

From Simple Data Sources to a Complex Information System: Integrating Heterogeneous Data Models into an Information Infrastructure for the Public Administration

Franz Schenk¹, Fred Kruse¹, Martin Klenke²

*¹Coordination Center PortalU, ²Fachgruppe Umweltinformationssysteme
at the Lower Saxony Ministry of Environment and Climate Protection*

Archivstraße 2, D-30169 Hannover

kst@portalU.de , martin.klenke@numis.niedersachsen.de

Abstract

Granting access to governmental data has become a matter of public interest since long. There are national and international regulations like the German environmental information act (UIG) or the European infrastructure for spatial information (INSPIRE). Both encourage governmental institutions to make information available to the public.

Despite the fast-paced development in information technology many problems with data integration are far from being solved. The availability of distributed data demands for standardised procedures in service discovery and invocation, in data processing and integration. Although there are many efforts for a standardisation of web services and data exchange formats, on the level of the administrative institutions heterogeneous data structures and service interfaces are still the normal case. Hence, the complexity of granting access to environmental information in a uniform way with respect to the aforementioned regulations is very high.

The German environmental information portal *PortalU* (www.portalU.de) is a publicly financed information infrastructure. It offers a single point of entry for all information with environmental relevance held by public authorities. Here, the key problem is not how to deliver the information or how to present the results of a query, but how to integrate the plethora of heterogeneous data sources. The aim is to allow for the integration of data in as many formats as possible and give access to all sources with a uniform query mechanism. The architecture of PortalU contributes to this situation with its flexible and extensible concept of specialised data source plug-ins. From the point of view of the query facility in PortalU, all information sources are treated uniformly as they all deliver their search results in the same way (the ranking of the results can be influenced, though). Data sources of very different structure can be connected, indexed and queried. The key component is the iBus, a communication broker that distributes queries to all connected data sources and collects, combines, and delivers the answers in return. One of the strong points in the architecture of PortalU is the wide range of data formats and the diversity of information systems that are supported.

1. Introduction

Queries to PortalU (Vögele et al. 2007) are answered based on a set of highly heterogeneous data sources, all from public administrations, all with relevance to environmental subjects, comprising

- Web pages (more than 3 million web pages),
- Metadata about environmental information objects,
- Research reports,
- Maps,

- Measurement readings, e.g. on air pollution, water quality, or radiation,
- OpenSearch result sets.

Because of the different nature of the data sources, PortalU deals with highly heterogeneous data structures. Web pages are relatively unstructured and text-centred whereas metadata that can be retrieved from metadata catalogues from most of the federal states, are highly structured in accordance with a well-defined data model. Moreover, a wide range of data storage formats are supported, ranging from highly structured relational database systems to semi-structured data stemming from XML databases or single XML documents. Even less sophisticated (from the point of view of database development) data storage formats like excel sheets can be integrated.

PortalU acts as an information broker and offers its services in form of an internet portal. New data sources are analysed with respect to the index structure in PortalU. By data source-specific mappings, all relevant information can be extracted and integrated into the index. The quality in query answering relies heavily on the specification of the mappings.

This paper describes the range of differently structured environmental information from public authorities and how the architecture of PortalU allows for the integration of all of the different information sources. First, an overview is given about the different kinds of data sources that are supported by PortalU. Then, the architecture is described in detail, followed by a description of the integration process. The paper concludes with an outlook on future developments.

2. Classes of Data Sources

PortalU integrates a wide range of data sources and offers uniform access to environmental information from the public administration. Hence, users are both people seeking and public authorities publishing information. The variety in data formats that can be accessed reflects the variety in information infrastructure in public administrations. There are administrative units with large-scale information systems, but also small public authorities with simple, often basic ways of dealing with data management. PortalU offers solutions for most of them. There are some information sources that are also supported but not described in this paper, e.g. SemanticNetworkService, RSS feeds, or the almanac. These information sources are not in the scope of this paper as they are available only in the portal and not in the result set that is returned by PortalU.

In the following, a characterisation is given for the most important data sources that are included in the architecture of PortalU.

Web Pages

PortalU includes a vast amount of selected web pages in its index. Although only information from public authorities is considered here, the web page index comprises more than 3 million entries. This pool of web pages is cultivated by the partners (which are, at the same time, the principals) of PortalU. Domain experts from all associated institutions provide a list of sites that characterise the crawling process, including web addresses as entry points, (patterns of) web addresses that should not be followed, and another class of addresses that are ignored completely. These address specifications define the set of web pages in the PortalU index. Thanks to the editorial work of experts the content of the web index of PortalU has an outstanding relevance with respect to environmental subjects.

Metadata

Nearly all (15 out of 16) federal states in Germany maintain metadata catalogues about information objects with environmental relevance. Moreover, also governmental institutions like the *Umweltbundesamt* or the *Bundesamt für Naturschutz* have extensive collections of metadata. These metadata are highly structured in accordance to a well-defined data model. Many catalogues use the editor and catalogue software *InGrid* (Klenke et al. 2007). Although *InGrid* can be employed independently from PortalU for the maintenance of collections of metadata information, it is typically used in conjunction with PortalU. Thanks to the tight coupling, *InGrid* metadata catalogues can be integrated into PortalU ad hoc, while other metadata catalogues need specific interfaces and individual schema mappings, depending on the DBMS that is used.

OpenSearch Result Sets

Another possibility for the integration of data sources is the use of OpenSearch (A9 2010) result sets. Here, a wide-spread and well-understood technology is employed for the communication of queries and search results. In this context, PortalU is not only an information broker but also a meta-search-engine.

A data source, which is connected via OpenSearch, can be any kind of information source. For example, an OpenSearch interface may connect a relational database, but also a web index can be connected. Some data sources already come with an OpenSearch interface. This is the most convenient situation. Using an existing infrastructure, there is no need to take care of the local data structure, the mapping definition, or the ranking scores.

In addition to that PortalU offers its own OpenSearch server component. This is a software installation that is installed remotely and connects to a local database. Implemented as a web service it offers, following a predefined mapping schema, an interface that returns query results from the underlying database in form of OpenSearch result sets. This interface gives complete control over the mapping schema and the ranking of the results. Note that, with respect to the mapping of local schemas to the PortalU index, OpenSearch server installations are quite similar to data source clients, which are explained in the next section.

In PortalU, OpenSearch data sources are also used in conjunction with the main web index, as some of the partners of PortalU already have an infrastructure for web indexing and make their index available through an OpenSearch interface.

Databases

Data source clients (DSCs) are adapter installations that directly connect to the data source. The installation can be either local (part of the PortalU service infrastructure) or remote (on the same server as the data source). The difference to OpenSearch server installations is that search results are returned in a format proprietary to PortalU. Hereby, the mapping to the fields of the PortalU index is much more flexible. With DSCs, the mapping from the data model to the result set is provided by the client, which communicates directly with the PortalU system (the iBus). The client generates a local index, which is an abstraction layer between any kind of database schema and the structure of the PortalU search index. This index is used for the answering of queries.

A wide range of information systems can be integrated in this way. For example, most of the common relational database management systems (RDBMS) can be connected. A mapping client, which connects to a RDBMS via JDBC, creates a local index upon the relational data according to the mapping schema. Queries to the client as well as the results are expressed in a specific format proprietary to PortalU.

While RDBMS often cover highly structured information with large numbers of entries, there are also information sources like Access databases and Excel sheets. The latter are widely distributed but, with

respect to data integration, often neglected. Although there are good reasons to limit the support of information systems to robust and scalable information systems, it can be useful to extend the range of supported database types to more simple data sources. Especially for small authorities with little resources and expertise in information technology it is of advantage to be able to integrate the data as it is. Again, the same principles apply as with other data source clients: A mapping is defined, which specifies the parts of the (local) data structure that will be mapped and made available to (global) queries.

XML Data Sources

The architecture of PortalU emphasises the integration of XML data. As XML has become a standard in information interchange, the need to support XML is obvious. For this reason, various kinds of XML-based data sources can be integrated. The most specific interface connects Web Services that implement the CSW (Catalogue Service Web) standard. These services can be integrated ad-hoc, as PortalU internally has a predefined mapping from CSW schema to its own index structure. In addition to that, there is support for specific XML databases (e.g. Tamino) and (a more generic) support for plain XML documents collections. The mapping of XML data is realised by a full support of XPath expressions. The query language XPath is very well suited for data extraction and data transformation, which are the key prerequisites in the process of data integration.

By the specification of mappings for information from all of these sources into a distributed, but homogeneous index structure, PortalU offers a uniform view on all these different sources. The architecture of PortalU is explained in the next section.

3. Architecture

PortalU as a service portal integrates many different kinds of data sources from public authorities with environmental relevance and offers a one-stop access via several different query interfaces. Before going into the details of data integration in PortalU, a brief overview is given on how information can be retrieved from PortalU.

3.1 Query Interfaces

PortalU offers several interfaces for searching the distributed data sources:

- Form-based interactive search portal: The portal presents a search form, which can be used for simple keyword searches, but also offers a wide range of (partly interactive) expert features which are available for the refinement of a query, e.g. by temporal or spatial constraints.
- Opensearch: PortalU offers an OpenSearch interface (Klenke et al. 2010), which basically accepts the same query syntax as the portal search interface. The main difference is that the result sets are delivered in a specific XML-syntax, which is optimised for machine readability and thus for reusability.

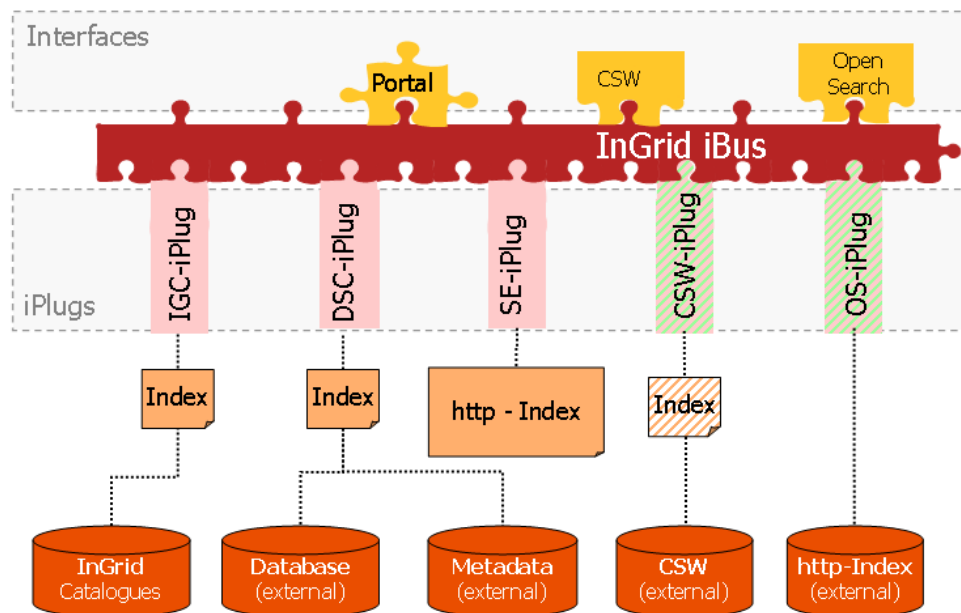


Figure 1: PortalU / InGrid Architecture. Note that the index for external CSW sources is optional but recommended for better search performance. InGrid catalogues are metadata catalogues with a more specific interface (IGC-iPlug) compared to other metadata catalogues (DSC-iPlug) and come with a pre-defined mapping.

- CSW interface: A metadata information retrieval interface, which delivers the result set in a highly structured way, compliant to the standards of CSW AP ISO 1.0 (OGC 2010) and INSPIRE (JRC 2009).

These interfaces clearly serve different purposes. Whereas the interactive search, which is available with the web page of PortalU, is obviously addressed at human users, the purpose of the OpenSearch and CSW interfaces is to allow for the integration of PortalU in web service infrastructures. Furthermore, the query languages of CSW and PortalU are very different. Queries in CSW are designed for searching in a single catalogue and use a very limited set of queryables for the filtering of the results, whereas the query language of PortalU is much more expressive and tailor made for a distributed information system. For example, the search can exclude (or be limited to) specific data sources or categories of information. The advantage of the CSW interface lies in the standardised access to information.

3.2 The iBus: Result Set Harmonisation

The central communicational component in the architecture of PortalU is the iBus. New data sources have to register at the iBus following a proprietary communication protocol. Incoming queries become distributed to all connected data sources, the answers are collected in return. The collected result set is treated in the following way:

- Duplicates are filtered, as the same result may originate from several sources.
- Results are ordered with respect to ranking scores.
- Results are grouped according to the specifications given in the query.
- Filters that are specified with the query are applied with respect to datatypes, domain, topic, value-ranges, or data source.

- Afterwards, the result set is passed to the query interfaces.

Information from remote sources can only be found if the native data structure can be mapped to a searchable index structure. This mapping process is described in the next section.

3.3 Data Integration: Schema Mappings

PortalU is designed as an information broker. Therefore, instead of fully integrating data sets, relevant parts of a data set are identified and made available for later search. This information system is realised by the definition of views on data sources. Consequently, information is integrated strictly virtually in the sense of a federated database system. Sometimes, however, for reasons of performance, the views on a data source are materialised in a local mirror. This is useful, for example, with CSW data sources, where query answers typically consist of larger amounts of data compared to other data sources.

This concept of information integration in PortalU relies on two key components: a uniform index structure (that is described in the next section) and schema mappings for each data source. The schema mappings specify how parts of the original data structures correspond to fields in the index. The mappings do not only define what kind of information is used for relating a search term to a data set but also what and how much information is returned with the search result. The amount of information that is contained in a search result as delivered by PortalU depends strictly on the mapping: It might consist solely of a link that allows retrieving the data set from the original data source, but it might as well consist of all parts of the original data set.

In the following we distinguish between data-centric and document-centric sources. This discrimination is well-known from a categorisation of XML data (as *semi-structured data*, XML data sources can be highly structured or virtually unstructured), but it is also suitable for the classification of data sources of PortalU.

Mappings for highly data-centric sources

The schema mappings reflect the data structure of a data source. For example, XPath expressions are used for the identification of relevant parts in XML data. Typically, the navigational elements of an XPath expression are used with well-structured XML. But there are also built-in string functions that can be used for less structured XML documents.

Mappings for relational databases are realised in form of SQL expressions. SQL expressions are per se highly structured, only string operations allow some degree of ambiguity. The complexity of a mapping therefore depends on the complexity of the underlying database. For example, a pre-defined mapping for an InGrid metadata catalogue relates over 250 fields from more than 50 tables of the underlying relational database to fields in the search index. Note that the DSCs also offer a graphical user interface for the specification of a mapping. This interface allows to interactively define the tables, relations between the tables and eventually those columns (or even down to the tuples by applying additional filters) that are needed in the mapping.

Mappings for document-centric sources

With a decreasing degree of structure the analysis of content becomes more important. When integrating relatively unstructured data like HTML pages as an information source, the content of the atomic information unit, the web page, has to be analysed intensively. The building of an index of selected web

pages is therefore more complex. The total of collected web pages is determined by three parameters: the entry points (start-URLs) for the crawling process, the link patterns that should be followed (limit-URLs) and link patterns that should never be followed (exclude-URLs). Collected HTML pages are fetched, indexed and semantically analysed by a combination of *Nutch search technology* (Nutch 2010) and the Semantic Network Service (SNS 2010) of the UBA.

The next section gives some details about how information is retrieved using a uniform index structure.

3.4 The Search Index

PortalU uses a distributed index structure where each connected data source maintains a search index of its own. All the indexes are built using the *Lucene* (Lucene 2010) framework and are therefore, with respect to the technical realisation, identical. However, the important aspect is not the technological framework itself. Rather, the fact that all components in PortalU use the same heuristic for building and maintaining a search index and ranking of the results.

Queries are evaluated using this collective of search indexes. A prerequisite for this operation is the structural uniformity of the distributed indexes. An index represents sets of information objects of the underlying data source. The exact nature of an information object is defined by the schema mapping. It can be given, for example, by a set of SQL queries, where each query maps a database entity to an index field. Each information object has to comprise of at least a title and an abstract. Optionally, an arbitrary number of additional properties can be assigned, each adding a new field to the index.

The query syntax in PortalU offers some special search keywords, e.g. spatial constraints can be expressed. These query parameters are related to specific index fields, which can be defined by the schema mappings of the data source clients. If these fields do not exist, the data source client will return no answers to queries with such parameters. For example, if a database entry *datasetlanguage* of a data source is mapped to an index field named *data_lang*, the query ‘air pollution data_lang=eng’ would only return answers from data sources that specify this field and answers would be limited to those items that have the appropriate value ‘eng’.

Another prerequisite for a distributed index is the ability of each single index to add a ranking score to the results. Given that the algorithm for the determination of the ranking score is the same for all data source clients, the iBus can deliver the collected results as a sorted result set to the query interfaces.

The ranking and the composition of the result set can be influenced in several ways. Firstly, a data source can be promoted by a bonus on the ranking score of all of its results. Secondly, each index field can be emphasised (or weakened) so that it has greater influence on the ranking score of the result. Moreover, filters can be defined, which exclude information objects inside or outside a specified range of values for a certain index field.

There are several ways in which the results to a query can be delivered. Usually, search engines return their results as links to the original content. This is also the case in PortalU for the results from one of the web indexes. Moreover, direct links to services like map viewers can be provided. Search results from the metadata catalogues, however, are returned as key-value pairs containing all information that is defined by the mapping. The OpenSearch interface delivers these results in an XML structure, while the portal dynamically generates a web page that is a transformation of the results into a human-readable form. The results delivered by the CSW interface are always complete CSW records, though pre-defined subsets of full records can be requested.

4. Discussion and Outlook

We described the basic steps that are needed for the integration of a new data source in PortalU. Especially with respect to web indexes, the OpenSearch interface offers a very convenient solution. It allows for a distributed web index and a combination of different search technologies. The main disadvantage of this solution lies in the fact that the algorithm, which determines the ranking score of the (remote) search results cannot be controlled. The only way to influence the measurement of the quality of search results from an OpenSearch interface lies in the manipulation of the score *after* the results are returned. This, however, can only be achieved in a static way for all results of the same interface, independent from the context of the query. Dynamic adjustments to result sets in relation to the ranking algorithm of PortalU are not yet possible in reasonable time during the generation of an answer to the original query. For this reason, the integration of a new web index demands for intensive testing of the search results and careful fine-tuning of the parameters of the interface.

We have shown how PortalU centralises access to information systems for environmental data from public administrations. Due to international legislation like INSPIRE and a growing interest in web service orchestration in general, the CSW interface of PortalU will be in constant development. At the time of writing, data exchange via CSW in the face of INSPIRE is resounded throughout the land, but in fact this infrastructure for data exchange is still in heavy development and will be a construction site for years to come. Therefore, besides the never-ending improvements in usability, better support for specific data sources, and better performance in general, the challenges in web service interoperability will be one of the main tasks for the PortalU team in the next future.

5. Summary

The aim of PortalU is to integrate environmental information from public authorities, independently from the underlying data formats. Although the integrated data itself is highly heterogeneous, standardised query facilities can be offered. Besides the form-based search in the portal, OpenSearch and CSW are supported as prominent interface standards.

The architecture of PortalU is designed to be flexible and extensible, new data sources can be added easily. A mapping client suitable for the data structure of the new source has to be installed, either locally or remotely. Additionally, a mapping from the source schema to the search index has to be defined.

One of the main requirements for data integration in PortalU is that authorities owning the data should have to apply only as little changes to their information systems as possible. This is achieved by standardised index structures and mapping clients that can be easily installed and act as interfaces to PortalU. The range of supported data formats covers most of the relevant information sources that are maintained by public authorities: HTML web pages, relational databases, XML databases and documents. It is one of the main achievements of PortalU how structurally highly diverse data sources are integrated into a homogeneous search index.

6. Literature

A9: OpenSearch Spezifikation, Internet: <http://www.opensearch.org/Specifications/OpenSearch/1.1>, last visited 01.07.2010

JRC European Commission Joint Research Centre: INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119, Version 1.1, 2009.

Lucene, Internet: <http://lucene.apache.org/>, last access: 01.07.2010, last visited: 01.07.2010

Klenke, M., Vögele, T., Kruse, F., and Lehmann, H.: PortalU® und InGrid® - Werkzeuge zur Erstellung, Recherche und Verteilung von Metadaten - In: Strobl, J., Blaschke, T. und Griesebner, G. (Hrsg.): Angewandte Geographische Informationsverarbeitung XIX, Beiträge zum AGIT-Symposium, Salzburg, 2007.

Klenke, M., Kruse, F., Schenk, F.: OpenSearch: Simple formats to share environmental information. EnviroInfo 2010, Bonn.

Nutch, Internet: <http://lucene.apache.org/nutch>, last visited: 01.07.2010

OGC Open Geospatial Consortium: OpenGIS® Catalogue Services Specification 2.0.2 – ISO Metadata Application Profile, Version 1.0. OGC document 07-045, 2007.

SNS: Semantic Network Service, Internet: http://www.semantic-network.de/doc_intro.html?lang=en, last visited: 01.07.2010

Vögele, T., Klenke, M., Kruse, F., Lehmann, H., and Giffei, C.: An Effort to Achieve Organizational Interoperability of Environmental Information - PortalU® and the Environmental Information Infrastructure in Germany. - 6th International Symposium on Environmental Software Systems (ISESS07), Prague, 2007.