Data Integration in VirGIS and in the Semantic Web

Linková, Zdeňka
2005

Institute of Computer Science
Academy of Sciences of the Czech Republic

Data Integration in VirGIS and in the Semantic Web

Zdeňka Linková

Technical report No. 922

January 2005

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 053 030, fax: +420 286 585 789,
e-mail:sima@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Data Integration in VirGIS and in the Semantic Web

Zdeňka Linková

Technical report No. 922

January 2005

Abstract:

Integration has been an acknowledged data processing problem for a long time. Integration is needed in many areas, but because various data descriptions, data heterogenity, and machine unreadability, it is not easy way. Some data from particular areas has been already integrated; for example, there is the VirGIS mediation integration system for particular set of Geographic Information Systems. However, there is no universal tool for general data integration. 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 Sementic Web is still a future vision, but there are already some features we can use. This report briefly describes how is integration solved in VirGIS and discusses usage of nowadays Semantic Web features to improve it.

Keywords:
Data Integration, Semantic Web, GIS

Contents

1 Introduction

Today's world is a world of information. Everything depends on information, whether science progress or business success. Expansion of World Wide Web has brought better accessibility to information sources. However, in the same time, the big amount of different formats, data heterogenity, and machine unreadability of this data have caused many problems. One of them is a problem of integration.

To integrate data means to provide one global view over several data sources and let them be processed as one source. The way is not easy. Yet, there is no universal tool or method that could be used every time when needed. Though, there are some partial solutions in many research areas. The same situation is also in the area where GIS (Geographic Information Sources) [1] are used. But in general, a resolution of the integration problem does not exist.

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, e.g. GIS.

I have studied VirGIS – a GIS integration system and main Semantic Web features as well. I started from pure VirGIS, how it was originally designed. Then I considered how some methods and techniques of Semantic Web could be used. And of course, how we can obtain from it.

2 Data Integration

To integrate data means to provide one global view over different data sources. This view can be either materialized, or virtual [3]. 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 solution of these partial problems depends on many things and conditions, among them is if the global view is materialized or virtual and how much information should be supported in the global source.

The possibility of materialized integration view is sometimes called data warehouse. In this case data are at first obtained and stored in a warehouse. After it, data processing does not need accessibility of primary data sources. However, task of memory space usage, task of information actuality etc. must be solved.

The main idea in usage a global virtual view 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. The issues of the schema integration are the sources heterogenity, global schema modeling and definition, the source semantics and the management of the coherence and the evolution of the schema. Essential in this case are mapping rules that express the correspondence between the global schema and the data source ones. So another issue is the definition and the management of these rules.

The global schema definition, that provides a uniform view of the different sources, can be done using two different approaches. The Global As View (GAV) approach consists in defining the global schema as a set of views over local schemas, while the Local As View (LAV) one consists in defining the local sources as a set of views made on the global schema. There are also approaches combining both.

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. The queries are executed over the sources. The reformulation problem can be solved by algorithms for answering queries using views [4]. Though in this context, a rewriting that is equivalent to the user query cannot be found because of the data sources' limited coverage. Instead, it is searched for a maximally-contained rewriting, which provides the best answer possible, given the available sources.

3 VirGIS

VirGIS [5] is a mediation platform that provides an integrated view of geographic data. VirGIS accesses GIS data sources via Web Feature Service (WFS) [6] server and uses WFS interfaces to perform communications with sources. WFSs play the role of wrappers in the mediation system. VirGIS uses GML [7] as an internal format to represent and manipulate geographic information. GML is a geographic XML-based language; therefore GQuery [8], 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.

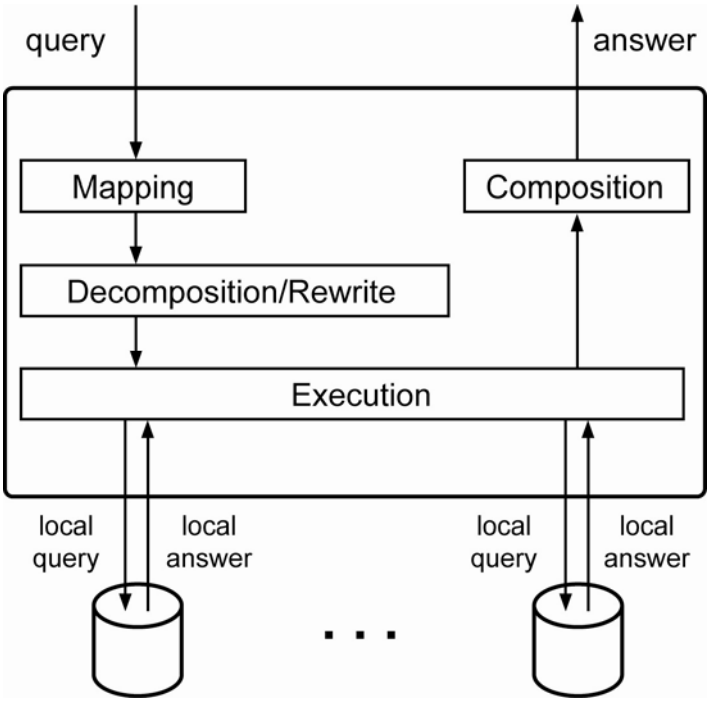

Figure 1: VirGIS

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. In current version of VirGIS, a LAV like approach is used, with 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 [9]. A global GQuery expression is used as a container for collecting and integrating results coming from local data sources. The query rewriting algorithm is inspired from the one used in the Styx system [10]. 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.

4 Semantic Web

The Semantic Web [2] is intended as an extension of today's World Wide Web. It should consist of machine readable, understandable and meaningfully processible 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 [11]). Semantic Web principles are implemented in layers of web technologies and standards. The layers are figured in Figure 2.

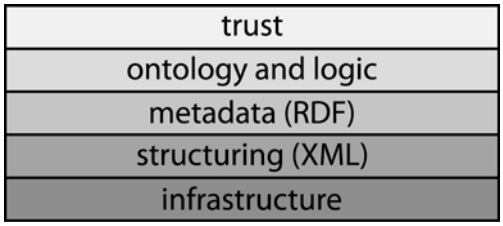

Figure 2: Semantic Web layers

The layer of infrastructure provides a source identification and location. The layer of structuring, the layer of metadata and the logic layer are essential for describing web sources content. The layer of trust is a thing of particular application. It considers proofs and trust about web information.

Infrastructure

The Semantic Web should consist of connected sources – it should contain sources and links. Every object should be identified (as on today's web) with identifiers URI (Universal Resource Identifier). The Semantic Web should be decentralized, of course with possibility of missing or incomplete information. What it should bring more, are not only classical web sources (web pages and documents), but also objects like people, places, and events. Moreover, it should be able to define source types and links types of course.

Data description

An important requirement of machine processible information is data structuring. On the web, the main structuring technique is using tags, which are parts of text containing information about the role of the text. Nowadays, the language XML (eXtensible Markup Language) [12] is used for making web document structure. It provides syntax for machine readable data.
But only XML is not enough to describe data. The technique to specify the meaning of information is RDF (Resource Description Framework) [13]. 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) [14] for encoding. Additionally, there is also an extension of RDF called RDF Schema [15] that is useful for class definition and class hierarchy description.
An instrument 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) [16] language. Thanks to usage of ontologies, applications

can share terms and so it enables application cooperation. Moreover, the Semantic Web idea considers also addition of logic and using inference rules. It brings a possibility to infer and to make conclusions.

Application operation

The real potential of the Semantic Web would express if people made many programs that would process web sources content and cooperate with other programs. These software agents would be as effective as the web data would be machine understandable and as automated services would be accessible. The Semantic Web should provide a basis for the other technologies.

5 Usage of Semantic Web features in mediation integration system

The Semantic Web promises a basis for machine understandable data. In consequence, it 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. And because the Semantic Web is about standards, we could reuse some tools, which are already made. If started from current state of VirGIS, several things must be changed. Inspired with this integration system, also the proposed system will be mediator-based.

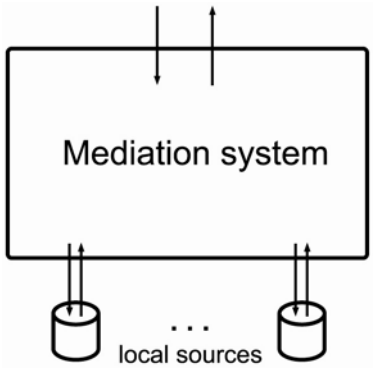

Figure 3: Mediation integration system

Data sources

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. According to DuCharme and Cowan [17], 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". Also for already existing documents, there is possibility to make some XML-RDF bridge. Of course, it has not to be always simple way. There is wide disconnect between the RDF world and most of today's data. RDF is focused on identifying the domain structure. In contrast, most existing data sources and applications export their data into XML, which tends to focus less on domain structure and more around important objects or entities. However, they often nest information within the descriptions of more important objects and in this way (using document structure) they express relationship between objects. In doing this, they sometimes leave the relationship type unspecified. Though, the name of relation is missing, some relation between object is expressed. We must add missing information, in order to keep maximum information contained in original XML source [18].
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.

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 [19].

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 processible 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 maybe, we could also improve this, according to chosen technique of mapping rules definition, cleverness of particular local sources query mechanism, and potentialities of an accessible tool that implements SPARQL.

6 Conclusion remarks

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. For in this paper proposed system, some tasks are planned: study of ontologies, query rewriting, and infer mechanism and tools. Because VirGIS is real, practical, and concrete integration system, first future step would be about this; particularly about VirGIS data and their ontologies.

Acknowledgements

I studied VirGIS system in laboratories “Laboratoire des Sciences de l’Information et des Systèmes” in Marseille, France. My visit was supported by the project of Czech-French cooperation Barrande 2004-003-1: “Integration de données sur le Web - applications aux Systémes d'Information Géographique (2004-2005)” and 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.”

Literatura

[1] Your Internet Guide to GIS (Geographic Information Systems). http://www.gis.com.

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

[3] Z. Bellahsene: "Data integration over the Web." *Data & Knowledge Engineering* 44 (2003), pp. 265-266.

[4] R. Pottinger and A. Levy: "A Scalable Algorithm for Answering Queries Using Views," in the *Proceedings of the 26th VLDB Conference*, Cairo, Egypt (2000).

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

[6] Open GIS Consortium Inc.: "Web Feature Service Implementation Specification." *OpenGIS Implementation Specification*, May, 2002. https://portal.opengeospatial.org/files/?artifact_id=7176.

[7] Open GIS Consortium Inc.: "OpenGIS Geography Markup Language (GML) Implementation Specification," January, 2003. https://portal.opengeospatial.org/files/?artifact_id=7174.

[8] O. Boucelma and F.-M. Colonna: "GQuery: a Query Language for GML," in the Proc. *24th Urban Data Management Symposium*, Chioggia-Venice, Italy, October, 2004.

[9] M. Essid, O. Boucelma, Y. Lassoued and F.-M. Colonna: "Query Processing in a Geographic Mediation System," in *Proceedings of The 12th International Symposium of ACM GIS*, Washingtin D.C., November, 2004.

[10] B. Amann, C. Beeri, I. Fundulaki, and M. Scholl: "Querying XML sources using an Ontology-based Mediator," in *On the Move to Meaningful Internet Systems, 2002 – DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, Springer-Verlag (2002), pp. 429-448.

[11] W3C (WWW Consortium). http://www.w3.org.

[12] Bradley, N.: XML kompletní průvodce. *Grada Publishing*, Praha, 2000. ISBN 80-7169-949-7

[13] Resource Description Framework (RDF). http://www.w3.org/RDF/.

[14] RDF/XML Syntax Specification (Revised). *W3C Recommendation*, February, 2004. http://www.w3.org/TR/rdf-syntax-grammar/.

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

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

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

[18] Z. G. Ives, A. Y. Halevy, P. Mork, and I. Tatarinov: “Piazza: mediation and integration infrastructure for Semantic Web data,” *Web Semantics: Science, Services and Agents on the World Wide Web* 1 (2004), pp. 155-175.

[19] SPARQL Query Language for RDF. *W3C Working Draft*, October. 2004. http://www.w3.org/TR/2004/WD-rdf-sparql-query-20041012/.