Doktorandský den ’05

Ústav informatiky
Akademie věd České republiky

Hosty – Týn nad Vltavou

5. – 10. říjen 2005

vydavatelství Matematicko-fyzikální fakulty
University Karlovy v Praze

Publikaci ”Doktorandský den ’05” sestavil a připravil k tisku František Hakl
Ústav Informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8

ISBN *– not yet –*

Obsah

Data integration in VirGIS and in the Semantic Web

Post-Graduate Student:
ING. ZDEŇKA LINKOVÁ

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2

182 07 Praha 8

linkova@cs.cas.cz

Supervisor:
ING. JÚLIUS ŠTULLER, CSC.

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2

182 07 Praha 8

stuller@cs.cas.cz

Field of Study:
Matematické inženýrství
Classification: 39-10-9

This work was supported by the project 1ET100300419 of the Program Information Society (of the Thematic Program II of the National Research Program of the Czech Republic) “Intelligent Models, Algorithms, Methods and Tools for the Semantic Web Realization”, by the project of Czech-French Cooperation Barrande 2004-003-1, 2: “Integration de données sur le Web - applications aux Systémes d’Information Géographique (2003-2005), and by the Institutional Research Plan AV0Z10300504 “Computer Science for the Information Society: Models, Algorithms, Applications”.

Abstract

Integration has been an acknowledged data processing problem for a long time. However, there is no universal tool for general data integration. Because various data descriptions, data heterogeneity, and machine unreadability, it is not easy way. Improvement in this situation could bring the Semantic Web. Its idea is based on machine understandable web data, which bring us an opportunity of better automated processing. The Semantic Web is still a future vision, but there are already some features we can use. The paper describes how is integration solved in mediation integration system VirGIS and discusses use of nowadays Semantic Web features to improve it. According to the proposed changes, a new ontology that covers data used in VirGIS is presented.

1. Introduction

Today’s world is a world of information. Expansion of World Wide Web has brought better accessibility to information sources. However, in the same time, the big amount of different formats, data heterogeneity, and machine unreadability of this data have caused many problems. One of them is a problem of integration. To integrate data could mean to provide one global view over several data sources and let them be processed as one source. To integrate data means to provide one global view over different data sources [1]. This view can be either materialized, or virtual. An important thing is to combine data in meaningful way and let them be accessible as one whole. There are two main problems resulting from the data integration. The first is the data modeling (how to integrate different source schemas); the second is their querying (how to answer to the queries posed on the global schema). The integration process is not easy. Yet, there is no universal tool or method that could be used every time when needed. Nevertheless, there are some partial solutions in many research areas.

As mentioned above, data features make automated processing difficult. Exactly from this base rises the idea of the Semantic Web [2]. It considers data to go along with their meanings. An addition of semantics would make data machine readable and understandable. The automation could be easier. This proposal is for general web data, so it offers to use it also for specialized kind of data.

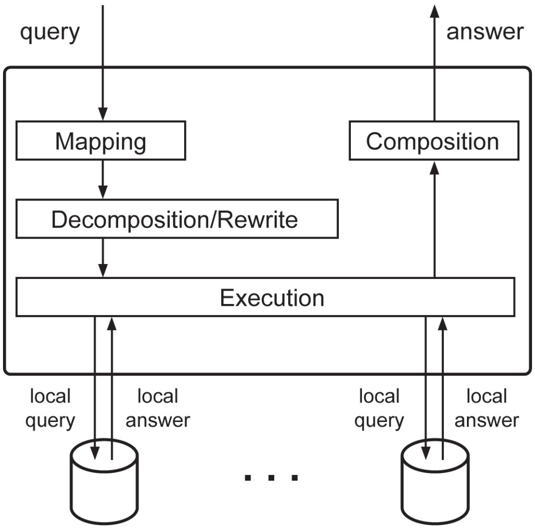

Figure 1: VirGIS System

Integration has been solved also in the area where GIS (Geographic Information Sources) [3] are used. Among these solutions, there is also VirGIS, an integration system that work with satellite images.

2. VirGIS System

VirGIS [4] is a mediation platform that provides an virtual integrated view of geographic data. In general, the main idea in a global virtual view use is a system of components called mediators. Mediators provide an interface of the local data sources. There are also other special components - wrappers, which play the roles of connectors between local source backgrounds and the global one. The principle of integration is to create a nonmaterialized view in each mediator. These views are then used in the query evaluation. Essential are mapping rules that express the correspondence between the global schema and the data source ones. The problem of answering queries is another point of the mediation integration - a user poses a query in terms of a mediated schema, and the data integration system needs to reformulate the query to refer to the data sources.

VirGIS accesses GIS data sources via Web Feature Service (WFS) server and uses WFS interfaces to perform communications with sources. WFSs play the role of wrappers in the mediation system. VirGIS uses GML as an internal format to represent and manipulate geographic information. GML is a geographic XML-based language; therefore GQuery, a geographic XQuery-based language, is used for querying. The integration system has only one mediator called GIS Mediator. It is composed of a Mapping module, a Decomposition/Rewrite module, an Execution module and Composition module.

The Mapping module uses integrated schema information in order to express user queries in terms of local source schemas. Each mapping rule expresses a correspondence between global schema features and local ones. For the global schema definition, a Local As View (LAV) approach is applied. This approach consists in defining the local sources as a set of views made on the global schema. In current version of VirGIS, there are used simple mapping rules that allow the specification of one-to-one schema transformations under some constraints: aggregations and one-to-many mappings are not considered. The Decomposition/Rewrite module exploits information about source feature types and source capabilities to generate an execution plan. A global GQuery expression is used as a container for collecting and integrating results coming from local data sources. The Execution module processes sub-queries contained in the execution plan and sends them to the appropriate source's WFS. The Composition module treats the final answer to delete duplicities and produces a GML document, which is returned to the user.

3. Use of Semantic Web features in mediation integration system

The Semantic Web is intended as an extension of today's World Wide Web. It should consist of machine readable, understandable and meaningfully processable data. The basis is addition of data semantics - there will be stored data meaning description together with data themselves. The Semantic Web idea belongs still to the future; however, there have been made already some features. It is based on standards, which are defined by W3C (WWW Consortium) [5]. The Semantic Web could improve or make easier to automate some operations. Hopefully it could bring something more also in data integration process. There are some areas, which could benefit by better automatization; for example addition of new sources, mapping rules generation and schema evolving.

3.1. Data sources

An important requirement of machine processible information is data structuring. On the web nowadays, the language XML (eXtensible Markup Language) [6] is used for making web document structure. But only XML is not enough to describe data. The technique to specify the meaning of information is RDF (Resource Description Framework) [7]. It is basic tool of web sources metadata addition. RDF data model gives an abstract conceptual framework for metadata definition and usage. It uses XML syntax (RDF/XML) for encoding. Additionally, there is also an extension of RDF called RDF Schema [8] that is useful for class definition and class hierarchy description. Instruments for definition of terms used either in data or in metadata are ontologies. In the context of web technologies, ontology is a file or a document that contain formal definitions of terms and term relations. The Semantic Web technique for definition of ontologies is the OWL (Ontology Web Language) [9] language.

In the VirGIS integration system, an XML-based language is used for data representation. If the integration is XML-based, why not bring more and, instead of simple XML, use RDF, which has bigger expressive power. So in the proposed integration system, the RDF is intended to represent information. Also XML document primarily not intended for RDF applications could be described using RDF. By observing several guidelines when designing the schema, he proposed how to make an XML ”RDF-friendly” [10]. For already existing documents, there is possibility to make some XML-RDF bridge. Of course, it has not to be always simple way.

As with data, the XML and RDF worlds use different formalism for expressing schema. The Semantic Web currently uses languages such as RDFS and OWL. So in the proposed integration system, OWL is used to publish sets of terms (called ontologies). Of course a source can use some richer ontology (richer than the source need as the schema). In this case, the source schema can be seen as a view of the ontology.

3.2. Querying

According to data description change, a change in querying is needed. Since RDF is defined using an XML syntax, it might appear on the first sight, that a query language and system for XML would also be applicable to RDF. This is, however, not the case, since XML encodes the structure of data and documents whereas the RDF data model is more abstract. The relations or predicates of the RDF data model can be user defined and are not restricted to child/parent or attribute relations. A query language based on XML element hierarchies and attribute names will not easily cope with the aggregation of data from multiple RDF/XML files. Also, the fact that RDF introduces several alternative ways to encode the same data model in XML means that syntax-oriented query languages will be unable to query RDF data effectively. Having motivated the need of an RDF query language, there was developed some query languages. A standardized query language for RDF data is called SPARQL [11].

3.3. Mapping and query rewriting

Essential task for the integration system are mapping rules and query rewriting, too. Closely related with it is also new sources addition and how (or whether) it could be done automatically. Mapping rules in VirGIS are expressed utilizing XML. However, the idea about the improvement of the integration system is to be able apply existing mapping rules, knowledge about already integrated sources, and knowledge about the new one to generate (automatically as much as possible) appropriate new mapping rules. Doing this, taking advantage of an inference mechanism tool would be practicable. But it requires machine processable data. Similarly to data sources, there is an idea to use RDF/XML instead of this pure XML. Nevertheless, even RDFS has no construct for terms or classes equivalency expression. There must be used some additional capabilities.

A possibility is own development to enrich RDF(S). Another possibility is to work with OWL, which is standard extension of RDFS. Using OWL provides at least two approaches. The first way is definition of mapping rules as a special class. The second way is to present mapping between schemas and concepts of sources by usage of OWL construct in order to express equivalency of some parts of different sources ontologies. The same situation is also in field of query rewriting. It needs further study. Of course, there some existing algorithms that could be used. Or, this could be improved, according to chosen technique of mapping rules definition, cleverness of particular local sources query mechanism, and potentialities of an accessible tool that implements SPARQL.

4. Building ontology for VirGIS system

The first step towards a Semantic Web-based version of integration system VirGIS was VirGIS ontology development. This task was joint work with Radim Nedbal[1]. Our aim was to build an ontology for a given data domain; it had cover at least data provided by VirGIS.

The term “ontology” has been used in many ways and across different communities. A popular definition of the term ontology in computer science is: an ontology is a formal, explicit specification of a conceptualization. A conceptualization refers to an abstract model of some phenomenon in the world. However, a conceptualization is never universally valid. Ontologies have been set out to overcome the problem of implicit and hidden knowledge by making the conceptualization explicit. An ontology may take a variety of forms, but it will necessarily include a vocabulary of terms and some specification of their meaning.

There are many tools and languages [12] that can be employed as means for ontology development. Among available ontology languages, Web Ontology Language (OWL) was chosen. OWL is proposed to be an ontology language for the Semantic Web. OWL, a XML based language, has more facilities for expressing meaning and semantics than XML, RDF, and RDF Schema, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web. OWL adds more vocabulary for describing properties and classes. A large number of organizations have been exploring the use of OWL, with many tools currently available.

As an ontology design tool, Protégé System [13] was used. Protégé is an integrated software tool used to develop ontologies and knowledge-based systems. Protégé has been developed by the Stanford Medical Informatics (SMI) at Stanford University.

4.1. VirGIS data

VirGIS is implemented as an integration system of satellite images. Figure 2 illustrates local and global sources of VirGIS. As local sources are used subsets of schemas drawn from SPOT and IKONOS catalogues and QUICK_LOOK database.

SPOT and IKONOS catalogues provide information about satellites; QUICK_LOOK refers to a sample of small images that give an overview of satellite images supplied in the catalogue. The role of the global source is played by the VIRGIS mediated schema. The VIRGIS schema contains just one entity VIRGIS with following attributes:

- string *id* (a common id for the different region photographed)
- string *name* (the name of the satellite that takes the photo)

[1] nedbal@cs.cas.cz

SPOT	
Attribute	Type
date_	Date
sun_elev	numeric
satellite	string
sat_id	numeric
key	string
the_geom.	Polygon

IKONOS	
Attribute	Type
date_acqui	Date
sun_el	numeric
satellite	string
sat_id	numeric
key	string
the_geom	Polygon

VIRGIS	
id	string
name	string
satid	string
date	Date
sun_elevation	numeric
url	string
geom	Polygon

QUICK LOOK	
Attribute	Type
key	string
filename	string

Figure 2: Local and global satellite schemas

- string *satid* (the id for the satellite)
- date *date* (the date when the photo was taken)
- numeric *sun_elevation* (the sun elevation when photo was taken)
- string *url* (the url where the real photo is saved)
- polygon *geom* (the geometry of the region photographed)

According to this schema description, the aim was a development of an ontology satisfying the VirGIS data semantics. It had to cover not only the global schema, but also the local ones and relationships among them.

4.2. The VirGIS ontology

The aim was a description of satellite image knowledge in a VirGIS ontology. In ontology re-use, we can consider only some general spatial ontology for basic geometric features. The VirGIS data area itself is not covered with any existing GIS ontology. A new ontology for this purpose is needed.

The proposed VirGIS specified ontology comes out of the data model described above. The main domain concepts and their relationships are depicted in Figure 3 by means of ISA tree.

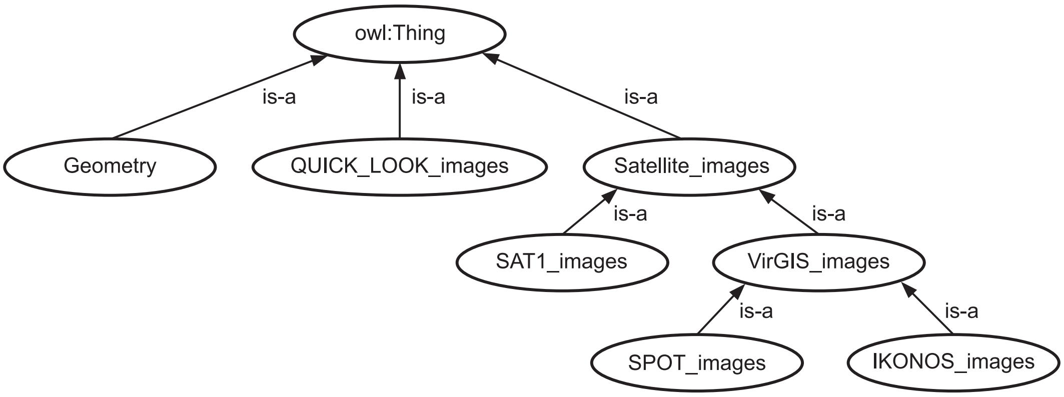

Figure 3: ISA diagram of the model

Observe that each node corresponds to one concept. `IKONOS_images` and `SPOT_images` refer to local sources; `VirGIS_images` refers to the global mediated source. The fact that every image contained in IKONOS or SPOT database is also contained in VirGIS induces the corresponding concepts relationship that can be understood as set inclusions:

$$\begin{aligned} IKONOS\_images \subseteq VirGIS\_images, \\ SPOT\_images \subseteq VirGIS\_images, \end{aligned} \tag{1}$$

Analogical relationship applies to `VirGIS_images` and `Satelite_images` concepts. Observe that there is an additional class `SAT1_images` in the model. It contains satellite images not integrated in VirGIS_images. Finally, an inherent feature of the OWL data model is the unique superclass `THING` being the superclass of all other classes.

In OWL, a `owl:Class` construct is used for concept indication and `rdfs:subClassOf` construct for expressing the concept relationships corresponding to set inclusion relations:

Example 1 *The OWL expression of the relationship of* `SPOT` *and* `VirGIS` *classes*

```
<owl:Class rdf:ID="SPOT_images">
   <rdfs:subClassOf rdf:resource="#VirGIS_images" />
</owl:Class>
```

The `rdfs:subClassOf` construct expresses inclusion relationship on both set and conceptual level. Therefore, the above OWL code example implies SPOT_images being conceptually more specific than VirGIS_images.

In OWL, classes are also characterized by means of properties, i.e. attributes of corresponding concepts. Properties definitions are to represent the semantic relationships of the corresponding concepts and their attributes.

Observe that SPOT and IKONOS use semantically equivalent attributes without any common name convention. In addition, VirGIS introduces its own identifiers for respective attributes. `date_` (SPOT), `date_acqui` (IKONOS) and `date` (VirGIS) represent semantically equivalent attributes for instance. This is solved with mapping of mediation integration in VirGIS. However, it can naturally be expressed on the semantic level, by means of OWL.

With regard to the above discussion and considering the inclusion (1), it follows:

$$(\forall image \in SPOT\_images)(date\_(image, DD/MM/YY) \rightarrow date(image, DD/MM/YY)),$$

which defines the semantic relationship of the binary predicates `date_` and `date`. The relationships between other predicates can be expressed analogically.

In OWL, `rdfs:subPropertyOf` construct is used for expressing such semantic relationships:

Example 2 *The OWL interpretation of the relationship of the properties* `date_` *and* `date`

```
<owl:DatatypeProperty rdf:about="#date_">
    <rdfs:subPropertyOf rdf:resource="#date" />
</owl:DatatypeProperty>
```

This relationship is more vague than the relationship of equivalence. However, the relationship of “subPropertyOf” mirrors SPOT_images being conceptually more specific than VirGIS_images.

For completeness, there is an additional class in the model. `Geometry` class contains geometric elements, designed for geometry type properties description. In case that richer geometry is needed, geometry classes from existing spatial ontologies can be imported. At this time, the presented ontology is suitable for VirGIS data description. It can be enriched in case more capabilities should be needed.

5. Conclusion

Data integration is a real problem of information processing for a long time. There were already done some solving steps, whether partial solutions in particular research areas, or development towards the Semantic Web. A lot of work must be still done. The first step for in this paper proposed system was done. A new ontology describing sources and data in the VirGIS integration system was developed. Further tasks are planned: mapping expression, query rewriting, and infer mechanism and tools.

References

[1] Z. Bellahsene, “Data integration over the Web”, *Data&Knowledge Engineering*, vol. 44, pp. 265–266, 2003.

[2] M.-R. Koivunen and E. Miller, “W3C Semantic Web Activity”, in the proceedings of the *Semantic Web Kick/off Seminar*, Finland, 2001.

[3] B. Korte George, “The GIS (Geographic Information Systems)”, *OnWord Press*, Santa Fe, 1994.

[4] O. Boucelma and F.-M. Colonna, “Mediation for Online Geoservices”, in Proc. *4th International Workshop Web & Wireless Geographical Information System, W2GIS 2004*, Korea, 2004.

[5] W3C (WWW Consortium), `http://www.w3.org`.

[6] Extensible Markup Language (XML), `http://www.w3.org/XML/`.

[7] Resource Description Framework (RDF), `http://www.w3.org/RDF/`.

[8] RDF Vocabulary Description Language 1.0: RDF Schema, *W3C Recommendation*, `http://www.w3.org/TR/2004/REC-rdf-schema-20040210`, February, 2004.

[9] Web Ontology Language (OWL), `http://www.w3.org/2004/OWL`.

[10] B. DuCharme and J. Cowan, “Make Your XML RDF-Friendly”, October, 2002, `http://www.xml.com/pub/a/2002/10/30/rdf-friendly.html`.

[11] SPARQL Query Language for RDF, *W3C Working Draft*, October, 2004.

[12] O. Corcho, M. Fernández-López, and A. Gómez-Pérez, “Methodologies, tools and languages for building ontologies. Where is their meeting point?”, *Data & Knowledge Engineering*, vol. 46, pp. 41–64, 2003.

[13] The Protégé Ontology Editor and Knowledge Acquisition, `http://protege.stanford.edu/index.html`.

Ústav Informatiky AV ČR
DOKTORANDSKÝ DEN '05

Vydal
MATFYZPRESS
vydavatelství
Matematicko-fyzikální fakulty
University Karlovy
Sokolovská 83, 186 75 Praha 8
jako svou *– not yet –* . publikaci

Obálku navrhl František Hakl

Z předloh připravených v systému LaTeX
vytisklo Reprostředisko MFF UK
Sokolovská 83, 186 75 Praha 8

Vydání první

Praha 2005

ISBN *– not yet –*