User-friendly Access to Structured Environmental Data

Andreas Abecker [1,2], Veli Bicer [2], Wassilios Kazakos[2],
Gabor Nagypal [1], Radoslav Nedkov [2,1] and Aleksei Valikov [1]

[1] disy Informationssysteme GmbH,
Erbprinzenstr. 4-12, D-76133 Karlsruhe, Germany
{Andreas.Abecker, Gabor.Nagycal, Aleksei.Valikov}@disy.net
[2] FZI Forschungszentrum Informatik,
Haid-und-Neu-Str 10-14, D-76131 Karlsruhe, Germany
{Veli.Bicer,Wassilios.Kazakos}@fzi.de

Abstract. We sketch the HIPPOLYTOS and the KOIOS system prototypes for simple, keyword-based search in structured environmental data. Built on top of a commercial Spatial Data Warehouse software, the prototypes apply lightweight Semantic Web techniques to facilitate search in complex environmental information systems.

Keywords: Semantic Search, Semantics and Environment

1 Motivation

With the increasing uptake of INSPIRE, the amount of publicly available environmental data is continuously growing. Nevertheless, the commercial and societal impact of open environmental data is still very limited. This has a number of reasons, but one of them is certainly the largely intransparent access to complex and distributed databases. This obviously holds true for interested citizen and companies, but even for employees of public authorities in a different domain, the heterogeneity and distribution of environmental data is often overwhelming. Hence, user-friendly and powerful search interfaces are a must-have in this area.

On the other hand, Semantic Web technologies promise advanced functionalities for intelligent search, integration, and processing of Web-accessible data, documents, and services [1]. Based on the annotation of complex machine-readable ***metadata*** to Web-based information resources, these resources can better be found and interpreted. Semantic metadata instantiate and refer to ***ontologies***, rich conceptual domain models, agreed within a certain user community and represented in expressive, logic-based languages, which are standardized by the World Wide Web Consortium (W3C) [2]. The power of semantic technologies relies on several factors – the specific importance of each depending on the specific use case – like standardization of representation languages, metadata or knowledge models; like automation of some reasoning services because of the logic-based semantics; or like empowering user or systems by the domain-specific background knowledge represented in ontologies. Ontologies for knowledge organization and human search and navigation, often comprise a so-called ***lexical layer***, which describes how the concepts are referred to by natural-language

expressions; such a lexical layer allows to cover the variability of human language, addressing phenomena like synonymy, or enabling multilingual knowledge access.
Although comprehensive metadata and extensive background knowledge for knowledge organization (in the form of thesauri) are widespread in environmental information systems, there are not many industrial-strength applications of semantic technologies in this area. This paper sketches the GUI approach of the HIPPOLYTOS project which aims at a practicable combination of semantic technologies and a commercial tool for geodata management and spatial reporting. Very generally spoken, the project goals of HIPPOLYTOS were:

- to map an intuitive text-based search interface at the front-end
- to complex data structures and relationships in the back-end (environmental information system/ spatial data warehouse)
- by exploiting existing expert knowledge (in form of domain ontologies and in predefined selectors and selector metadata),
- taking into account real-world constraints and requirements.

The presentation of the HIPPOLYTOS search interface is completed by a short description of the KOIOS schema-agnostic search, another prototype which has complementary characteristics.

The paper is structured as follows: in Section 2, we sketch the look-and-feel of the HIPPOLYTOS prototype; in Section 3, the same is done for KOIOS; in Section 4, we summarize and discuss current status and some future work.

2 Look-and-Feel of the HIPPOLYTOS Semantic-Metadata Search

In contrast to other semantic search projects, HIPPOLYTOS and KOIOS do not focus on text, documents or multimedia information (as we do in complementary research [3,4]), but on ***structured data*** in relational databases or a Data Warehouse. We develop a search layer on top of such data repositories realized, e.g., by disy GmbH's Cadenza software.[1]

Fig. 1 below illustrates the current prototype of the HIPPOLYTOS system: Assume the user types in "Eisenschrott Ballungsraum Stuttgart" ("iron junk city region Stuttgart") at a Google-like query interface. The system reasons as follows:

- "Iron junk" is not a technical term in environmental information systems, but "recyclable fraction FE scrap" is – which is represented in the ontology, with "iron junk" as a synonymous wording.
- The ontology also contains the taxonomic knowledge that "potential recyclables" is a super-concept of FE scrap and that "metal" is a super-concept of iron/FE whereas "waste" is a super-concept of scrap.

[1] disy Cadenza (http://www.disy.net/produkte/cadenza.html) is a system for building search, analysis, and visualization solutions for spatial data. At its core stands a re¬po¬si¬to¬ry system, which manages the back-end data sources. An important Cadenza concept are so-called Selectors, pre-defined query templates for the back-end systems which are designed by domain experts for specific query and analysis tasks. Selectors can be described with text metadata. They stand at the heart of many special applications that disy has built for environmental agencies and other public authorities.

- It also contains in its taxonomy the knowledge that "recyclable fraction Aluminium scrap" and "recyclable fraction glass" may be siblings to "recyclable fraction FE scrap" in the taxonomy.
- Furthermore, the lexical part of the ontology knows that "city region" is a synonym for "metropolitan region" or for "urban agglomeration", which is an informal term that can be mapped to several spatial interpretations, such as the city of Stuttgart, the Stuttgart region constituted by 6 neighboring administrative districts, or the geographic area within a certain radius around Stuttgart city center.

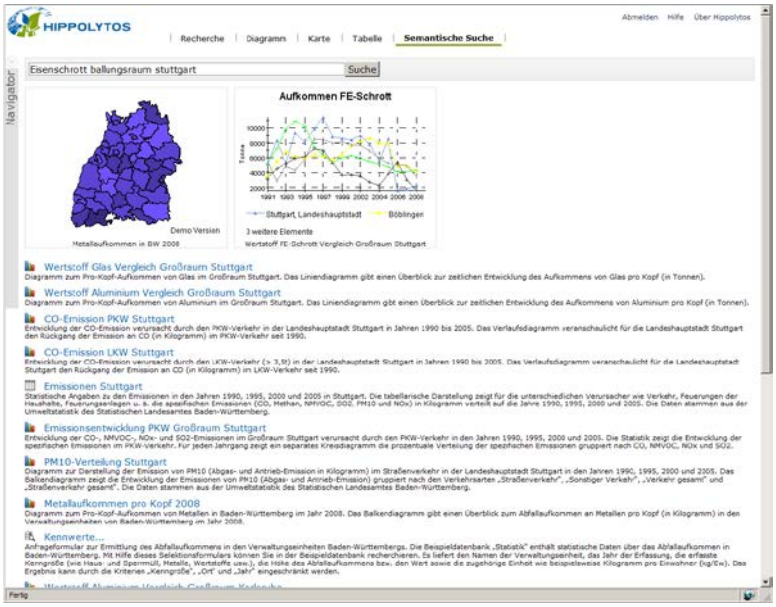

Fig. 1. Query for "Eisenschrott Ballungsraum Stuttgart".

Using the lexical and conceptual background knowledge, the system can identify a number of stored, semantically indexed ***selectors*** – parameterized, predefined query templates, accessing the data sources in the back-end. The match between query concepts and annotation concepts of stored selectors can be based on:

- The ***subject matter*** of the selector (e.g., there may be a selector querying for the amount of certain recyclable materials [which is a parameter of this selector] in sorted waste of a given region [2nd parameter] in a given timeframe [3rd parameter] – here a proximity match could be made with the "potential recyclables" concept in the set of super-concepts of the query concepts.
- The co-domain of the selector ***parameters*** (e.g., "FE" could be a parameter value for the 1st parameter of the example selector above, and "Stuttgart" for the 2nd parameter).
- The ***visualization*** or presentation type (data value, data series, data table, map-based visualization, specific diagram type, …) for the results. For

instance, if the query would contain terms like “comparison”, “trend”, or “distribution”, this could give hints to the expected kind of presentation.[2]

Then – for the given query – the most appropriate selectors and parameter settings can be identified and sent to the back-end system. The result screen in Fig. 1 shows a ranked list of potential result selectors as well as previews of the visualized results of the two top-ranked ones.

3 Look-and-feel of the KOIOS Schema-Agnostic Search

The KOIOS approach is fundamentally different (see Fig. 2). It applies a so-called ***schema-agnostic search***, which takes a set of keywords and heuristically creates a number of potential SQL-queries which *might* have been meant by the user when launching his keyword query. These hypotheses are based on the given DB-schema and the statistical distribution of DB-values occuring in the concrete, actual DB-content. Based on that hypothetical SQL-queries, we can then select those Cadenza Selectors which come close to them.

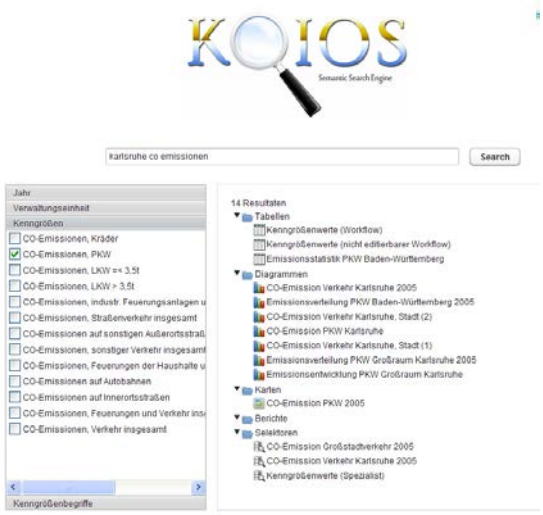

Fig. 2. KOIOS Faceted-Search Interface for Query “Karlsruhe CO Emissionen”.

In practice, with large DB-schema and often occurring data values, normally many different hypotheses will be possible. In order to quickly deliver probably highly relevant hypotheses to the user, ranking mechanisms are of utmost importance for such approaches (cp. [5]). Furthermore, we can offer to the user a very structured interface to navigate through the hypotheses, which is done by a ***faceted-search GUI*** (see Fig. 2): If different potential query interpretations differ in several dimensions, each dimension will be represented by one tab in the left-hand part of the search GUI in Fig.

[2] This kind of „query hints“ for the visualization type is not yet implemented.

2. There, a search for “CO emissions in the city of Karlsruhe” yields a number of possible selectors which can be differentiated according to:

- The year for which information is sought (only years for which we have actual DB content will be offered).
- The administrative region which is examined by the selection (concretely, Karlsruhe city versus Karlsruhe county).
- The actual measured value considered (e.g., CO emissions from motor-bikes, CO emissions from passenger cars, CO emissions from trucks, …).

A selection on this left-hand side will immediately refresh the list of possible resulting queries shown at the right-hand side of Fig. 2. As shown before in Section 2, clicking one of these results on the right-hand side will then evaluate this selector and deliver the actual query results from the back-end database.

4 Summary and Future Work

Summary. We have sketched functionality and realization of two prototypes for semantic search over geo-referenced environmental data. The goal is a “third kind of information access” for disy’s Spatial Reporting products, which currently offer map- and form-based access to geodata. This third kind of access shall be a Google-like, simple text query interface, which automatically finds and instantiates available selectors and thus automatically configures appropriate structured queries to the back-end data sources.

Following the HIPPOLYTOS approach, repository elements and their textual descriptions are semantically annotated (by hand, by the public servants who create the Selectors and attach already metadata for human reading, in the current installations) with ontology concepts. The ontological background knowledge about taxonomic and non-taxonomic relationships between concepts and about lexical variations of concept references allows to build a semantic index – such that also vague, too abstract, too specific, or wrongly expressed queries can be resolved. In order to provide a solution suitable for real-world usage, we aim at largely automated ontology creation and annotation processes.

KOIOS is a second prototype, which does not require any pre-modeled extra knowledge in form of ontologies or metadata, but instead exploits existing schema information and value distributions of database content in order to find possible query interpretations. The results are offered to the user in a faceted-search GUI which facilitates orientation in the space of possible query interpretations.

Obviously, both approaches, HIPPOLYTOS and KOIOS, exhibit different strengths and have different prerequisites and basic assumptions. This makes it an interesting idea to think about their synergetic integration.

Status. Both presented prototypes still have some “hardwired” aspects. But they show that, also for realistic data volumes and ontology sizes, it is possible to deliver reasonable results with acceptable performance. The evaluation of the retrieval *quality* still has to be evaluated in long-term experiments. Obviously, the quality of the HIPPOLYTOS retrieval depends on the used ontologies and annotations. Here, the practicability of *fully*-automated ontology creation and semantic annotation still has to

be verified – and, probably, user-friendly editors for manual corrections must be implemented. Regarding KOIOS, the ranking heuristics are the most important critical aspect to evaluate in practice. From the HCI points of view, both prototypes must be seen as design studies which explore technical feasibility, but still lack evaluation from the usability point of view. However, one thing seems to be clear: that mask-based, browsing-based, or also clumsy map-based search interfaces will hardly be accepted in the near future, by the members of the "Google generation" of end users. Regarding KOIOS, there are also positive prior results about the usability of faceted-search in Semantic Web search approaches.

Future work. There are still many areas for potential future work, to mention only two: (1) In the SUI and SUI II projects [3, 4], more usage and design studies for ontology-based access to environmental information have been performed, including *unstructured* information and the *links* between information sources, as well as *navigational* support for end users through ontological knowledge. A combination with HIPPOLYTOS/KOIOS could make sense. (2) The current approach mainly employs background knowledge about the domain of geo-referenced *environmental information*. It does not yet go very deeply into the semantic analysis of the *spatial concepts* themselves in the query. Though the use of ontologies is a longstanding research topic in GIS (see, e.g., [6,7]), it has not yet found its way very far into OGC or W3C standardization. Pragmatic steps into this direction may be a thrilling long-term goal.

Acknowledgment. HIPPOLYTOS has partially been funded by the German Federal Ministry of Economics and Technology (BMWi) in the project HIPPOLYTOS which runs within the "SME sub-programme" of the BMWi research programme THESEUS; KOIOS has been developed within the CTC-WP3 sub-project of THESEUS. Both have been supported by the Ministry of Environment, Nature Conservation and Transport of the Federal State of Baden-Württemberg and by the "Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg (LUBW)", as part of the KEWA cooperation for environmental informatics research. Both have been supported by the Ministry of Economics of Baden-Württemberg which gives a base funding for FZI.

References

1. J. Domingue, D. Fensel, J.A. Hendler (eds), Handbook of Semantic Technologies, Springer-Verlag, Berlin, Heidelberg, 2011.
2. S. Staab, R. Studer (ed.), Handbook on Ontologies (2nd ed), Springer-Verlag, Berlin, Heidelberg, 2009.
3. A. Abecker et al., SUI – Ein Demonstrator zur semantischen Suche im Umweltportal Baden-Württemberg, In: R. Mayer-Föll, A. Keitel, W. Geiger (eds), *Kooperative Entwicklung wirtschaftlicher Anwendungen für Umwelt, Verkehr und benachbarte Bereiche in neuen Verwaltungsstrukturen, Phase IV 2008/09*, Forschungszentrum Karlsruhe, Wissenschaftliche Berichte, FZKA 7500, 2009, 157-166. *In Gernan.*
4. U. Bügel et al., SUI II – Weiterentwicklung der diensteorientierten Infrastruktur des Umweltinformationssystems Baden-Württemberg für die semantische Suche nach Umwelt-informationen. In: R. Mayer-Föll, R. Ebel, W. Geiger (ed), *Kooperative Entwicklung wirtschaftlicher Anwendungen für Umwelt, Verkehr und benachbarte Bereiche in neuen Verwaltungsstrukturen, Phase V 2009/10,* Karlsruher Institut für Technologie, KIT Science Reports, FZKA 7544, 2010, 43-50. *In Gernan.*
5. R. Nedkov, *Schlüsselwortsuche über relationalen Datenbanken,* Diploma thesis, KIT Karlsruhe Institute of Technology, May 2011. *In German.*

6. F. Fonseca, M. Egenhofer, P. Agouris, G. Camara, Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS*, **6**, 231–257, 2002.
7. T. Bittner, M. Donnelly, B. Smith, A Spatio-Temporal Ontology for Geographic Information Integration, *International Journal of Geographical Information Science* 23**(6)**, 765-798, 2009.