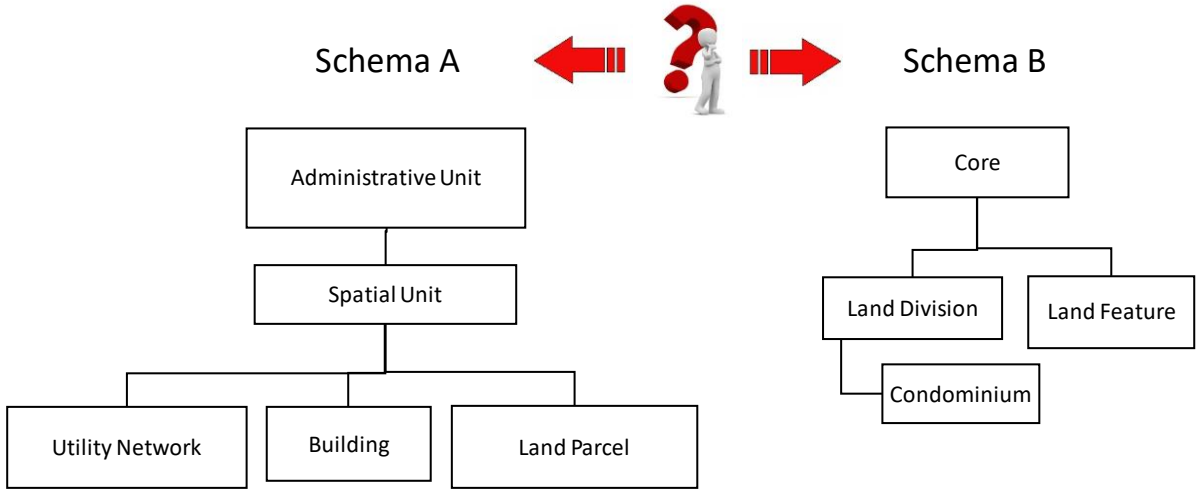

Figure 1 – Example of simple schemas to compare

Schema matching techniques may enable the comparison based on three levels, which can be strategically combined (Casanova et al., 2007; Hossain et al., 2014; Rahm and Bernstein 2001; Shvaiko and Euzenat 2005):

- **Structure level**: Compare the structure of the schema, the hierarchy of classes and attributes. It usually includes data type.
- **Syntactic level**: Compare string by string or group of strings of the words at the level of a language spelling. Acronym is taken into account at this level.
- **Semantic level**: Compare the meaning of the words; it usually requires having access to dictionary, thesaurus, and lexical knowledge base. This level much depends on the quality of the external resources used.

Schema matching techniques are not new (Batini et al., 1987; Milo and Zohar 1998) and it is used in various contexts like data and database integration (Beneventano and Bergamaschi 2007; Ibrahim, et al., 2014; Uluta et al., 2016), data updating (Wang et al., 2015) and in semantic Web and ontologies (Euzenat and Shvaiko 2007; Sala and Bergamaschi 2009).

At a first glance, the comparison between two standards may look simple but it is not (Shvaiko and Euzenat 2005; Hasani et al., 2015). A major obstacle concerns semantics: we can find the same word for referring to two distinct concepts, and distinct words to refer to the same concept. Some concepts are used in a more general way while others are more specific. The same word can be used to describe a class of objects while in another standard the same word will refer to the name of an attribute. As mentioned by Pouliot et al. (2018) and referred by many authors

such as Do and Rham (2007), the application of schema matching becomes a difficult task when schemas are large.

4. COMPARISON OF THE THREE STANDARDS

As mentioned, we performed tests in order to estimate the syntactic and semantic overlap between the LADM, LandInfra (InfraGML) and LandXML. Since all the standards are available in English, we selected this language. The XML schema (XSD) files of LandInfra (InfraGML) and LandXML are available and we can easily have access to them on the respective Web site of the organization. We did not find XSD for LADM (at the conceptual level) and we decided to produce ourselves the XSD file. To perform the comparison, we used the free and open tool OpenII (Open Information Integration), version 2015, developed by MITRE Corporation[4] (Seligman et al., 2010; Smith et al., 2009).

4.1 The XSD standards to be compared

LADM

LADM XSD schema comprises 45 classes to enable the comparison. Figure 2 shows the overall structure of the LADM schema (only some classes are shown).

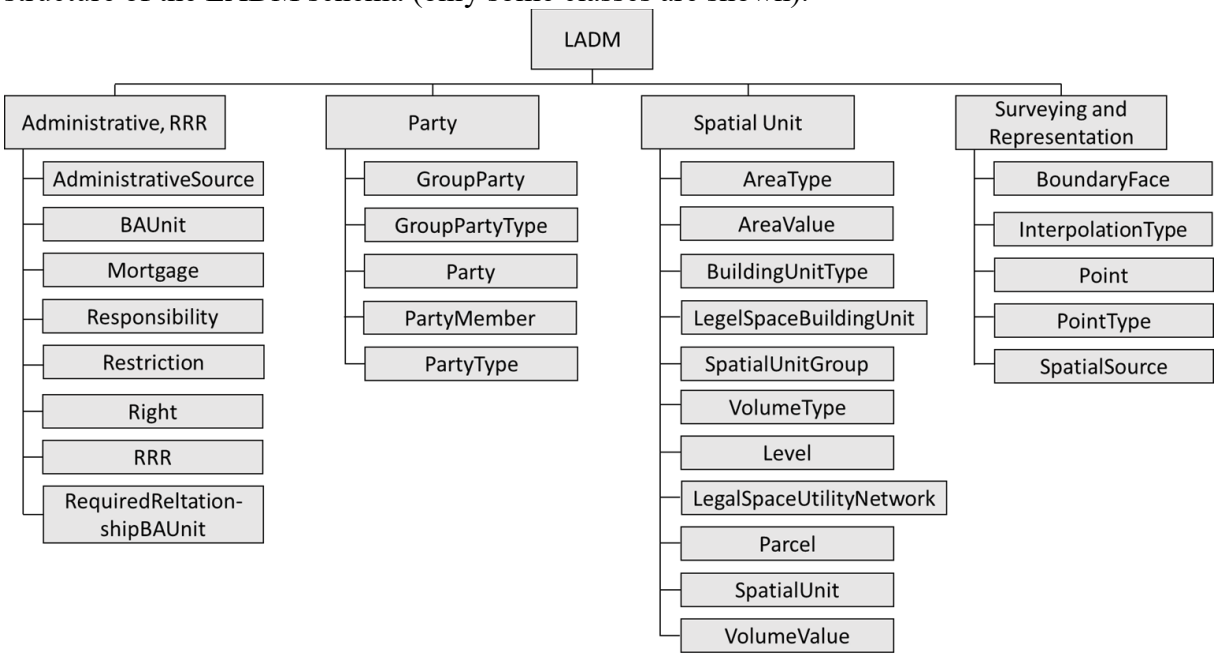

Figure 2 – Hierarchy of classes included in the schema LADM

LandInfra (InfraGML)

LandInfra contains 8 parts: 0 LandInfra Core / 1 LandInfra LandFeatures / 2 LandInfra Facilities and Projects / 3 LandInfra Alignments / 4 LandInfra Roads / 5 LandInfra Railways / 6 LandInfra Survey / 7 LandInfra Land Division. They are modeled in InfraGML with 15 XSD files. The comparison was performed on each XSD files and afterward grouped. It results with 446 distinctive classes that can be compared with the classes of LADM.

[4] http://openii.sourceforge.net/.

LandXML

LandXML2.0 XSD schema comprises 223 classes. LandXML in version 2.0 includes packages as: Alignments / Application / CgPoints / CoordinateSystem / GradeModel / Monuments / Parcels / PipeNetworks / PlanFeatures / Project / Roadways / Surfaces / Survey / Units.

4.2 Overall comparison

The first step in the process of standard's comparison is to perform an overall comparison without the intervention of the user. This will allow us to get an overview of the overlap between the content of the standards. To achieve the overall comparison, OpenII offers two options.

Option 1. Affinity Diagram

The affinity diagram displays associations between members of a generic group (clusters) of schemas. The algorithm used to create the cluster is the TF-IFD (Term Frequency-Inverse Document Frequency) (Sparck Jones 1972). TF-IFD is a weight often used in information retrieval and text mining. Schemas that appear close together may present the most semantical similitude. Figure 3 illustrates the various clusters of the compared schemas. In the upper part, the clusters of the 15 XSD files of InfraGML are shown, which somehow confirm that the XSD files of InfraGML are closer compared with LADM and LandXML. With this first analysis, it is not clear to state if one standard is closer to another. A cluster is proposed between the XSD files of InfraGML and LandXML, which may be perceived to indicate proximity.

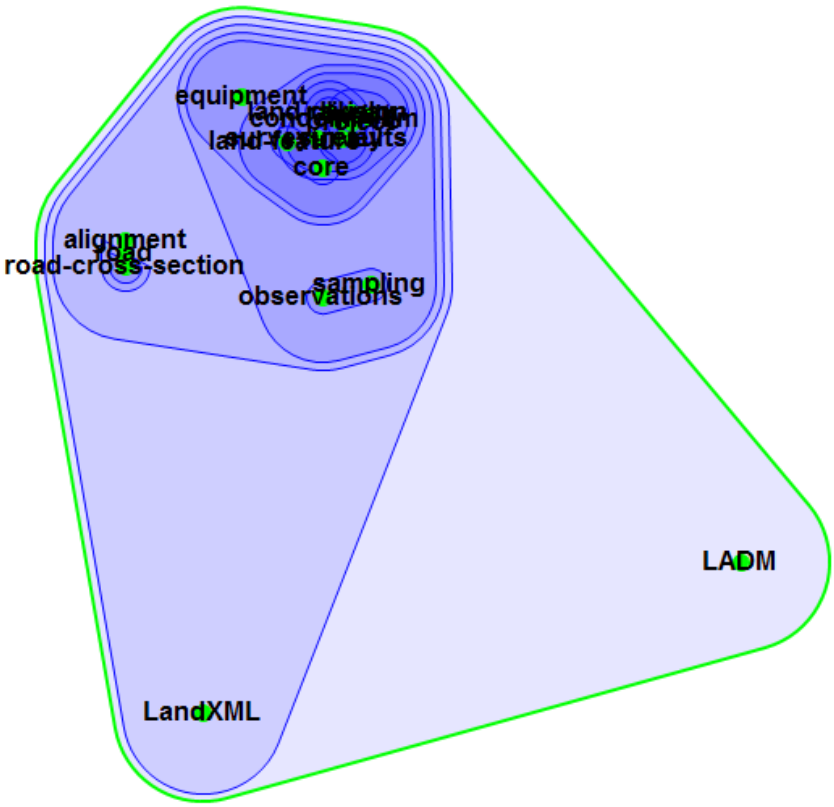

Figure 3 – Affinity diagram explaining the overlap between LADM, LandInfra (InfraGML) and LandXML (based on term frequency)

Option 2. Proximity alignment scores

OpenII also proposes Proximity views and the computation of alignment scores between one schema and others schemas. The alignment score corresponds to the maximum number of overlapping elements (syntax), normalized, between pairs of schemas. Table 1 shows the alignment score between LADM and LandInfra and with LandXML. Since LandInfra contains 15 XSD files, we took the average score. This second overall comparison reveals that LADM is closer to LandInfra (InfraGML) compared with LandXML.

Table 1 - Global comparison of LADM with other geospatial standard schemas

LADM with ...	**Alignment Score**
LandInfra (InfraGML)	0.83 (average)
LandXML	0.55

4.3 Detailed comparison

4.3.1 Levels of similarity

Detailed comparison is performed with what OpenII calls Harmony diagram and users must set various parameters. Figure 4 illustrates Harmony diagram between the XSD of LADM (in the left part) and the XSD of InfraGML-LandDivision (in the right part). The algorithm computes a matching scores (Evidence) varying between 0 and 1 and matching links are added in the diagram. In the figure 4, we highlighted in yellow the link between LA_Parcel (LADM) and LandParcelType (LandDivision), the evidence score was 0.3.

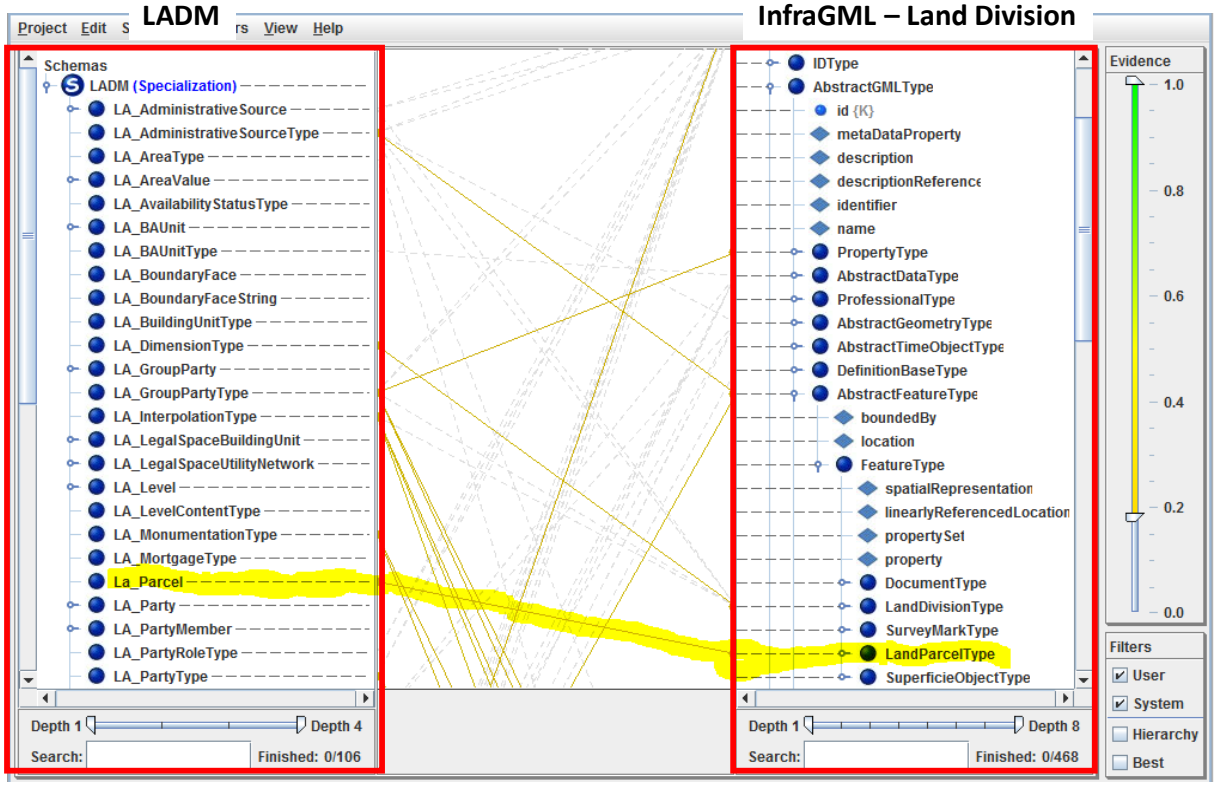

Figure 4 - Example of Harmony diagram that present matching links and matching scores between the XSD of LADM (left part) and the XSD of InfraGML-LandDivision (right part).

The matching scores are furthermore interpreted and classified by our team (this step is done manually outside OpenII). In order to facilitate the interpretation of the results, and to progress in the appreciation of the concept of similarity between standards, we identify three groups of matching scores; what we call level of similarity (Mork et al., 2006; Pouliot et al., 2018):

- **Tightly match**: Matching scores higher than 0.4
- **Loosely match**: Matching scores between 0.2 and 0.4
- **Not match** : Matching scores between 0 and 0.19

This strategy consisting in grouping the matching scores may bias the results, but it was perceived as required since the analysis on a case by case score was long and not effective. It also helped us to converge through the definition of similarity levels and the final decision i.e. matched or not matched.

4.3.2 Options and Parameters

To empower the comparison and calculate the matching score, it exists various options and parameters to set in OpenII. For example, we can decide to compare only the syntax of the class, the syntax of all the terms, the semantic in using external resource for linguistic, etc. Based on the work done in 2018 (Pouliot et al., 2018) and empirical tests, the options used are:

Option 1. Only syntax: Matching score is computed with edit distance between names (only the name of the classes are taken into consideration).

Option 2. Only Semantic: Exploit the sense of the class name and the available relations synonym, hyponym, hypernym of a lexical database. We used Wordnet[5] (Fellbaum 2005; Miller 1995). The matching score is estimated by looking up terminology relationships between what OpenII called a "bag of words" (Mork et al., 2006).

Option 3. Syntax and Semantic (Wordnet): Combination of options 1 and 2.

In using these three options, this will allow us to easily illustrate the advantages of using or not the semantics. Note that only the name of the classes are used at this stage of comparison. Indeed, using the name of the attribute cause lot of confusion since many classes contain the same attribute like ID, NAME or TYPE for example. Therefore, the matching results when using the attributes are less relevant and this is why in this experiment we only present the results based on class name matching.

Also, note that multiple matches are possible (e.g. 1 class of LADM may match with 1 to n classes of LandInfra and 1 class of LandInfra may match with 1 to n classes of LADM). Multiple matches is part of the issues to address when working with matching procedure.

4.3.3 LADM versus LandInfra (InfraGML)

Table 2 presents the number of single matched classes when comparing the XSD of LADM and LandInfra (all the XSD InfraGML) and the single match rates (number of tightly+loosely matched/ number of classes). Table 3, 4 and 5 illustrate a sample of matched classes between LADM and LandInfra for each option.

[5] https://wordnet.princeton.edu/

We can first notice, when only using syntactic option, that 54% of the LADM classes are matched with the classes of LandInfra, among them 0 tightly matched. This first result reveal that the syntax between both standards present some overlap from the point of view of syntax but yet they are pretty distinct. The highest single match rate is obtained with the syntax+semantic options (92%), while 22 LADM classes out of 45 are tightly matched with LandInfra. This matching rate is quite high and its reveals that when using Wordnet a large part of the classes of LADM find a correspondent in LandInfra. The highest score of LandInfra matched classes with LADM is obtained with syntax+semantic option (57%) in which 56 LandInfra classes over 446 find a tightly match with LADM. This is a relatively low rate of matching.

We can see that the matched classes (number of, and content) between option 1 (only syntax) and option 2 (only semantic) are relatively distinct. The semantic option increases the number of matches and in most cases, it brings in the comparison links that appear relevant like:

- *LA_Right->InterestInLandType*
- *LA_SpatialUnitGroup->LandDivisionType*
- *LA_SpatialUnitGroup->AggregationAttributeGroup*
- *LA_LegalSpaceBuildingUnit->InterestInLandType*
- *LA_RequiredRelationshipSpatialUnit->OwnershipAttributeGroup*

In some cases, the semantic option also propose new matches which, at a first glance, looks strange like *LA_LegalSpaceBuildingUnit->TimePositionUnion* or *LA_LegalSpaceBuildingUnit->DirectPositionType*. Using semantic, originated from Wordnet in our case, improves a lot the number of matches compared with only syntax option, but two aspects need to be reminded. Using external resources, as Wordnet, make the results dependant on the completeness and the accuracy of this external resource. Furthermore, we have no real control on Wordnet content, some time the term is included, sometime not. For example, *LA_BAUnit* did not find any relevant match. This term is not syntactically comparable with classes in LandInfra, neither recognized by Wordnet.

Finally, in combining the option syntax and semantic (Wordnet), it generally reduces the number of tightly matched and its reintegrate a number of matches between some classes perceived as relevant, which were surprisingly withdrawn with only the semantics (like *LA_AdministrativeSource->AdministrativeDivisionPropertyType* or *LA_Parcel->LandParcelType*).

Table 2 - Number of matched classes between LADM and LandInfra

	Syntax-Name		Semantic (Wordnet)		Syntax+Semantic (Wordnet)	
	LADM	LandInfra	LADM	LandInfra	LADM	LandInfra
Tightly match	0	0	14	63	22	56
Loosely match	27	75	28	103	24	197
No match	23	371	8	280	4	191
Single match rate	54%	17%	84%	37%	92%	57%

Table 3 – Examples of matched classes between LADM and LandInfra (option 1. syntax)

Level of match	LADM	LandInfra (InfraGML)
Loosely match	LA_AdministrativeSourceType	AdministrativeDivisionPropertyType
Loosely match	LA_ResponsibilityType	CI_ResponsibleParty_Type
Loosely match	LA_AdministrativeSource	AdministrativeDivisionPropertyType
Loosely match	LA_BuildingUnitType	BuildingType
Loosely match	LA_Parcel	LandParcelType
Loosely match	LA_RequiredRelationshipBAUnit	FacilityPartRelationshipType
Loosely match	LA_SpatialUnit	SpatialUnitType

Table 4 – Examples of matched classes between LADM and LandInfra (option 2. semantic)

Level of match	LADM	LandInfra (InfraGML)
Tightly match	LA_Right	InterestInLandType
Tightly match	LA_RightType	InterestInLandType
Tightly match	LA_LegalSpaceBuildingUnit	TimePositionUnion
Tightly match	LA_LegalSpaceBuildingUnit	DirectPositionType
Tightly match	LA_SpatialUnitGroup	LandDivisionType
Tightly match	LA_LegalSpaceBuildingUnit	LandSurfaceType
Tightly match	LA_RequiredRelationshipSpatialUnit	OwnershipAttributeGroup

Table 5 – Examples of matched classes between LADM and LandInfra (option 3. syntax+semantic)

Level of match	LADM	LandInfra (InfraGML)
Tightly match	LA_SpatialUnitGroup	AggregationAttributeGroup
Tightly match	LA_SpatialUnitGroup	SpatialUnitType
Tightly match	LA_LegalSpaceBuildingUnit	TimePositionUnion
Tightly match	LA_Right	InterestInLandType
Tightly match	LA_SpatialUnitGroup	LandDivisionType
Tightly match	LA_SpatialUnit	SpatialUnitType
Tightly match	LA_Parcel	LandParcelType

In order to estimate the accuracy of the matching results, we performed a comparison between our Syntax+Semantic results and the table proposed in the annex D (p.275) of LandInfra official document (OGC 15-111r1, 2016). Table 6 shows the matches proposed by OGC 15-111r1 (2016) (columns LADM and Landinfra), and the results we obtained. We present the results in counting the correct matches, the omission (match omitted), the commission (matched but not relevant of OGC 15-111r1). The number of commission did not necessary reveal problem in our results and might even be perceived as possible links between both standards not reveal in OGC 15-111r1 (2016). For example, we find interesting links as *LA_Responsibility->CI_ResponsibleParty_PropertyType* and not relevant links like *LA_Restriction* loosely matched with *SC_CRS_PropertyType, StringLineSetType, SurfaceSetPropertyType*.

If we sum-up the correct matches and the number of omission, we arrive at a final rate of success of 60% and omission error of 40% on 15 possible matches to verify. This success rate is encouraging and it partly confirm our hypothesis about the ability of schema matching techniques to extract similar content between LADM and LandInfra (InfraGML).

Table 6 – Accuracy assessment based on the annex D of LandInfra (OGC 15-111r1. 2016))

		Number of matches Option 3 (syntax+semantic)		
LADM	**LandInfra**	**Correct**	**Omission**	**Commission**
~LA_RRR (that include LA_Right, LA_Responsibility and LA_Restriction)	7.10.2 InterestInLand	1	3	22
LA_BAUnit	7.10.2.1 PropertyUnit	0	1	2
~LA_Parcel alias LA_SpatialUnit	7.10.2.3 LandParcel	2	0	34
~LA_Restriction	7.10.2.5 Easement	0	1	9
LA_SpatialUnitGroup	7.10.3 AdministrativeDivision	1	1	69
LA_AdministrativeSource	7.10.4 Statement	0	1	8
~LA_SpatialUnit	7.10.6 SpatialUnit	1	0	27
LA_Source	Document (+LI_Source)	1	1	1
LA_MonumentationType	7.10.5 SurveyMonument	1	0	11
LA_LegalSpaceBuilding Unit	7.11.1 CondominiumUnit and building. 7.11.4 BuildingPart,	3	0	69

4.3.4 LADM versus LandXML

Table 7 presents the number of single matched classes that occur in the comparison of the XSD of LADM and LandXML and the single match rates. It is clear that the LADM content is not cover by LandXML (only 10% of their classes find a match). Again, when using the semantic option, it clearly increase the matching rates illustrating that Wordnet bring correlated information in the matching process. The best single match rate is 59% and it corresponds to the percentage of LADM content cover by LandXML with the option semantic only. Similarly to the comparison of LADM with LandInfra, a number of links raise up with the semantic option are relevant while others are not. Table 8 shows examples of matched classes between LADM and LandXML for the option syntax+semantic.

Table 7 - Number of matched classes between LADM and LandXML

	Syntax-Name		Semantic (Wordnet)		Syntax+Semantic (Wordnet)	
	LADM	LandXML	LADM	LandXML	LADM	LandXML
Tightly match	0	0	11	26	3	4
Loosely match	5	4	23	72	22	55
No match	45	219	16	125	25	164
Single match rate	10%	2%	59%	37%	47%	25%

Table 8 – Examples of matched classes between LADM and LandXML (option 3. syntax+semantic)

	Syntax+Semantic (Wordnet)	
Level of match	LADM	LandXML
Tightly matched	LA_LevelContentType	TargetPoint
Tightly matched	LA_LegalSpaceBuildingUnit	Corner
Loosely matched	LA_SpatialUnit	Location
Loosely matched	LA_Point	DataPoints
Loosely matched	LA_LegalSpaceBuildingUnit	FieldNote
Loosely matched	LA_AdministrativeSource	AdministrativeArea
Loosely matched	LA_RequiredRelationshipSpatialUnit	Location
Loosely matched	LA_MonumentationType	Monument
Loosely matched	LA_Parcel	Parcel
Loosely matched	LA_AdministrativeSource	AdministrativeDate

In order to assess the accuracy of these matches, we use the tables proposed by Stubkjaer (2015) that has manually compared the content of LADM and LandXML. Table 9 shows the matches proposed by Stubkjaer (2015) (columns LADM and LandXML) and the number of correct, omission and commission matches we obtained with the option syntax+semantic. With 11 possible matches to verify, we ended with 20% of correctness. This correctness is very low and a large portion of the proposed matches by Stubkjaer (2015) were not detected. As mentioned, the accuracy of the matching depends on the completeness of Wordnet and the convention used for labeling the classes. For instance, *LA_BAUnit* did not find any relevant match (this term only syntactically match with Units in LandXML), but the similarity score was too low (0.004) to reveal any applicable links.

This comparison between LADM and LandXML did not help us to positively prove our hypothesis. Instead, it may even demonstrate the inverse i.e. schema matching techniques are not proposing accurately results when comparing XSD files of standards.

Table 9 - Accuracy assessment based on the proposal of Stubkjaer (2015)

		Number of matches Option 3 (syntax+semantic)		
LADM	**LandXML**	**Correct**	**Omission**	**Commission**
LA_BAUnit	Parcels	0	1	1
LA_Parcel	Parcel	1	0	7
LA_LegalSpaceBuildingUnit	Parcel	0	1	16
LA_LegalSpaceUtilityNetwork	Parcel	0	1	5
LA_PartyMember	Parcel	0	1	0
LA_Restriction (~easement)	Parcel	0	1	2
LA_Monumentation	SurveyMonument	1	0	0
LA_Point	SurveyMonument	0	1	11
LA_SpatialSource	Core::FieldNote	0	1	3
LA_SpatialSource	SurveyorCertificat	0	1	0
LA_AdministrativeSource	SurveyorCertificat	0	1	2

5. DISCUSSION AND CONCLUSION

In this paper, we presented experiments in order to demonstrate the applicability, the accuracy and the usefulness (rapidity and automation) of schema matching techniques applied for the

comparison of three standards as LADM, LandInfra (InfraGML) and LandXML. XSD files were used to map the schema of the standards. For this experiment, three options to run the comparison were tested as syntax, semantic (Wordnet) and syntax+semantic and only the class name (not the attributes, neither the description or the structure) were used in the comparison process. To the best of our knowledge, we are the only authors who propose such a work. Here are the conclusions or lessons to learn we draw:

1. How applicable are schema matching techniques to compare geospatial standards?
   The application of schema matching with XSD files is quite simple but we faced a number of practical issues. First, the XSD files obviously need to be available, which is generally the case for official standards (except for LADM). Second and as expected, the level of detail in the schema modeling between LADM, LandInfra and LandXML varies. For instance, XSD of LandXML encompasses geometry features as curve, metric, symbol, while LADM formalize the features at a conceptual level. Also, a number of features were not correctly modeled in the XSD schema, we had to fix them manually. Third, the size of the schema, or the number of XSD files (local or integrate) to represent one standard require specific procedures and thus increase the processing time. Fourth, and probably the main issue to address, how to interpret the matching scores and converge to a final decision i.e. "do they match or not". For instance, the management of multiple matches (relation n to n) is the main difficulty. Finally, it exists very few tools to run matching procedures; most of them are found as supplementary tools in database management systems and a limited number of algorithms for matching are available (most of them are based on the computation of edit distance between terms). Open II, even though the last release date is 2015 and thus would require up-to-date developments and improve documentation, was an appropriate tool offering a diversity of options. Nevertheless, it would require improvements in order to assist the interpretation of the matching results and to perform some statistical analyses and visual exploration.

2. What is the usefulness (rapidity and automation) of schema matching techniques?
   Regarding the automation, the overall comparison is fast (few minutes) and fully automated. Detailed comparison requires the users to select options for the matching process. In this paper, we presented three options that perform very distinctively. The algorithm used to enable the comparison influences the results. Consequently, once the options are selected, running the comparison and calculating the matching score is rapid and automatic. Yet, the interpretation of the matching score needs to be performed by the users and it may require time and a certain level of tenacity, since the number of matches is quite high (ranging from 22 to 200 matches between LADM and LandInfra). This is why we suggested grouping the matching scores in three levels as *tightly match*, *loosely match* and *no match*. In doing so, we facilitated the interpretation of the matching scores. We were able to generate tables showing the matching links classified as levels of similarity between LADM and Landinfra (InfraGML) and LADM and LandXML.

3. What is the accuracy of schema matching techniques?
   High match rate obviously does not necessary indicate accurate results. The accuracy was assessed in comparing our matching results with independent and trustworthy works of

OGC 15-111r1 (2016) and Stubkjaer (2015). The results in comparing LADM and LandInfra and LandXML are not converging, in one case the accuracy is perceived as good (60% of correctness LADM-LandInfra), while in the other case (LADM-LandXML) the correctness is low (20%). Consequently, it is hazardous to conclude if schema matching techniques offer accurate results. Likewise, the numbers of samples to compare, respectively 15 and 11 classes, are not very high and further comparisons are required. In the literature for schema matching we did not find, similar approaches for accuracy assessment; thereby this is another contribution of our work. The accuracy of our results dependants on three main aspects. First, the selected matching options for this first experiment (only name, only syntax, only semantic Wordnet) largely influence the quality of the results. The grouping strategy, which was required to converge to a decision (match or not), influence the accuracy of the results. Furthermore, having complete description (definition) of elements in the XSD files would be of great help in the matching process, but it was missing in a number of cases. Encouraging people to complete the definition of elements when designing a standard, and the XSD files, is a must.

Second, naming convention largely influences the matching results. If naming principles exist (as proposed by ISO), there is no real agreement and this aspect will always been of concern for matching procedure. Having access to the conceptual name of features and not only the implementation name (often contracted for technical reasons) would considerably improve the quality of the results. In this sense, schema matching could even be perceived as an interesting source of information when selecting the name of classes in the designing process of the standard.

Third, the completeness of the external source used (Wordnet) directly impact the quality of the matching results. When the name of the class and synsets are available in Wordnet, the matching score and the accuracy increase. Participating in populating such lexical databases in our domains of expertise might be encouraged to face this matter.

4. How do we define similarity levels between geospatial standards?
   In this paper, we identified the combination of syntax and semantic as the best manner to establish the comparison of our standards. We also proposed specific thresholds applied to the matching scores resulting from the OpenII Harmony diagram. We suggested three levels of similarity mentioned above as *tightly match*, *loosely match* and *no match*. This notion of level of similarity is a clear contribution of our work.

5. Does LADM propose similar contents with LandInfra and LandXML (what concepts and quantity)?
   A first conclusion can be promptly stated; there is no perfect syntactic match between LADM, Landinfra and LandXML i.e. there are no exact similar terms among the schemas, one of the closest is *LA_Parcel (LADM)* and *Parcel (LandXML)*. Even the term "land" did not find any match since LADM has no classes that contain the term *land*, while LandInfra has nine classes with the term land (as *InterestInLandType, LandCrossSectionPropertyType, LandElementPropertyType*). The are no classes including the term "land" in LandXML. We clearly see that the name convention has a direct impact on the matching score, especially with the syntax option. Even if we know that "LA" means Land Administration, the schema

comparison use the acronym “LA” and try to match it but no match is found. Furthermore, with the syntax option, the longer is the name, the better is the matching score. The level of similarity we proposed tends to reduce this effect. Having integrated the documentation and semantics would help to get better matching results.

We previously present the matching classes between LADM with LandInfra and LandXML. In general, the number of matches between LADM and LandXML is lower compared with LandInfra. For example, 22 LADM classes out of 45 are tightly matched with LandInfra while only 3 are tightly matched with LandXML. With these results, we can now conclude that the content of LADM better matches with LandInfra (InfraGML) compared to LandXML.

Further work

These preliminary experiments were performed in a relative short period of time (less than 2 months, part-time with an internship). Further works are currently planned as:

1) Populate the documentation (definition of classes and attributes) in the XSD files;
2) Complete supplementary accuracy assessment by analyzing the attributes and also the code lists that provide a basis for terminology standards;
3) Consider others standards in the comparison (like INTERLIS (ref : www.interlis.ch);
4) Exploit specific list of keywords and explore incremental schema matching;
5) Use machine learning algorithms applied in Natural Language Processing (NLP) for semantic matching in order to improve the results;
6) Develop our own schema-matching tool and implement alternate matching algorithms.

Acknowledgements

This project was funded by CRSNG RGPIN-2015-05514. We would also thank Alaa Boudhaim, who helped us in the production of the XSD files.

REFERENCES

Batini, C., M. Lenzerini, S.B. Navathe, 1986. A comparative Analysis of Methodologies for Database Schema Integration. ACM Computing Surveys, 18(4), pp.323-364.

Beneventano, D., and Bergamashi, S. 2007. Semantic Search Engines Based on Data Integration Systems. In e. J. C. Idea Group Publishing, IGI Global (Ed.), Semantic Web Services: Theory, Tools and Applications, pp. 317-343.

Casanova, M. A., Breitman, K. K., Brauner, D. F., Marins, A.L.A. 2007. Database Conceptual Schema Matching. Computer, 40(10), pp.102-104.

Do, H.-H., and Rahm, E. 2007. Matching large schemas: Approaches and evaluation. Information Systems, 32(6), pp.857-885.

Euzenat, J. and Shvaiko, P. 2007. Ontology Matching. Springer-Verlag.

Fellbaum, C. 2005. WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, pp.665-670.

Hasani, S., Sadeghi-Niaraki A., Jelokhani-Niarak M. 2015. Spatial Data Integration using Ontology-Based Approach. International Conference on Sensors & Models in Remote Sensing & Photogrammetry, 23–25 Nov 2015, Kish Island, Iran, pp.293-296.

Hossain, J., Fazlida, N., Sani, M., S.A., L., Ishak, I., Kasmiran, K. A. 2014. Semantic schema matching approaches: a review. Journal of Theoretical and Applied Information Technology, 62, pp.139-147.
Ibrahim, H., Karasneh, Y., Mirabi, M., Yaakob, R., & Othman, M. (2014). An automatic domain independent schema matching in integrating schemas of heterogeneous relational databases. Journal of Information Science and Engineering, 30(5), 1505-1536.
ISO 19152. 2012. Geographic Information - Land Administration Domain Model (LADM). International Organisation for Standardisation (ISO), International Standard, Ed 1.
Kumar, K. A Labetski, K. Arroyo Ohori, H. Ledoux, J. Stoter, 2019. The LandInfra standard and its role in solving the BIM-GIS quagmire. Open Geospatial Data Software and Standards, 4:5, pp.1-16.
LandXML 2.0, 2016. Change for LandXML Schema Version 2.0. January 19, 2016. The Land Development and Transportation Industry, LandXML Organization: http://www.landxml.org.
Miller, G. A. 1995. WordNet: a lexical database for English. Communication ACM, 38(11), pp.39-41.
Milo, T., and Zohar, S. 1998. Using schema matching to simplify heterogeneous data translation. International Conference on Very Large Databases, New York, USA — August 24-27, pp.1-21.
Mork, P., Rosenthal, A., Korb, J., Samuel, K. 2006. Integration Workbench: Integrating Schema Integration Tools. 22nd International Conference on Data Engineering Workshops (ICDEW'06), 3-7 April, Atlanta, Georgie.
OGC 15-111r1. 2016. Land and Infrastructure Conceptual Model Standard (LandInfra). Open Geospatial Consortium (OGC®) Implementation Standard, V1.
OGC InfraGML 2017. OGC InfraGML 1.0: Part 0 – LandInfra Core – Encoding Standard 2017. Internal reference number of this OGC® document: 16-100r2.
Pouliot, J., C. Ellul, S. Larrivée, A. Boudhaim, 2018. Exploring Schema Matching to the Selection of Geospatial Standard: Application to Underground Utility Networks. 13th International Conference on 3D GeoInformation, Delft, Netherlands, 1-2 October.
Rahm, E., and Bernstein, P. A. 2001. A survey of approaches to automatic schema matching. The VLDB Journal, 10(4), pp.334-350.
Sala, A., and Bergamaschi, S. 2009. A Mediator-Based Approach to Ontology Generation and Querying of Molecular and Phenotypic Cereals Data. Int. Journal of Metadata and Semantics Ontologies, Special Issue on Agricultural Metadata and Semantics, 4(1), pp.85-92.
Seligman, L., Mork, P., Halevy, A., Smith,K., Carey, M.J., Chen K., Wolf C., Madhavan, J., Kannan, A. 2010. OpenII: An Open Source Information Integration Toolkit. SIGMOD '10, June 6-11, 2010, Indianapolis, Indiana, USA.
Shvaiko, P., and Euzenat, J. 2005. A survey of schema-based matching approaches. J. Data Semantics IV 3730, pp.146-171.
Smith, K., Mork, P., Seligman, L., Rosenthal, A., Morse, M., Allen, D.M., Li, M. 2009. The Role of Schema Matching in Large Enterprises, Conference on Innovative Database Research (CIDR '09), Jan.
Stubkjaer, E. 2015. A comparison of LandXML, ISO 19152:2012 LADM, and the draft LandInfra conceptual model. Geospatial-BIM-Indoor Workshop, Boulder, Colorado, June 2th.

Uluta, D., Kara, G., & Comert, C. 2016. Semantic definition and matching for implementing national spatial data infrastructures. Journal of Spatial Science, 61(2), 441-459.
Wang, Y. H., Zhang, H. B., & Xu, J. 2015. A Survey of Applications and Researches on Schema Matching between GIS Spatial Data (Vol. XL-7/W4).

BIOGRAPHICAL NOTES

Jacynthe Pouliot is a full professor at the Department of Geomatics Sciences and vice-dean of research at the Faculty of Forestry, Geography and Geomatics, at Université Laval, Quebec, Canada. Her main interests are the development of GIS systems, the application of 3D modeling techniques in the domain of cadastre, and the integration of spatial information and technologies. She has been a member of the Professional association of the Quebec land surveyors since 1988.

Claire Monney is a student in geomatics and land management at the School of Business and Engineering Vaud, University of Applied Sciences Western Switzerland, and completing a summer internship at the Department of Geomatics Sciences at Université Laval, Quebec, Canada.

Jens Ingensand is a full professor at the Territorial Engineering Institute (insit) at the Department of Environment, Construction and Geoinformation at the School of Business and Engineering Vaud, University of Applied Sciences Western Switzerland (HEIG-VD; HES-SO). His main interests are the development and the evaluation of geospatial applications and processes, as well as citizen science and visualization techniques such as augmented and virtual reality.

Suzie Larrivée is a research professional at the Department of Geomatics Sciences at Université Laval, Quebec, Canada. Her expertises are GIS, spatial databases, spatial data quality, spatial data integration, and geospatial business intelligence (GeoBI).

CONTACTS

Professor Jacynthe Pouliot a.-g.
Université Laval
Casault Building, office 1325
1055 avenue du Seminaire,
Quebec (Quebec), Canada, G1V 0A6
Phone: +1 418 656-2131, ext. 402146
Jacynthe.pouliot@scg.ulaval.ca
https://www.scg.ulaval.ca/jacynthe-pouliot/

Claire Monney
GIS Laboratory University of Applied Sciences Western Switzerland (HES-SO)
School of Business and Engineering Vaud (HEIG-VD)
claire.monney@heig-vd.ch

Professor Jens Ingensand
GIS Laboratory University of Applied Sciences Western Switzerland (HES-SO)
School of Business and Engineering Vaud (HEIG-VD)
jens.ingensand@heig-vd.ch

Suzie Larrivée
Université Laval
Casault Building, office 1353
1055 avenue du Seminaire,
Quebec (Quebec), Canada, G1V 0A6
Suzie.Larrivee@scg.ulaval.ca