

## **An Effort to Achieve Organizational Interoperability of Environmental Information - PortalU® and the Environmental Information Infrastructure in Germany**

T. Vögele, M. Klenke, F. Kruse, H. Lehmann, C. Giffei  
Coordination Center PortalU  
at the Lower Saxony Ministry for Environment  
portalu@portalu.de

### **Abstract**

PortalU®, the new German Environmental Information Portal, provides one-stop access to publicly-held environmental information. The portal is a central access point to heterogeneous and geographically as well as organizationally distributed information sources. A user-friendly interface features advanced search- and visualization tools to enable experts and non-experts alike to find and view texts of national and regional legislation, information about environmental policies and programmes, environmental reports, monitoring data, digital maps, and many other types of environmental information and data. PortalU® is part of the administration's strategy to comply with EU-Directive 2003/4/EC (EU 2003), which calls for better public access to environmental information. The portal helps citizens to find relevant information about the national and regional environment in a fast and effective way.

### **1. Introduction**

Expert systems, modelling tools, GIS, and other information systems play an important role for problem solving and decision making in the environmental field. Environmental- and computer scientists that develop these tools typically focus on aspects of information processing and user interaction, while the availability of adequate input-data is often seen as a given. In reality, however, the necessary data may exist but they may not be available at a specific time to solve a specific problem. The reason for this shortcoming is closely related to a lack in technical, semantic and organizational interoperability of environmental data and information. Of these three, technical interoperability is the easiest to achieve. With the recently passed INSPIRE

Directive, the EU goes an important step forward in the direction of technical interoperability, mainly in the sector of geospatial data, but also including a large amount of environmental data. Solutions to problems of semantic interoperability are still in the realm of research, and initiatives like INSPIRE do address them only marginally. Organizational interoperability, on the other hand, is neither addressed by science nor by international initiatives as it is often assumed to follow the achievement of technical interoperability. However, long-term practical experience with the management of government-owned environmental information and data in Germany show that this is a rather optimistic assumption and that the integration of environmental information distributed among a large number of government agencies is hampered by many organizational obstacles.

To solve this problem, the environmental administration in Germany has established an organizational framework that supports a technical information infrastructure aimed at the harmonized and integrated access to information, data and metadata. The Environmental Data Catalogue (UDK) is in use since the late 1990's (Swoboda 1999, Karschnik 2003). Today, it contains more than 20.000 records of environmental datasets in 16 federal states and several federal agencies. With the Environmental Information Network (gein), an effort was made to establish national common gateway to environmental information (Vögele 2004). Since June 2006, both UDK and gein have been intergrated and replaced by PortalU® (www.portalu.de), the new central Internet-portal to publicly-held environmental information in Germany. In both the technical development and the content-selection for PortalU®, stake-holders from the environmental agencies in all 16 federal states and the federal government are involved. Technically, the new information portal relies heavily on state-of-the-art Open-Source tools and newly developed data-source interfaces that allow to tap into different types of data sources including websites, data catalogs and databases. Ontology-based semantic services are used to enhance automatic metadata annotation and user-friendly query expansion. This paper will describe the main features of the organizational and the technical set-up of PortalU®.

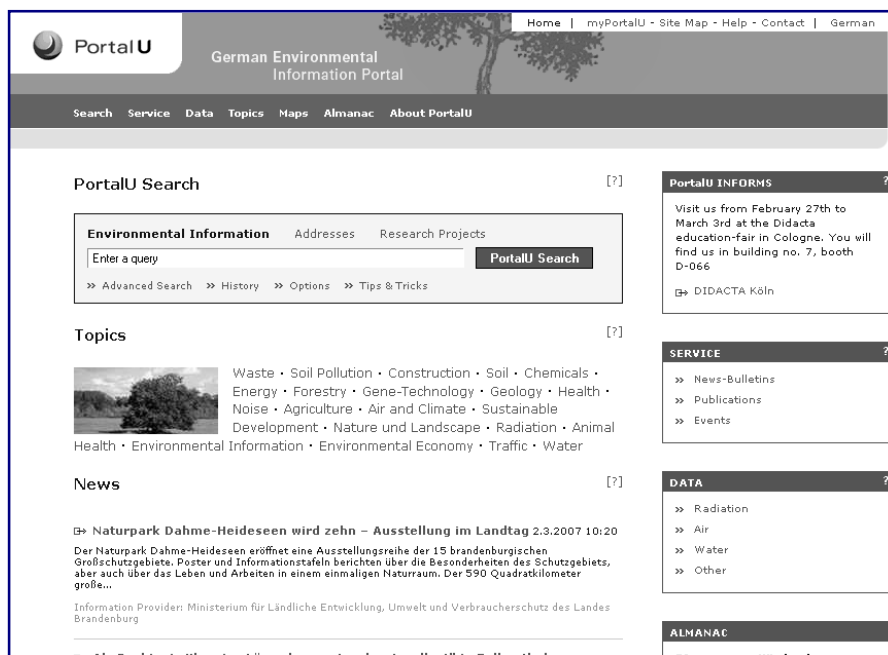


Figure 1: PortalU® start-page

## **2. Improved Access to Environmental Information**

Even before the INSPIRE Directive was adopted, PortalU® was designed to help the environmental (and other) administration in Germany to comply with another piece of EU legislation, EU Directive 2003/4/EC on Public Access to Environmental Information (EU 2003). A key requirement of this directive is the provision of a structured and simple access to publicly-held environmental information. Unlike in other European countries, environmental information in Germany is within the responsibility of many different agencies, on all levels of the administrative hierarchy (federal, state, and municipal). PortalU® thus tries to meet the requirements of the Directive by providing a central access point to information hosted by the numerous public authorities dispersed throughout the country. As of spring 2007, PortalU® has access to more than 140 federal- and state agencies, and the goal is to ultimately include all public agencies that hold environmental information, including those on the municipal level.

As an Internet-based information system, PortalU® is open to the public on a 24-hour basis. The portal tries to make access to information in a structured and simple way by providing two main functionalities: a search function for unstructured, query-based information requests, and a catalog-function for structured, theme-based information browsing.

### *2.1 Search function*

The PortalU® search function features a simple-search mode, and an extended-search mode for expert queries. Experience with PortalU®'s predecessor, the German Environmental Information Network (gein), showed, that the simple search function attracts more users than the extended-search function. Nevertheless, the latter is important because a (small) group of expert-users rely on its functionality to specify complex information requests.

One feature of the extended-search mode is the ability to limit the search-space to specific data sources. For example, the user may want to search only the (meta)data catalogs connected to PortalU®. Or the information request may be limited to web-pages only.

Another key feature of the extended-search mode is the ability to perform query expansion using terms from an environmental ontology. PortalU® makes use of an external service, the Semantic Network Services (SNS), an environmental knowledge-base operated by the German Federal Environment Agency (Umweltbundesamt) (Bandholtz 2003). The SNS incorporates an environmental thesaurus, a geothesaurus, and an event-database called the "environmental chronicles". All three ontologies are implemented as topic maps. In PortalU®, they can be used to retrieve thematic keywords and similar terms, to disambiguate place names, and to specify time periods.

To make these valuable expert-tools available for non-expert users that do not venture beyond the simple-search interface, parts it were directly integrated in the simple-search interface: For each term in a simple-search query, a set of similar terms is derived on-the-fly by SNS. The user may select one or multiple terms from this list to expand the query string.

## 2.2 Focussed Access to Environmental Topics

In addition to the search functions described above, PortalU® opens a second path to environmental information through a number of catalogue functions. They feature lists of selected information items like web-sites with up-to-date environmental monitoring data, events and new publications, news bulletins, and other particularly relevant information. The user may browse through these lists and apply a number of filter-functions to find the most interesting information items. Catalog-entries of type monitoring data, events/publications, and news-bulletins can be filtered according to their main thematic and regional (i.e., geographic) coverage<sup>1</sup>. Information items that are listed as “environmental topic pages” support a third filter that allows to select specific information categories, like “Maps and Data” or “Legal Documents” etc.

The environmental topic pages are the most comprehensive list in PortalU®, although they represent only a sub-set of the PortalU® information space. Members of this set are selected manually, i.e. the PortalU® content managers may nominate selected web-pages and/or metadata records for the thematic catalogue. The 21 environmental themes and 6 functional categories are modelled along the lines of the thematic and functional categories specified by EU-Directive 2003/4/EC. Thus on the one hand, the environmental topic pages are a tool to focussed catalogue of content-rich and valuable information items in direct response to the requirements of the EU Directive. On the other hand, the manual maintenance of the catalogue requires considerable efforts. In the long term, tools for auto-classification may offer a practical solution for a less expensive maintenance of the PortalU® thematic catalogue.

## 2.3 Regionalized Access to Environmental Information

Experience with gein and a number of papers and studies (e.g. Zinnbauer 2005) indicate, that many users of environmental information systems are mostly interested in regional or local information. Very often, questions related to “How is the environment in my city / my neighborhood?” are the most relevant ones for the average user. Regionalized access to information and spatial search options therefore play an important role in PortalU®.

With respect to its main information space, i.e. the index of web-pages, metadata and database records, PortalU® supports basically two types of spatial-search functions: a map-based spatial search function and a tool for spatial searches through place-names. Both tools can be used to define a search region in terms of bounding box coordinates (for web-pages and database records) and spatial identifiers (for metadata). Place names are translated to bounding-boxes and/or identifiers with the help of the geothsaurus provided by SNS.

The geographic reference of metadata records and database-entries is provided directly by the data sources. In the UDK, for example, each metadata record is assigned a geo-reference through the ISO-compatible field *geographic-extent*. Catalog pages are regionalized according to the responsible information provider. The underlying assumption here is that each information provider holds information only about a specific geographical area. Because public authorities

---

<sup>1</sup> In PortalU, regional coverage of an information item is closely related to its provenance. This is feasible because most information providers in PortalU are government organizations with a clearly defined region of operation.

typically have a well-defined geographical area of operation, this assumption is valid in most cases. The geo-reference of unstructured information items, such as web-pages and documents, is assigned through SNS on the basis of a “semantic” text analysis. SNS parses the content, extracts place names and tries to match them to the terms of its geothsaurus. This method works reasonably well, but further development of the algorithms applied is necessary to improve the results.

However, the biggest challenge in terms of providing regionalized information is not a technical, but an organizational one: The best and sometimes only information sources for environmental information and thematic geospatial data are local, i.e. on the municipal and county level. Organizations on this level are therefore explicitly addressed by the EU-Directive on environmental information. They are likely to play an important role in the context of INSPIRE as well. Although the need to include them in PortalU® is clearly seen, this has not been possible for organizational reasons.

However, the technical basis for their inclusion has already been established: The PortalU® software is flexible and scaleable enough to accommodate a large number of information providers and heterogeneous information sources. If necessary, new information nodes may be established on demand, for example to build environmental information portals for a cluster of municipalities or even an individual community. All information portals based on the InGrid software can be interlinked into one homogeneous information network. In the following sections, we will outline the basic principles of the flexible and scaleable PortalU® software.

### **3. Technical Implementation of PortalU®**

#### *3.1 User Interface*

PortalU® was equipped with a web-interface designed to provide an intuitive and simple access multiple heterogeneous information sources. The goal was to hide the underlying complexity from the user and to present information in a structured, simple and understandable way. It has to be pointed out, however, that PortalU® does not provide a harmonized view on data or information. The portal operates as a mere broker for information items. A semantic harmonization and integration of information and data is beyond the scope of PortalU®, at least at the current stage of the project.

Nevertheless, the PortalU® web-crawler can access and index a large number of different information sources. This includes web-pages, (meta)data catalogues, online information systems, and databases. Under the condition that these information sources fulfill specific technical requirements<sup>2</sup>, all retrieved information items can be integrated into one single, globally ranked result set, notwithstanding their technical and semantic heterogeneity. The user may thus select the most relevant information item from the result-set without having to care about it’s format or origin (although, format and origin of each information item are displayed for information purposes).

---

<sup>2</sup>To participate in the global ranking, data sources have to be connected to PortalU via a specific (Data-Source Client) interface.

Most objects are displayed in their native format, but specialized visualization tools are available for certain types of information items: Record sets from a data catalog are displayed using a report-like format. OGC-compatible Web Mapping Services (WMS) can be viewed, overlaid, and queried using a built-in mapping client.

### 3.2 InGrid Architecture

The software behind PortalU® was named “InGrid®”. This name, which stands for “Information-Grid”, reflects on the grid-like architecture of the system. To build InGrid®, standard OpenSource solutions adapted to the needs of PortalU® were used where possible. For example, to store configuration information and local metadata, InGrid® relies on MySQL as RDBMS; the InGrid® search engine is based on Nutch and Lucene (Apache 2007), and the built-in GIS utility applies the UMN Mapserver (UMN 2007).

A schematic view of the InGrid® 1.0 architecture is given in Figure 2. The core component of this architecture is an information-broker dubbed “iBus” that functions as a link between user- and data interfaces and multiple distributed data sources.

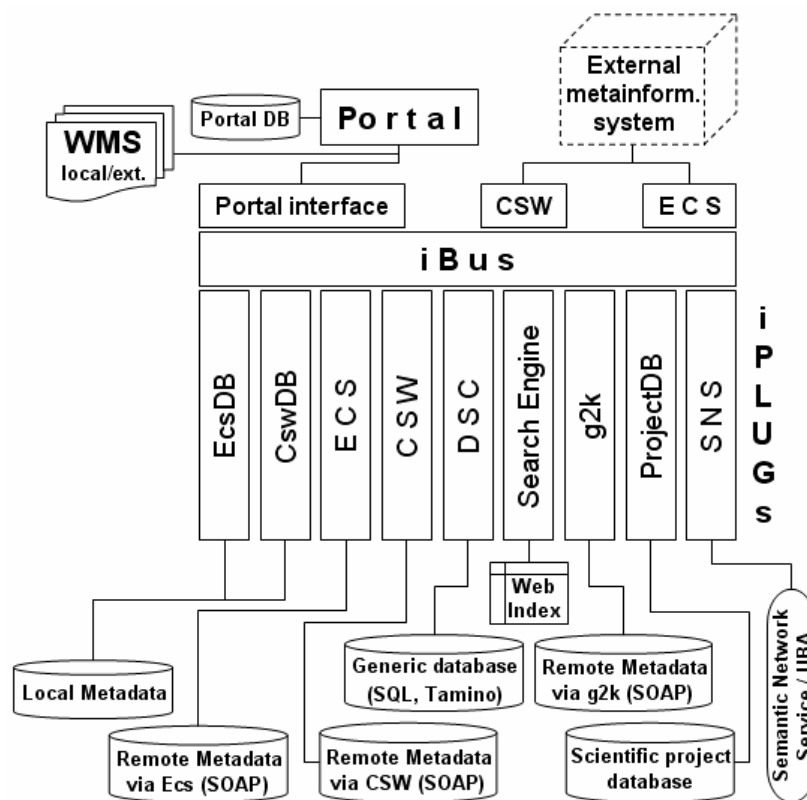


Figure 2: Schematic view of InGrid 1.0 architecture

### 3.2.1 User- and Data Interfaces

An information-request can reach the broker component over one of three interfaces: *(i)* the portal-interface, where users may specify search queries, *(ii)* the CSW catalog-interface, which allows other information systems and data catalogs to submit queries to PortalU®, and *(iii)* the Environmental Catalog Service interface (ECS-interface), which basically is another SOAP interface covering the complete InGrid metadata model.

The portal user interface was implemented using the Apache Jetspeed framework featuring Velocity as template engine. This setup offers a lot of built-in and very useful functionality, in particular in the area of interface administration and customization. As a consequence, it is easy to personalize the PortalU® user interface and to change content and query characteristics.

Attached to the PortalU® user-interface is a built-in Web Map Server (WMS) and an associated visualization client. The WMS module builds on the popular UMN Mapserver (UMN 2007) and conforms to the OGC WMS1.1.1 specification (OGC 2007). The stand-alone client was developed on the basis of the Mapbender (Open Source Geospatial Foundation, 2007) and Maplab (Maptools.org, 2007). Within PortalU®, the map-client is used for geographic query extension (which is an important part of the PortalU® detailed search functionality), and to visualize geospatial data that are retrieved through a search query. The map-client can also be used to visualize remote WMS by specification of the respective get-capabilities URL.

The other two interfaces (the CSW and the ECS interface) are new developments. The ECS interface evolved out of the so-called “semantic XML interface” that was used by the predecessor of the PortalU® data catalog, the Environmental Data Catalog UDK (Karschnick et al. 2003). The CSW interface applies an application profile built upon OGC’s Catalog Service Web 2.0 (CSW 2.0) specification (OGC, 2005). In particular setting up the CSW interface proved to be problematic because the available specifications were, at the time, not finalized state. As a result of changing and inadequate specifications, the CSW interface had to be modified several times. Only now, almost one year after PortalU® went online, the interface is in a state where it can communicate with (at least some) other CSW data catalogs. Our experience with the lack of general interoperability between different CSW interfaces coincides well with an (unpublished) study conducted by the EU’s Joint Research Center (JRC). It showed for several data catalogs in Europe that communication and metadata exchange is still not possible without the help of customized catalog interfaces. It can only be hoped that the maturing of OGC specifications and ISO standards, assisted by the soon-to-be-released implementing rules for INSPIRE, will be able to improve the situation.

### 3.2.2 Distributed Datasources

The information broker, or iBus, applies a number of customized data interfaces to connect to different data sources. In InGrid®, these adapters are called “Data Source Clients” (DSC’s) and their technical implementations “iPlug’s”. The iPlug has two functions: it acts as a translator between the iBus and a data source, and it provides the functionality needed for indexing and ranking query results. The latter is needed by the iBus to integrate query results from different data sources and to build a globally-ranked result list.

Each iPlug builds on a generic DSC-module and has to be configured to match a specific data source. The data-source interface of the generic DSC currently supports databases of type MySQL, Oracle and Microsoft SQL Server. Moreover, a special Tamino DSC can be used to connect to Tamino-based XML data stores. In PortalU®, this is needed mainly to connect to the “Umweltobjektkatalog” (UOK) operated by the Bavarian Ministry for Environment.

A specialized DSC is the Search-Engine iPlug which controls the crawling and indexing process over the defined web space (start-URLs). It generates a flat, file-based high performance index of the specific page content. Parsing of web pages and common document formats, e.g. pdf and doc, is supported. The search engine iPlug has been built on the Apache Nutch framework featuring the Apache Lucene indexer (Apache, 2007).

Communication between iBus and iPlugs is established using the Java JuXTApose (JXTA) P2P platform (JXTA 2007). JXTA is able to specify a message syntax without being tied to a specific transmission protocol. In distributed systems without a unique definition of the network characteristics, intelligent and flexible brokering is of prime importance for the overall performance of the system. Firewall-friendly “protocols” like SOAP can be a performance killer. On the other hand, more efficient protocols can be blocked by firewalls. With JXTA, communication channels can be established over the most efficient protocol available. It can use TCP/IP sockets, but also, if two information nodes are separated by a firewall, http and the standard port. Moreover, JXTA enables intelligent routing between two subsystems, if necessary even over another intermediate system which acts as protocol translator.

Using JXTA it is possible to physically distribute iBus and iPlug’s on different machines in different organizations. An iPlug connecting a remote data source can be installed locally with the iBus, or remotely on top of the data source itself. Nevertheless, there are some good reasons to physically store the index close to the respective data source, so that the data provider has full control over the index generation process (what, when, etc.) and the indexing does not have to deal with too many network bottlenecks.

#### **4. Summary and Conclusions**

In a federal state like Germany, the responsibility for publicly-held data and information is distributed among many different public authorities on all levels of the administrative hierarchy. This can be the cause for a considerable lack of “organizational interoperability”, which hampers the access to and exchange of information. PortalU is operated by the environmental administration to improve organizational interoperability in the environmental sector, and to cope with new legislation like the EU-Aarhus Directive and INSPIRE.

As the central information portal of the environmental administration in Germany, PortalU® provides access to data and information from federal- and state agencies. Currently, the portal acts as an information broker that harmonizes the access to data and information. Harmonization and semantic integration on the data level has not been implemented yet.

## 5. References

- Apache Lucene Project. 2007. <http://lucene.apache.org/>. Last accessed Feb. 28, 2007.
- Apache Lucene Project. 2007. <http://lucene.apache.org/nutch/>. Last accessed Feb. 28, 2007.
- Apache Portals. 2007. <http://portals.apache.org/jetspeed-2/> . Last accessed Feb. 28, 2007.
- Apache Velocity Project. 2007. <http://velocity.apache.org/>. Last accessed Feb. 28, 2007.
- Bandholtz, T. 2003. Erstellung eines semantischen Netzwerkservice (SNS) für das Umweltinformationsnetz Deutschland (gein®). Abschlussbericht, Umweltbundesamt.
- EU 2003 Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on Public Access to Environmental Information. In: Official Journal of the European Union, L 41/26, 14.2.2003.
- JXTA. 2007. <http://www.jxta.org/>. Last accessed Feb. 28, 2007.
- Karschnick, O., Kruse, F., Töpker, S., Riegel, T., Eichler, M., and Behrens, S. 2003. The UDK and ISO19115 Standard. In: Proceedings of the EnviroInfo 2003, Cottbus.
- Maptools.org. 2007. <http://www.maptools.org/maplab/>. Last accessed Feb. 28, 2007.
- OGC Open Geospatial Consortium 2005: OpenGIS® Catalogue Services Specification 2.0 - ISO19115/ISO19119 Application Profile for CSW 2.0, Version 0.9.3. OGC document 04-038r2.
- OGC Open Geospatial Consortium. 2007. <http://www.opengeospatial.org/standards/wms>. Last accessed Feb. 28, 2007
- Open Source Geospatial Foundation. 2007. [http://www.mapbender.org/index.php/Main\\_Page](http://www.mapbender.org/index.php/Main_Page) Last accessed Feb. 28, 2007.
- Swoboda, W., Kruse, F., Nikolai, R., Kazakos, W., Nyhuis, D., and Rousselle H. 1999. The UDK Approach: the 4th Generation of an Environmental Data Catalogue Introduced in Austria and Germany. In: Proceedings of the 3rd IEEE Meta-Data Conference, Bethesda, Maryland.
- UMN University of Minnesota. 2007. <http://ms.gis.umn.edu/>. Last accessed Feb. 28, 2007.
- Vögele, T., Kruse, F., and Karschnick, O. 2004. The gein® 2.0 Information Broker for Environmental and Geospatial Data. In: Proceedings of the 10th EC GI & GIS workshop, Warsaw.