# EXPLORING SCHEMA MATCHING TO COMPARE GEOSPATIAL STANDARDS: APPLICATION TO UNDERGROUND UTILITY NETWORKS

J. Pouliot[1], S. Larrivée[1], C. Ellul[2], A. Boudhaim[1]

[1] Dept. of Geomatics, Université Laval, Québec, Canada (jacynthe.pouliot; suzie.larrivee; alaa.boudhaim.1)@ulaval.ca
[2] Dept. of Civil, Envir. and Geomatic Engineering, University College London, Gower Street, London, UK – c.ellul@ucl.ac.uk

**KEY WORDS:** Schema matching, geospatial standard, underground utility network, CityGML UtilityNetwork ADE, IFC, InfraGML.

**ABSTRACT:**

This paper proposes a preliminary analysis of whether a schema matching approach can be applied for the comparison and possible the selection of geospatial standards. Schema matching is tested in the context of underground utility network modelling and, as an initial experiment, three geospatial standards are compared with user requirements: CityGML UtilityNetwork ADE, infraGML and IFC. The schema comparison is enabled by XSD files, and carried out from syntactic, structural and semantic points of view, making use of existing software. The findings of this preliminary investigation show that schema matching is applicable for the comparison of user needs and existing geospatial standards, and does show some potential, but the matching results are varied and not easy to interpret. In particular, the similarity scores between user needs and standards are very low and the comparison and the selection is not straightforward. Having a strategy - an iterative process - is required. While for this preliminary examination, the focus of this paper is on assessing the schema matching approach (which parameters to take into consideration, how to proceed, tools available, automation aspect), further work will include examining software options and performance, as well as exploring how to take the relatively complex preliminary results obtained here and use them to assist the selection of a specific standard.

## 1. INTRODUCTION

### 1.1 Context

Given global trends towards, and the importance of, sharing spatial data and applying standardized and semantic modelling, we can find a large variety of geospatial standards, mainly proposed by OGC and ISO. This makes it more complicated to compare and then select a standard for spatial data modelling that best fits the needs for a specific application, in particular for users with little expertise in this domain.

For example, if the targeted features are underground infrastructures, which conceptual model should underpin the spatial model? This issue was clearly demonstrated during the 2017 workshop organized by the OGC Underground Concept Development Study[1]. A quick review of standards proposed by OGC and ISO reveals at least nine sources of information possibly suitable to model underground networks:

- Land and Infrastructure DWG/SWG with LandInfra (OGC 15-111rl). InfraGML is the GML implementation version.
- CityGML (OGC 12-019). Utility network ADE and tunnel ADE are CityGML-ADEs specifically related to underground infrastructures.
- PipelineML SWG.
- Energy and Utilities DWG.
- 3Dim DWG.
- ISO 19107 Geographic information.
- ISO 19152 Geographic information - Land Administration Domain Model (LADM).
- ISO 16739 Industry Foundation Classes-IFC. IFC-Infra is a research initiative to standardize BIM for infrastructure, among them underground infrastructures.
- INSPIRE.

[1] http://www.opengeospatial.org/projects/initiatives/undergroundcds

### 1.2 Problem statement

The time required to read the documentation associated with each standard is significant (for example the document for CityGML OGC 12-019 has 344 pages), and selecting a standard requires the reader to understand all of the standards among which a user may select. Even for specialists in data modelling, this is a huge task.

In comparing some documentation related to geospatial standards, we also observe that for a concept that looks similar (e.g. network features), the heterogeneity in terms of structure and meaning is surprisingly high. We find the same word referring to two distinct concepts, and distinct words referring to the same concept. We also notice inconsistencies in the hierarchical relationships between concepts. Some concepts are used in a more general way while others are specific. The same word can be used to describe a class of objects while in another standard it will refer to an attribute name. Consequently, it becomes complex to understand a standard, to compare its content with others, and finally to decide which geospatial standard best fit the needs.

### 1.3 Objectives

As geospatial standards are by definition using formal description and language and may contain similar concepts, we hypothesize that automatic schema matching is a valuable approach to compare geospatial standards with user needs. A schema is the formal description of the arrangement of classes, attributes, domain of values and relationships between classes,

(Rahm and Bernstein 2001). Schema matching is an extensive research field and it plays a central role in the context of database and data integration, metadata management and semantic Web (Bellahsene et al., 2011). Finding schemas for geospatial standards is straightforward (many standards are described using XML Schema Descriptors, XSD). However, as will be seen in Section 2, schema matching is mainly designed and exploited by domain experts, not fully automatically performed, and as far as we know, this technique is not exploited in the context of comparing and selecting geospatial standards.

Therefore, the objective for this first phase of the study is to explore schema matching techniques, and to determine whether the outputs of these techniques can in turn facilitate the comparison of, and eventually the selection of, geospatial standards, as a first step towards a solution to help organisations to select an appropriate standard.

More specifically, this paper illustrates the concrete application of schema matching approaches to the specific problem of selecting a geospatial standard suitable to modelling underground utility networks (UUN), such as water and sewer pipes and valves, gas conduits, communication cables. For this preliminary exploration of schema matching techniques, the paper illustrates the results obtained when comparing user needs and three geospatial standards: CityGML Utility Network ADE, InfraGML and Industry Foundation Classes (IFC). In this context, the paper tries to answer the following questions:

- Is schema matching applicable for the comparison of user needs and existing geospatial standards, in particular for the non-expert user?
- How to apply XSD schema matching (what are the key parameters to consider)?
- What are the lessons to learn from schema matching applications (in terms of modus operandi, the quality of the results, in replicability of this work)?

Our work and further investigation may also contribute to users and designers of geospatial standard in formalizing user needs as XML Schema, in revealing the overlapping in standard's offers, in stimulating the communication and exchange of information and knowledge.

## 2. LITERATURE REVIEW

Schema matching is not new (Miller 1995; Milo and Zohar, 1998) and it occurs either manually or automatically depending on the applications and the complexity of the system under study and the performing vary much (Benerecetti et al., 2005). Schema matching can be applied at the schema level solely or at the instance-level- i.e. the data itself (Rahm and Bernstein 2001; Yu-hong et al., 2015). For this preliminary exploration, our research project is only interested in the schema level since we assume that it will be more convenient and consistent when comparing standards and user requirements. Consequently, the following review will mainly focus on this aspect.

### 2.1 Overview of schema matching

Schema matching is the process of comparing two schemas and producing possible mappings between elements that correspond (Do et al., 2003; Doan 2002; Rahm and Bernstein 2001). Figure 1 illustrates, in a simple manner, the overall principle of schema matching.

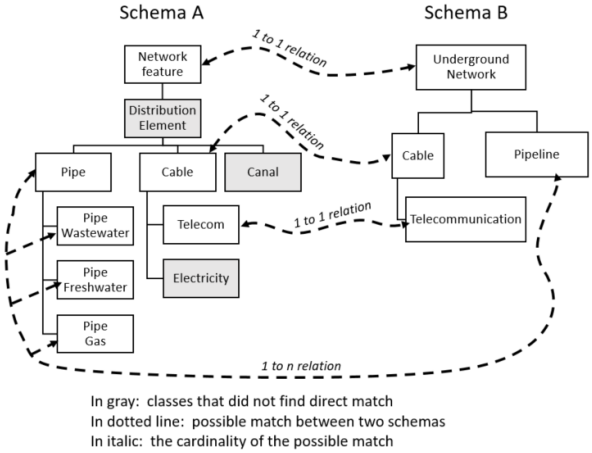


Figure 1 – A simple example of possible matches between two schemas.

Three levels of schema comparison can be identified (Casanova et al., 2007; Hossain et al., 2014; Rahm and Bernstein 2001; Shvaiko and Euzenat 2005):

- **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.
- **Structure level**: Compare the structure of the schema, the hierarchy of classes and attributes. This usually includes data types.
- **Semantic level**: Compare the meaning of the words; this usually requires having access to dictionary, thesaurus, and lexical knowledgebase. This level is dependent on the quality of the external resources used.

The three levels can be strategically combined and the complexity may differ inside each level (Rahm and Bernstein 2001).

### 2.2 The concept of similarity

A similarity index may be produced by different heuristics and computational techniques depending on the level of comparison used (Chen et al., 2012, Fan et al., 2016; Rada et al., 1989; Smiljanic 2006). The similarity measures usually ranges from 0 for fully distinct objects to 1 being assigned to a match. Intermediate values can be obtained for example by semantic distance (e.g. 1 = synonym, 0 = antonym) with intermediate values based on semantic path weight distance (Lin, 1998), path cost (e.g. exploiting the hierarchy order of strings in the tree parent-child) or string matching (see next paragraphs). Various relationship cardinalities between matched candidates also exist (1 to n or n to n), and in this case, the similarity distance exploits the frequency (counting co-occurrence of terms in search patterns) and generates a normalized similarity (range between 0 and 1). At the end of the comparison process, the matching approach will typically aggregate (combine) local similarity measures into one global indicator of similarity; distinct functions can be used (Euzenat and Shvaiko 2013; Peukert et al., 2010).

String Matching
Most commonly, schema matching relies on string matching techniques with edit distance is the most common and basic method (Cohen et al., 2003 and Navarro, 2001 offer an exhaustive comparison of those techniques). The principle of

string matching is to compute the number of operations (single character insertion, deletion or substitution) required for transforming one string to another. A number of extensions of this edit distance approach can be found in literature (Algergaway et al., 2010; Do et al., 2003; Tiwari and Trivedi 2012), with one of the most common being the Levenshtein distance (insertions, deletions, substitutions).

The techniques of string matching and linguistic can be applied for the name of classes, attributes and domain of values, or in exploiting the annotation or documentation part of schema. In this last case, the frequency of a search term in both schemas is the similarity measure mostly commonly used (Algergawy et al., 2010; Cohen et al., 2003; Sorrentino et al., 2011; Yi et al., 2005). The semantic level will go a step further considering the position of a word within a sentence as a given significance, and in using lexical annotations that assign a certain meaning to a word (Giunchiglia and Yatskevich 2004; Hossain et al., 2014; Li et al., 2003; Martinez-Gil and Aladan-Montes 2013; Yi et al., 2005).

### 2.3 Use of external resources

A step further in the comparison, which could be applied at the syntactic or semantic levels, is the use of external resources such as linguistic resources, thesauri and taxonomies, lexical database or formal ontologies (Euzenat and Shvaiko 2013; Fan et al., 2016; Hossain et al., 2014; Rahm 2011). For instance, Wordnet is a well-known lexical database in which the distinct ways of expressing the same concept –based on the meaning - is described from a pre-defined set of nouns, verbs, synonyms, etc (Fellbaum 1998; Miller 1995). Such systems usually recognize hierarchical and non-hierarchical relationships between the match candidates. Figure 2 illustrates a simple example of the content of Wordnet for the term "network". It shows that the word "network" as a noun refers to five general meanings. We highlight the direct hypernym of network#2.


Figure 2 –Wordnet search for the term "network" (extracted from http://wordnetweb.princeton.edu/perl/webwn).

### 2.4 Limitations of Schema Matching

The schema matching approach also presents some limitations in terms of the size and the complexity of the schema (going beyond a 1-1 match relation), which can be a resource bottleneck. Additionally, schema matching is not easily adaptable to a specific domain and the users' choices through the process make it subjective. The process is difficult to totally automate (Doan 2002; Hossain et al., 2005; Rahm 2011; Smiljanic et al., 2006). One possible avenue to improve schema matching efficiency is to reduce the number of candidates and the complexity of the schema; a technique called clustering (Do and Rahm, 2007).

## 3. METHODOLOGY

As noted in Section 2, a large number of parameters and factors influence the application and performance of schema matching techniques. In this first study, we are mainly interested in exploring all three levels of comparison (syntax, structure and semantic) and assessing their value in the context of comparing user needs and geospatial standards. Our work is dependent on existing and available online information regarding geospatial standards.

After reviewing documentation related to the geospatial standards listed above, it is observable that most of them (if not all), use reports (text files) and UML formalism (packages and class diagrams) to graphically show the content of the standard. Furthermore, in the majority, the standard is presented as structured XML Schema (XSD) or if not, the retro-engineering conversion from XML files or UML packages to XSD is feasible. An XML Schema Definition (XSD) is a W3C recommendation (https://www.w3.org/TR/xmlschema11-1/) and somehow a "de facto" standard for describing XML documents. XSD expresses in text format the exact structure, content and definition of the documents (not the data itself), and given its common use this was an appropriate selection of input format for schema matching.

Before running any schema matching, the language also has to be settled. In our experiment, English language is selected since it offers a larger diversity of standards available in XSD format. The English language also allows us to take advantage of existing and valuable English lexical databases such as Wordnet.

The overall approach is organised as follows:

**Step A. Identify and formalize user requirements.** The formalization consists in declaring, as an XML schema, the user requirements: (a) classes of objects, (b) attributes and domains of values, (c) possible relationships between classes and (d) short definition of relevant items (called annotation in XSD). This can be done manually, by import/export functionality from data modeller tools or by retro-engineering if a database already exists. The user needs are then formalized as one XSD schema (called the global user schema-GUS).

**Step B. Retrieve the XSD of the geospatial standards** (called the geospatial standard schema-GSS). Some standards propose more than one XSD file and selection may be required.

**Step C. Select the schema matching tool.** We are interested in assessing the three levels of comparison (syntax, structure and semantic), consequently the tools for schema matching will have to enable these options. With regards to the external

sources to improve semantic matching, we decided to use Wordnet, since it is the most well-known and used. After considering a number of options, we selected the schema matching tool OpenII, http://openii.sourceforge.net/ (Seligman et al., 2010; Smith et al., 2009). OpenII (Open Information Integration tool suite) is developed by MITRE Corporation and proposes a free and open solution for matching schemas as XSD files.

**Step D. Perform the schema matching process.** The schema comparison is achieved between the GUS and the GSS. A number of tests are performed to assess the schema matching tool and approaches.

## 4. APPLICATION TO UNDERGROUND UTILITY NETWORKS

### 4.1 Global User Schema (GUS)

As noted above, the first step in schema matching consists of defining the user requirements and transform them into an XML Schema. For the purpose of this initial experiment, we used the requirements of a municipality with whom we have previously worked on data modelling of their underground utility network (Coté and Boucher 2016). The users provided a list of classes and attributes (attributes are presented in []):

- Damage [claim ID, provider, damage date, address, infrastructure type, status, geometry]
- Delivery pipe [ID, owner, serial number, type, status, diameter (mm), length (m), depth (m), slope, equipment, installation date, repair date, geometry, pumping station ID]
- Floor lamp [ID, type, model, power watt, geometry, streetlight cable ID]
- Gas pipe [ID, serial number, type, status, diameter (mm), length (m), depth (m), slope, equipment, installation date, repair date. owner, geometry]
- Hydro network [ID, depth (m), length (m), installation date, repair date, geometry]
- Manhole sewer [ID, geometry, sanitary pipe ID, storm leads ID]
- Pumping station [ID, name, installation date, geometry, sanitary pipe ID]
- Sanitary pipe [ID, owner, serial number, type, status, diameter (mm), length (m), depth (m), slope, equipment, installation date, repair date, geometry, sewer junction ID1, sewer junction ID2]
- Sewage sump [ID, geometry, storm leads ID]
- Sewer junction [ID, geometry]
- Standpipe [ID, diameter (mm), brand, model, flow available, installation date, geometry, waterworks leads ID]
- Storm leads [ID, owner, serial number, type, status, diameter (mm), length (m), depth (m), slope, equipment, installation date, repair date, geometry]
- Streetlight cable [ID, provider, type, depth (m), length (m), installation date, geometry]
- Telecommunication cable [ID, provider, type, depth (m), length (m), installation date, geometry]
- Water valve [ID, type, brand, owner, model, diameter (mm), depth (m), installation date, pressure, geometry, waterworks leads ID]
- Waterworks leads [ID, owner, serial number, type, status, diameter (mm), length (m), depth (m), slope, equipment, installation date, repair date, geometry]

To enable the comparison with the standard, we constructed the XSD schema of the user (GUS) from this list. The GUS contains 10 root elements, 5 hierarchical depths (or levels) for a total number of 169 possible elements to be compared.

### 4.2 Global comparison

The second step in the approach is to collect XSD of geospatial standards that include features about underground utility networks. These are widely available, and we decided for this first experiment to use the following:

- **CityGML UtilityNetworkADE** is an application domain extension (ADE) of the CityGML standard for the modelling of utility network (Kutzner and Kolbe 2017). This ADE is still under development. We used the latest version of the ADE (http://www.citygmlwiki.org/index.php/CityGML_UtilityNetworkADE). This XML schema contains 5 XSD files. For example, NetworkComponents XSD has 27 root elements, 3 hierarchical depths for a total number of 71 possible elements to be compared.
- **InfraGML** is the GML implementation version of Land and Infrastructure LandInfra (OGC 15-111rl). It models objects including civil engineering infrastructure facilities (e.g. UUN) and land. We used version 1.0 http://schemas.opengis.net/infragml/. The InfraGML schema contains 15 XSD files. As an example, the Core XSD has 103 root elements and 6 hierarchical depths for a total number of 290 possible elements to be compared.
- **IFC** (Industry Foundation Classes) from buildingSMART is a specification for *Building Information Modeling* (BIM) data. This standard is well known and used in the building construction or facility management projects. We used the last version of the specification IFC4 Add2 (http://www.buildingsmart-tech.org/specifications/ifc-releases/ifc4-add2). It contains only one XSD file and has 893 root elements, 10 hierarchical depths for a total number of 1874 possible elements to be compared.

Before running a detailed comparison, we first conducted a global evaluation between GUS and all the XSD of geospatial standard. OpenII includes Proximity views which allow an overall examination of the alignment of one single source (GUS in our case) and other schemas (standards). Table 1 shows some of the results. The alignment scores correspond to the maximum values of the similarity scores (either computed with edit distance or the number of similar words found in the documentation of both schemas). It is a high-level correspondence or weak semantic matching but can be used to reveal interesting elements for further investigation and discussion. This first evaluation shows that the best alignment of GUS is obtained with InfraGML, followed by IFC. However, the alignment scores are dependent on the number of elements available to compare and these results should therefore be interpreted with caution.

Table 1. Overall comparison of global user schema (GUS) and geospatial standard schemas

| **Global comparison GUS with ->** | **Alignment Score** |
|---|---|
| InfraGML Core | 0.65 |
| InfraGML Road cross-section | 0.59 |
| InfraGML Land feature | 0.55 |
| IFC | 0.55 |
| ADE NetworkComponents | 0.45 |
| ADE UtilityNetworkProperties | 0.41 |
| ADE NetworkCore | 0.32 |
| InfraGML Condo | 0.20 |
| ADE FeatureMaterial | 0.16 |
| ADE UtilityHollowSpace | 0.02 |

OpenII also proposes Affinity Diagrams that find associations between members of a generic group of schemas. It creates clusters without the intervention of the user based TF-IFD (frequency-inverse document frequency) that calculates the

frequency of terms in a schema (Sparck Jones 1972). Schemas that appear close together may present the most semantical similitude. We found in this second comparison that GUS is closer to CityGML UtilityNetworkADE, in particular with the schema NetworkComponents. The terms *slope* and *diameter* are the most frequent terms found.

This global comparison gives a first overview of the semantical overlaps between the schemas and already reveal relevant information about the proximity of standards and the GUS. It allows us to rank the standards based on their overall semantic similarity.

### 4.3 Detailed Comparison

To perform a more detailed comparison and because OpenII is running comparisons between pair of schemas, we decided to reduce the number of XSD files to process. We selected one XSD files per standard - the one that best matched the GUS in the global examination. Reducing the number of schemas to compare also helped us to keep the focus on schema matching options and the interpretation process. In the following sections, the comparison is thus accomplished between GUS and Utility Network ADE NetworkComponents, InfraGML Core and IFC.

Furthermore, for reasons of simplicity and space, we do not present all the matching options and combinations offered by OpenII. We will discuss only one matching option per level of comparison (syntax, structure, semantic). In the following tables, we present the five best scores for each option. To compute those matching scores, we used the Harmony views option of OpenII that matches pairs of schemas. The results show both hierarchical schemas and the links of match found (see Figure 3 for one example).

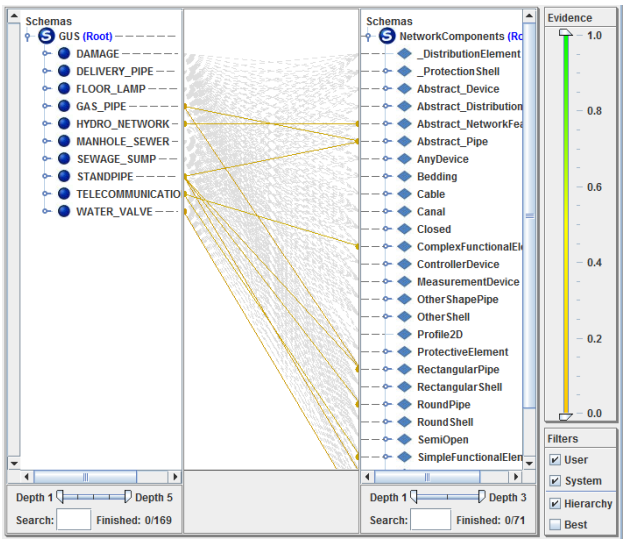

Figure 3. Example of the Harmony interface of OpenII

#### 4.3.1 Syntax comparison

A syntax comparison is achieved within OpenII by using the option "Name", and computes similarity score based on string comparison. The following tables show partial matching elements and scores. In the tables, *Type* may be interpreted as classes of entity while *Property* corresponds to attribute. The score refers to similarity scores computed with *Edit Distance*. With this option, NetworkComponents has 221 matches, InfraGML has 882 matches, while IFC has 4092 matches. This number of results is complicated to interpret.

For information, if we run a name matching on the exact same schemas, we obtain similarity scores ranging from 0.1 to 0.5 while the best matching scores visually identified were located close to 0.4 to 0.5. Scores in this latter range may be then interpreted as excellent similarity scores.

We can see in Tables 2, 3 and 4 that the similarity scores are in most cases very low. This may be due to the fact that every character is counted in this comparison including "_" or spaces. For example for NetworkComponents, we could have expected higher scores for the property *diameter* that exists in both schemas, but because of the presence of special characters, the score is lower when compared with *status*. The score for the property *slope* is a bit lower compared with *status* because there are fewer characters to compare. In general, in the matching results, we notice that the property largely influences the results. Since the same property may belong to distinct types, we can conclude that the result of name matching is weakly relevant.

Table 2. Name matching between GUS-NetworkComponents

| GUS Type (property) | ADE Network Type (property) | Score |
|---|---|---|
| Telecommunication_cable | Cable (isCommunication) | 0.29 |
| Delivery_pipe (status) | Abstract_NetworkFeature (status) | 0.23 |
| Gas_pipe (status) | Abstract_NetworkFeature (status) | 0.23 |
| Delivery_pipe (slope) | Canal (slope) | 0.20 |
| Gas_pipe (diameter_mm) | RoundShell (exteriorDiameter) | 0.12 |

Table 3. Name matching between GUS-InfraGML

| GUS Type (property) | InfraGML Type (property) | Score |
|---|---|---|
| Damage (infrastructure_type) | AbstractCurveType | 0.26 |
| Damage (infrastructure_type) | AbstractSurfaceType | 0.20 |
| Delivery_pipe (type) | ReferentType (type) | 0.17 |
| Floor_lamp (type) | ReferentType (type) | 0.17 |
| Damage (infrastructure_type) | LandInfraDatasetPropertyType | 0.17 |

Table 4. Name matching between GUS-IFC

| GUS Type (property) | IFC Type (property) | Score |
|---|---|---|
| Damage (infrastructure_type) | IfcSpatialElementTypePropertyHolder (IfcSpatialStructureElementType) | 0.28 |
| Telecommunication_cable- | IfcFlowTerminalPropertyHolder (IfcCommunicationsAppliance) | 0.26 |
| Damage (infrastructure_type) | IfcConstructionResourceType | 0.23 |
| Delivery_pipe (geometry) | EntityPropertyHolder (IfcGeometricSet) | 0.23 |
| Gas_pipe (geometry) | EntityPropertyHolder IfcGeometricSet) | 0.23 |

We can also use a Group By Type option and run the matching process again. The following tables present some of the results. With this option, NetworkComponents has 17 matches, InfraGML has 20 matches, while IFC has 74 matches.

When grouping by Type, the number of possible matches is much reduced. Additionally, while the matching process does not result in higher scores, it allows us to first match entities that have similar name. In this way, it is much easier for the users to identify relevant matches, and then progress to finding further candidate matches in using attributes. Even so, this approach is constrained by the string matching rules used. For example, in Table 5, we can clearly see the match between *Sewer_Junction* and *SimpleFunctionalElement* results from the common character sequence *UNCTION*; which is not a pertinent match in this case. For the InfraGML schema, the matching scores are very low, and are not strong enough to infer any similarity with GUS. This observation contradicts the first result in the global comparison.

Table 5. Group By Type matching between GUS-NetworkComponents

| GUS Type | Network ADE Type | Score |
|---|---|---|
| Gas_pipe | Abstract_Pipe | 0.13 |
| Hydro_network | Abstract_NetworkFeature | 0.10 |
| Sewer_junction | SimpleFunctionalElement | 0.10 |
| Standpipe | Abstract_Pipe | 0.10 |
| Sanitary_pipe | Abstract_Pipe | 0.09 |

Table 6. GroupBy Type matching between GUS-InfraGML

| GUS Type | InfraGML Type | Score |
|---|---|---|
| Telecommunication_cable | ObjectIdentification | 0.06 |
| Standpipe | DistanceExpressionType | 0.05 |
| Telecommunication_cable | LinearlyReferencedLocationType | 0.04 |
| Waterworks_leads | LateralOffsetDistanceExpressionType | 0.03 |
| Pumping_station | SpatialRepresentationType | 0.03 |

Table 7. GroupBy Type matching between GUS-IFC

| GUS Type | IFC Type | Score |
|---|---|---|
| Telecommunication_cable | IfcCommunicationsAppliance | 0.26 |
| Telecommunication_cable | IfcCommunicationsApplianceType | 0.25 |
| Standpipe | IfcSectionedSpine | 0.04 |
| Water_valve | IfcWasteTerminalType | 0.04 |
| Floor_lamp | IfcCooledBeamType | 0.03 |

#### 4.3.2 Structural comparison

OpenII offers the possibility to identify the exact same hierarchical naming of elements all of the way to the root (parent-child relationships). However, in this case the matcher produces no matching results for all standards. This illustrates the restrictive action when using structural hierarchy. In fact, structural comparison is most appropriate when schemas come from the same source or share lot of content. It should also be noted that the option GroupBy used in the previous section somehow considers the schema structure as part of its process.

#### 4.3.3 Semantic comparison

OpenII offers semantic comparison with the options *Thesaurus*, *Documentation* and *Wordnet*. With these options, instead of comparing characters, the software compares, without regard to order or syntax, the word and its meaning (if lexical information is available). The similarity is estimated by looking up terminology relationships between what they called a "bag of words" (Mork et al., 2006). In our experiment, we used Wordnet Thesaurus. Typically, the similarity is fixed to 1 if two terms are set as synonym, at 0 if they are antonym and intermediate values are computed based on semantic path weighted distance.

Tables 8, 9 and 10 present the results for schema matching with the option Wordnet (with GroupBy option on Type for ease of comparison with previous results). With this option, NetworkComponents has 24 candidate matches, InfraGML has 54 candidate matches while IFC has 174 matches.

Compared with the Groupby option, the total number of matches with Wordnet is higher; this is expected since the external source augments the amount of information to compare. The augmentation rate is higher with IFC. This may be explained by the fact that IFC contains many more elements and these are possibly more prevalent in Wordnet. Additionally, the candidate matches are not identical to previous results. Since the comparison is carried out with the support of Wordnet matching rules, it is somehow difficult to control the matches and the results are determined by the completeness of Wordnet for relevant themes. Some of the matches are difficult to explain e.g. most of the candidates of InfraGML.

Table 8. Wordnet matching between GUS-NetworkComponents

| GUS Type | Network ADE Type | Score |
|---|---|---|
| Floor_lamp | ControllerDevice | 0.31 |
| Damage | RoundShell | 0.19 |
| Floor_lamp | AnyDevice | 0.19 |
| Standpipe | Abstract_Pipe | 0.19 |
| Water_valve | Canal | 0.19 |

Table 9. Wordnet matching between GUS-InfraGML

| GUS Type | InfraGML Type | Score |
|---|---|---|
| Pumping_station | SetType | 0.54 |
| Floor_lamp | AbstractObject_Type | 0.31 |
| Gas_pipe | SC_CRS_PropertyType | 0.31 |
| Pumping_station | PropertySetPropertyType | 0.31 |
| Sewer_junction | DirectedNodePropertyType | 0.31 |

Table 10. Wordnet matching between GUS-IFC

| GUS Type | IFC Type | Score |
|---|---|---|
| Damage | IfcCenterLineProfileDef | 0.54 |
| Water_valve | IfcAirToAirHeatRecovery | 0.52 |
| Floor_lamp | IfcWorkControl | 0.31 |
| Gas_pipe | IfcTransportElement | 0.31 |
| Hydro_network | IfcTelecomAddress | 0.31 |

## 5. DISCUSSION

The main objective of this first phase of a larger study was to explore schema matching techniques, and to determine whether the outputs of these techniques can in turn facilitate the comparison of, and eventually the selection of, geospatial standards, as a first step towards a solution to help organisations to select an appropriate standard. This preliminary exploration of schema matching highlighted the pros and cons of the approach for comparing and selecting a standards-based schema. Important outcomes are summarised here:

- Global analysis that allows the rapid comparison of many standards, although using weak semantic measure, is a helpful and efficient approach.
- String matching and structural matchers operated in isolation of other results do not offer good results in the context of the comparison of geospatial standards.
- Schema matching becomes a difficult task when schemas are large, when the number of possible matches is high. Automation is not well implemented and is inappropriate in some situations.
- Grouping items when performing the schema comparison is a valuable approach to easiness the interpretation of the results.
- An iterative process is recommendable, i.e. first find the 1-1 relationships, and try to reduce the cardinality matching as much as possible.
- A non-match score is a useful source of information.
- Having the definition of classes of entities and attributes is important for semantic comparison in particular.
- Having more descriptive information (e.g. annotation, documentation, thesaurus) to support the matching process makes the scores higher and produce more correspondences, but the interpretation of the results by the user will be more complex.
- Using external sources, such as WordNet, to determine the semantic similarity between element names makes the matching scores higher but the results are then dependent on the quality (relevance) of the external source.
- A high matching score does not implicitly mean a good match. We found a large number of contradictory results, so the user still needs to be involved in the process; full automation of all the process of comparison is not possible.
- It is not possible to identify one standard that best matches the use need or other standards solely through schema

matching, but schema matching can be used as an exploratory tool (for example to better understand standards and possibly rank them).

Based on our experiment, it is quite clear that the semantic heterogeneity for both class names and attributes of geospatial standards is manifest. Additionally, the strategy of using XSD, as delivered by the standardisation organisation also causes a few issues. For instance, the terminology, the level of detail and annotation varies a lot from one standard to another, and also within the XSDs. This aspect of heterogeneity results in low similarity scores and uncertainty in the comparison analysis. The number of schemas used is also problematic since most of the schema matching tools compare pairs of schema rather than many schemas in the same time.

In this experiment, we used OpenII software (having selected this following an exploration of other options). We found OpenII easy to run, useful and the diversity of viewing diagrams interesting. As indicated in the documentation of OpenII, the main purpose of the tool is not to automatically and fully match schemas but instead to give hints to the user to support the process of matching schemas. Currently, the semantic measures used are not clearly explained in the OpenII documentation and the consistency for some results was questionable. However, as the tool is open source it may be possible in future to explore the code in depth to further understand the algorithms and approaches used, as well as to improve and customise the tool, which is a great advantage.

In this first experiment and as explained, the focus was not on identifying one standard but more on exploring schema matching techniques. Since many results were contradictory between the various levels compared, and the number of elements to be compared varies a lot from one schema to another, it is not possible to state which one of the three standards best match the GUS. For example, if we assume that a score of 0.3 is a good match, IFC results in a larger number of matches with GUS (a total of 415) while InfraGML is 132 matches and Network ADE is 79 matches. However, if we place this value in the context of the total number of elements to match, IFC has a match rate of 0.1% while InfraGML is 0.3% and Network ADE is 1.7%. Obviously, the size of the schema impacts the matching results and these values do not reflect the accuracy of the matches.

Despite the above limitations, and even though the results are for now not able to guide the user directly to the selection of one specific geospatial standard, we believe that schema matching for the comparison of user needs and existing geospatial standards is a valuable approach. For example, the global comparison of many standards can be performed in few minutes and the results highlight both similarities and differences between schemas. We note, however, that schema matching apply to standard comparison should not be foreseen as an independent activity but instead as a phase in the design process and as a tool to facilitate the alignment of standards, updating of standards or the enhancing of standards. It should be seen as a way to narrow the number of standards to select from and to rapidly identify overlaps and gaps between them. We estimate the approach to be helpful for an initial triage of geospatial standards and even for promoting the reuse of concepts and semantics between domains of expertise (or conceptual models). Schema matching could also be perceived as rapid manner to discover the content of existing geospatial standards.

## 6. CONCLUSION

To date, this research project has concluded that while not offering an end-to-end solution that will provide a full schema selection process for non-expert end users, schema matching could form a key part of this process. While we did not identify a rigorous modus operandi to apply schema matching for the selection of geospatial standards, we have identified key parameters to consider.

However, it is clear that the quality of the matching results are to date difficult to understand. The outputs from the schema matching tools are not unique, not always clear, the users have to be involved in the process. Thus, better approaches and strategies will have to be recommended, both in terms of communicating the outputs for non-experts as well as potential further automation. In particular, we note that a significant amount of additional work is required before the existing tools and methods can be widely deployed to help organisations identify the best schema for their needs, with an appropriate level of customisation to take the level of user expertise into account. This is, however, worth the effort as it would help avoid the current ‘we will develop our own’ approach which is a major limitation to data interoperability.

### 6.1 Future Work

As mentioned, this is a preliminary experiment. As an initial component of further work, we are testing other schema matching tools, we are using different themes (e.g. buildings) and a larger number of geospatial schemas. We are also testing the generation of thesauri and annotation to improve the comparison process. An additional area to explore is the fact that the matcher can “learn” once the user explicitly accepts or rejects a link, and this option my produce interesting results. Furthermore, to date the quality (the accuracy) of the matching results was has not been appraised, as we were simply comparing numerical results without discussing which one is more accurate. Accuracy considerations could be seen as one possible way to extend the assessment of schema matching. We also plan to investigate the “spatial” characteristics of geographic features to enrich the effectiveness of the matching techniques.

## ACKNOWLEDGEMENTS

This project was funded by CRSNG RGPIN-2015-05514. We would also thank J. Coté and D. Boucher, two students who have produced the conceptual model of the users in the context of a previous project.

## REFERENCES

Algergawy, A., Nayak, R., Saake, G. 2010. Element similarity measures in XML schema matching. Information Sciences, 180(24), pp.4975-4998.

Bellahsène, Z., Bonifati, A., Rahm, E. 2011. Schema matching and mapping. Springer, Berlin.

Benerecetti, M., Bouquet, P., Zanobini, S. 2005. Soundness of Schema Matching Methods. The Semantic Web: Research and Applications, eds Gomez-Perez and Euzenat, Berlin, Heidelberg, pp.211-225.

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

Chen, N., He, J., Yang, C., Wang, C. 2012. A node semantic similarity schema-matching method for multi-version Web Coverage Service retrieval. International Journal of Geographical Information Science, 26(6), pp.1051-1072.

Cohen, W.W., Ravikumar, P., Fienberg, S. E. 2003. A Comparison of String Metrics for Matching Names and Records. SIGKDD Conference.

Coté, J., & Boucher, D. 2016. Élaboration d'une base de données spatiales et d'une interface cartographique pour la représentation de réseaux souterrains. Rapport technique, Projet de génie, Département des sciences géomatiques, Université Laval.

Do, H.-H., Melnik, S., Rahm, E. 2003. Comparison of Schema Matching Evaluations. A.B. Chaudhri et al. (Eds.): Web Databases and Web Services 2002, LNCS 2593, Springer-Verlag Berlin Heidelberg, pp.221–237.

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

Doan, A., 2002. Learning to Map between Structured Representations of Data. Ph.D. thesis Computer Science & Engineering, University of Washington.

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

Fan, H., Liu, J., Deng, K. 2016. Towards a composite XML schema matching approach using reference ontology. Paper presented at the 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), 8-10 July, Los Alamitos, CA, USA, pp.724-728.

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA, MIT Press.

Giunchiglia, F., & Yatskevich, M. 2004. Element level semantic matching. Proceedings of Meaning Coordination and Negotiation workshop at the International Semantic Web Conference (ISWC), 7-11 November, Hiroshima, Japan.

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.

Kutzner, T., & Kolbe, T.H. 2017. CityGML Utility Network ADE - Scope, Concepts, and Applications. Infrastructure Mapping and Modeling Workshop, NY City, April 24-25.

Li, Y., Bandar, Z. A., Mclean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering, 15(4), pp.871-882.

Lin, D. 1998. An Information-Theoretic Definition of Similarity. Proceedings of the Fifteenth International Conference on Machine Learning, pp.296-304.

Martinez-Gil, J., & Aldana-Montes, J. F. 2013. Semantic similarity measurement using historical google search patterns. Information Systems Frontiers, 15(3), pp.399-410.

Miller, G. A. 1995. WordNet: a lexical database for English. Communication ACM, 38(11), pp.39-41.

Milo, T., & 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.

Navarro, G. 2001. A Guided Tour to Approximate String Matching. ACM Computing Surveys, 33(1), pp.31-88.

Peukert, E., Maßmann, S., König, K. 2010. Comparing Similarity Combination Methods for Schema Matching. Conference GI Jahrestagung.

Rada, R., Mili, H., Bicknell, E., Blettner, M. 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1), pp.17-30.

Rahm, E. 2011. Towards Large-Scale Schema and Ontology Matching. In Z. Bellahsene, A. Bonifati, & E. Rahm (Eds.), Schema Matching and Mapping. Springer Berlin Heidelberg, pp.3-27.

Rahm, E., & Bernstein, P. A. 2001. A survey of approaches to automatic schema matching. The VLDB Journal, 10(4), pp.334-350.

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., & Euzenat, J. 2005. A survey of schema-based matching approaches. J. Data Semantics IV 3730, pp.146-171.

Smiljanic, M. 2006. XML Schema Matching Balancing Efficiency and Effectiveness by means of Clustering. Ph.D. thesis, Dutch Graduate School for Information and Knowledge Systems. Twente University, Enschede - The Netherlands.

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.

Sorrentino, S., Bergamaschi, S., Gawinecki, M. 2011, 11-16 April 2011. NORMS: An automatic tool to perform schema label normalization. 2011 IEEE 27th International Conference on Data Engineering, pp.1344-1347.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1), pp.11-21.

Tiwari, A., & Trivedi, V. 2012. Estimating Similarity of XML Schemas using Path Similarity Measure. International Journal of Computer Applications & Information Technology, 1(1), pp.34-37.

Wu, Z., & Palmer, M. 1994. Verb Semantics and Lexical Selection. ACL Conference, pp.133-138.

Yi, S., Huang, B., Tat Chan, W. 2005. XML application schema matching using similarity measure and relaxation labeling. Information Sciences, 169(1), pp.27-46.

Yu-hong, W., He-bing, Z., Xu Jun, X. 2015. A Survey of Applications and Researches on Schema Matching between GIS Spatial Data. International Workshop on Image and Data Fusion, 21-23 July, Kona, Hawaii, USA, pp.175-179.