Towards a Universal Search in Environmental Information Systems

Clemens Düpmeier[1], Werner Geiger[1], Thorsten Schlachter[1], Rainer Weidemann[1], Renate Ebel[2], and Ulrich Bügel[3]

[1] Karlsruhe Institute of Technology (KIT)
Institute for Applied Computer Science
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
{clemens.duepmeier,werner.geiger,thorsten.schlachter,
rainer.weidemann}@kit.edu
[2] Baden-Wuerttemberg State Institute for Environment, Measurements and Nature
Griesbachstraße 1, 76185 Karlsruhe, Germany
renate.ebel@lubw.bwl.de
[3] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)
Fraunhoferstraße 1, 76131 Karlsruhe, Germany
ulrich.buegel@iosb.fraunhofer.de

Abstract. Full-text search functions in environmental portals make a large amount of environmental data accessible. Many data sources, however, are not suited for indexing by search machines or the data themselves are not suited for access by full-text search. A possibility to make such data of the "dark web" accessible consists in addressing the data sources in the environmental portal directly. The procedure presented here starts with a formal description of data sources (e.g. from the point of view of the portal, these are the target systems). Based on this description, a special component of full-text search, the so-called search broker, can extend and detail a search query, such that all necessary parameters (if possible) are compiled to address these systems and to guide the user directly to the data desired. The presentation component of the environmental portal is responsible for the adequate compilation and display of these data, the so-called result mash-up.

Keywords: Public Search Portals, Semantic Technologies, Result Mash-up, Search Broker, OpenSearch Description, Dark Web.

1 Introduction

Since the adoption of the EU directives on the citizens' access to environmental data in 1990 and 2003 at least, active dissemination of this information has been a duty of environmental administrations [5], [6]. Many authorities provide central environmental portals via which the citizen is given access to data that are often distributed over many individual environmental information systems.

J. Hřebíček, G. Schimak, and R. Denzer (Eds.): ISESS 2011, IFIP AICT 359, pp. 319–326, 2011.

Unfortunately, these portals often provide links to the start or search pages of the target systems only. While integrated full-text search engines offer access to all text information, other types of data like database tables, geographic information systems, and multi-media files often are not covered by these full-text search engines or they cannot be represented adequately in the list of results.

Huge parts of many environmental information systems have to be connected by specialized adaptors or remain hidden to the users of environmental portals [11].

In practice, very few systems offer a real semantic description or links of data in terms of the semantic web [2]. A "real" semantic search is not yet possible. Analysis of environmental information systems in the state of Baden-Württemberg has even revealed differences in the semantics of individual terms in several systems of a single authority, although this semantics is never expressed explicitly.

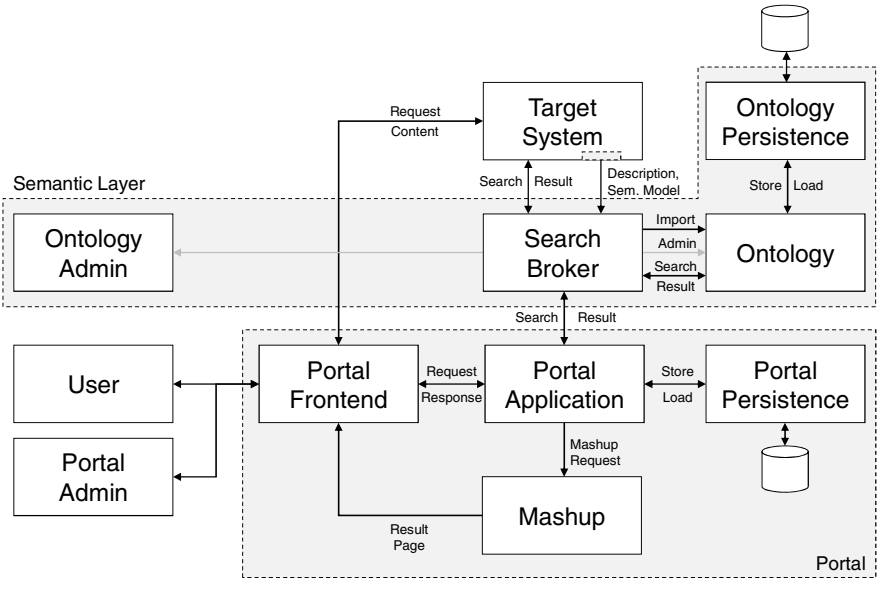

Fig. 1. Architecture of the Environmental Portal with SearchBroker, Semantic Component (Ontology), and Target Systems

The idea of the approach presented here is to deposit some semantics in the description of systems and, thus, to make queries that are as intelligent as possible.

Search requests to environmental web portals typically consist of the following three components:

1. Subject
2. Spatial reference (e.g. coordinate, administrative unit, professional object)
3. Temporal reference (e.g. point or period of time)

Very often, queries refer to one or more topics alone or in combination with a spatial reference. Queries relating to temporal references, also in combination with a subject or spatial reference, are rare.

Mobile end devices, such as smart phones or tablet PCs often allow for a more or less precise determination of the own location (and, hence, of the location of the user) with components like GPS receivers, WLAN, and mobile radio communication devices. This location information is available in principle as a context of the search query, even though these data are not input explicitly by the user.

Within the project "Semantic Search for Environmental Information (SUI)" [1], [3], an architecture was developed (Figure 1), which

- provides and processes descriptions of various target systems,
- delegates preprocessing of search terms to specialized components for environmental issues, spatial, and temporal references, and
- presents multiple data formats in an integrated view (mash-up).

The core of this architecture is the "SearchBroker" that acts as a search engine for the environment portal.

2 Target System Descriptions

By a description, the individual target systems are made known to the SearchBroker (Figure 2). An extension of the OpenSearch description XML format [9] is used as a vehicle for these descriptions.

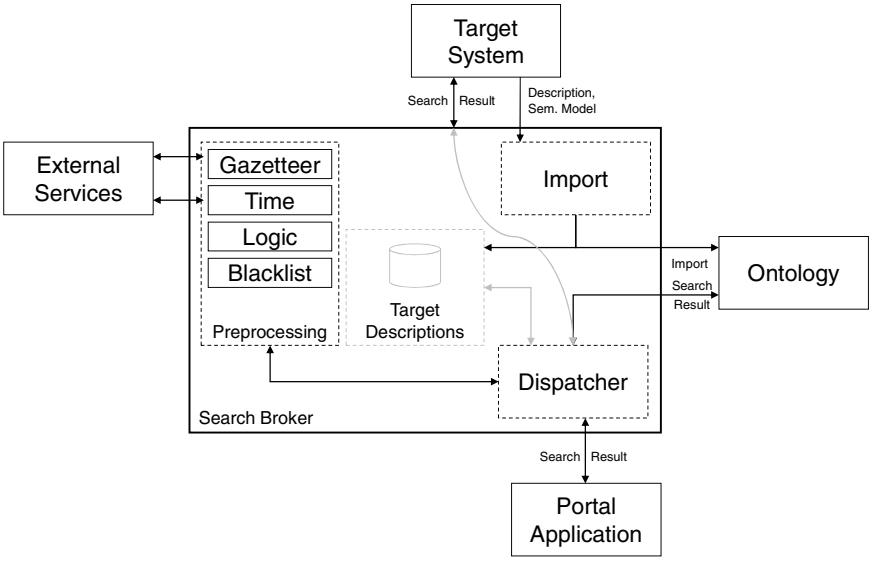

Fig. 2. Architecture of the SearchBroker with Plug-ins and Semantic Component (Ontology)

This format is based on URL patterns[1] which may contain wild cards for all necessary parameters (see Listing 1).

```
<Url type="text/html" template=
"http://foo/?q={searchTerms}&amp;pw={startPage?}"/>
<Url type="application/atom+xml" template=
"http://foo/atom?q={searchTerms}&amp;pw={startPage?}"/>
```

Listing 1. OpenSearch URL templates

Generally, several URL patterns may be contained in an OpenSearch description. The patterns are distinguished by giving the reply format (type). If several reply formats are available, the application may decide which to use and select the corresponding URL pattern.

URL patterns in OpenSearch descriptions normally refer to the addressing of search engines. This mechanism can be extended easily by the admission of any parameters instead of the standard parameter set. While the semantics of standard parameters is set already, the "free" parameters have to be explained more explicitly. Again, the extension of an OpenSearch description element is used, as shown in Listing 2.

```
<Url type="text/html" template=
"http://foo/dienst?repId=biotope&amp;value={commune}"/>
<Query role="substitution" ui:name="commune"
ui:type="geo:commune:id"/>
```

Listing 2. URL template and corresponding semantic description

The "query" element is extended by two attributes containing the name of the variable ("ui: name", here: "commune") and a unique identifier for the semantics of this variable ("ui: type", here: "geo: commune: id").

The unique identifier has to be known to the search broker and to be used by the associated plug-ins for preprocessing of the query.

An OpenSearch description may contain multiple URL templates that may differ in the target format given by the "type" attribute. This enables the environmental portal to retrieve data from a target system in different formats, e.g. HTML, XML, GeoRSS, or Atom.

In this sense, even a full-text search engine can be treated as a target system. It can also be described by an OpenSearch description and provides a corresponding hit list, e.g. "Application/atom+xml" or "application/rss+xml". In practice, many websites will not be described as specific target systems, but they will be indexed and queried by a full-text search engine.

In the case of the state environmental portals [10], it may be reasonable to define certain tasks previously executed by the full-text search machine as separate target systems. The Google Search appliance [8] used as full-text search machine is capable

[1] Addressing systems via URL indeed excludes some systems (e.g. such as those based on web services). If required, however, these can be integrated via appropriate adapters.

of submitting queries to other systems parallel to the search in the own index. This mechanism called "OneBox" delegates the search query to other systems. If they reply within a defined period of time, these results are delivered in addition to the search results from full-text index and can be represented near them by the portal (see right column in Figure 3).

If these OneBoxes are addressed by the search broker or the mash-up component of the environmental portal and not by the search machine, many unnecessary queries can be avoided, as it is possible to decide which queries are promising or not after preprocessing the search query already (Section 3).

3 Preprocessing

The search broker knows the descriptions of all systems connected to the environmental portal and, hence, the syntax and semantics of their calls. Upon receipt of a query by a user from the environmental portal, the search broker has to provide for all relevant parameters being available. Otherwise, target systems cannot be called up.

To identify the semantics of a given query, the SearchBroker is assisted by a series of specialized plug-ins. The plug-ins analyze each constituent (e.g. single words or a series of words) of the query and try to allocate an explicit semantics to them.

In this way, one or more gazetteer plug-ins can resolve the spatial reference in a query. Many internet search engines use gazetteer services available online from different vendors. Some environmental agencies also offer specialized gazetteer services which, for example, can resolve field names, names of water bodies or names of natural areas in addition to place names.

In the above example, a gazetteer plug-in can recognize a place name, e.g. "Karlsruhe", and allocate several properties to it, such as:

- `geo:commune:name = Karlsruhe`
- `geo:commune:id = 08212000`

The latter can now serve as a parameter for addressing the target system shown in Listing 2.

Another gazetteer plug-in may supply further information about the same place name "Karlsruhe":

- `geo:lon = 8.4037563`
- `geo:lat = 49.0080848`
- `geo:bbox = 8.2756969,48.9494975 8.5318157,49.0666033`

Similarly, another plug-in can explain temporal terms. For a search, including "summer 2010", the explicit start and end dates are assigned:

- `datetime:calendar:day:first = 2010/06/21`
- `datetime:calendar:day:last = 2010/09/22`

Apart from spatial and temporal references, resolving of environmental issues is the biggest challenge in the preprocessing of search terms. Resolution aims at mapping search terms onto one or more elements of a well-defined vocabulary, as it is used in

the connected target systems. For the purpose of restricting or expanding a query, also the neighborhood of each issue is supplied to the SearchBroker, e.g. synonyms as well as superordinate and subordinate concepts.

The environmental issues are modeled in an ontology. This ontology initially is restricted to the terms of a certain domain (here: "Umwelt" (environment)) that reflects the contents of the portal. The ontology consists of several partial ontologies. The backbone is the GEMET environmental thesaurus [7]. It is extended by a thematic catalog, the entries of which contain further metadata, e.g. frequently used key values. This structure is covered by another partial ontology that contains so-called "life situations". This corresponds to the approach of many portals to meet the citizen in a life situation and to guide him from this situation to the relevant services and data [12]. These partial ontologies are linked by mapping. Due to this mapping, the environmental issues found generally contain entries from all partial ontologies. Hence, a maximum of information is available for queries by target systems. A more detailed description of this semantic approach is given in [4].

4 Presenting Search Results (Mash-up)

After the completion of preprocessing, the SearchBroker can decide which target systems are available and how to request them. The SearchBroker can now query the data by itself or return the full addresses to the environmental portal. Currently, the latter approach is being used.

Within the environmental portal, a mash-up component is responsible for the presentation of search results. Depending on the results supplied, it can decide how these will be presented.

Essentially, the following target formats are distinguished:

- spatial data, e.g. displayed by a web map client
- links, e.g. in the form of link lists
- tabulated data and charts that may be converted into HTML
- multi-media contents, for example in the form of a gallery view
- text messages (e.g. RSS), for example in the form of summary lists
- HTML pages, HTML fragments and microformats displayed at certain points in the layout
- results of a full-text search, for example in the form of hit lists

The target formats are assigned using the MIME types given in the target system descriptions. In case more than one target format are available, the mash-up component can decide how the hits will be displayed. The same data may be displayed at several places, e.g. within a hit list and a map view.

Furthermore, the gathered information may be used to provide the user with additional navigation steps in the portal. These include, for example, the restriction or expansion of the query based on subordinate or superordinate concepts (Figure 3). It is also possible to resolve ambiguities resulting from the preprocessing of place names.

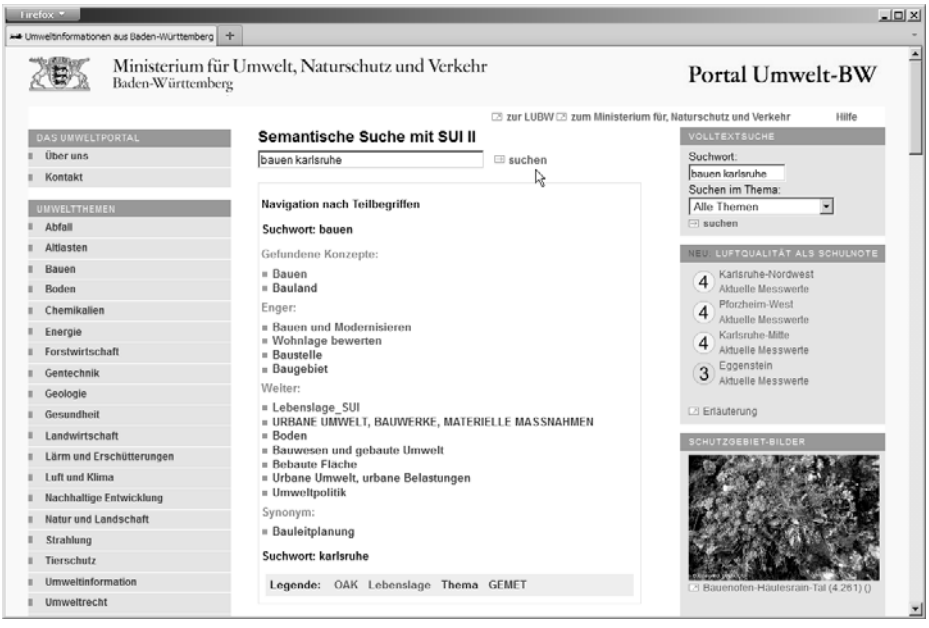

Fig. 3. Screenshot of the SUI II prototype: Subjects gathered from the ontology component for the provision of additional navigation steps within the full-text search results.

5 Conclusion and Outlook

The architecture proposed provides environmental portals with the opportunity to access different target systems with different target formats. The description of the target systems and their addresses are stored in an XML document that is based on and extends the OpenSearch description format.

The parameters necessary for addressing the target systems are obtained from the search query by various specialized plug-ins during preprocessing. An ontology component is included for the resolution and explanation of the thematic reference.

The complete addresses for target system requests are supplied to the environmental portals together with format information. Depending on the formats supplied, the portals display an integrated results page.

Thus, a broad variety of formats can be represented adequately within the search facility of an environmental portal, e.g. map layers with nature reserves, geo-point objects, such as measuring points, full-text search results, tables, or multi-media data.

This “universal search” approach offers a significant value added compared to a conventional full text search.

References

1. Abecker, A., et al.: SUI - Ein Demonstrator zur semantischen Suche im Umweltportal Baden-Württemberg, UIS Baden-Württemberg. In: F+E Vorhaben KEWA, Phase IV 2008/09, Wissenschaftliche Berichte, FZKA-7500, pp. 157–166 (Juli 2009)

2. Berners-Lee, T.: Semantic Web Road Map (1998), http://www.w3.org/DesignIssues/Semantic.html (visited January 19, 2011)
3. Bügel, U., et al.: SUI II - Weiterentwicklung der diensteorientierten Infrastruktur des Umweltinformationssystems Baden-Württemberg für die semantische Suche nach Umweltinformationen, UIS Baden-Württemberg. In: F+E Vorhaben KEWA, Phase V 2009/10, KIT Scientific Reports, KIT-SR 7544, August 2010, pp. 43–50 (2010) ISBN 978-3-86644-540-6
4. Bügel, U., et al.: Leveraging Ontologies for Environmental Information Systems. In: Hřebíček, J., Schimak, G., Denzer, R. (eds.) ISESS 2011. IFIP AICT, pp. 372–379. Springer, Heidelberg (2011)
5. European Union, Richtlinie 90/313/EWG des Rates vom 7. Juni 1990 über den freien Zugang zu Informationen über die Umwelt (1990), http://www.umwelt-online.de/recht/allgemei/90_313gs.htm (visited January 19, 2011)
6. European Union, Richtlinie 2003/4/EG des europäischen Parlaments und des Rates vom 28. Januar 2003 über den Zugang der Öffentlichkeit zu Umweltinformationen und zur Aufhebung der Richtlinie 90/313/EWG des Rates (2003), http://www.umwelt-online.de/recht/eu/00_04/03_4gs.htm (visited January 19, 2011)
7. GEMET, http://www.eionet.europa.eu/gemet (visited March 23, 2011)
8. Google Search Appliance, http://www.google.com/enterprise/search/ (visited March 23, 2011)
9. OpenSearch, OpenSearch description document, http://www.opensearch.org/Specifications/OpenSearch/1.1#OpenSearch_description_document (visited January 19, 2011)
10. Schlachter, T., et al.: LUPO - Ausbau der Suchfunktionalität der Landesumweltportale und Vernetzung mit dem Umweltportal Deutschland, UIS Baden-Württemberg. F+E Vorhaben KEWA, Phase V 2009/10, KIT Scientific Reports, KIT-SR 7544, 9–20(August 2010) ISBN 978-3-86644-540-6
11. Schlachter, T., et al.: Erschließen von Datenbank-Inhalten durch die Volltextsuche in Landes-Umweltportalen, Umweltinformationssysteme: Suchmaschinen und Wissensmanagement - Methoden und Instrumente. In: Workshop 'Umweltdatenbanken/ Umweltinformationssysteme', Dessau-Roßlau, June 5-6, Umweltbundesamt, Berlin (2009)
12. Service-BW, http://www.service-bw.de (visited March 23, 2011)