

InGrid 1.0 – The Nuts and Bolts of PortalU

M. Klenke¹, F. Kruse¹, H. Lehmann¹, T. Riegel¹ and T. Vögele¹

Abstract

After the official launch in May 2006, PortalU will act as the new internet portal of the German environmental authorities. The “U” in PortalU stands for “Umwelt”, the german word for the environment. PortalU will provide access to government-owned environmental information of the 16 federal German states and the German federal government.

The application replaces the actual “German Environmental Information Network” gein[®] (Vögele et al. 2004). Moreover, the “Virtual-UDK” (Swoboda et al. 1999), currently acting as the top-level environmental metadata query component of the German environmental authorities, will be merged with PortalU. The well-established UDK metadata model, however, will retain its importance as de-facto standard within the environmental community in Germany. Hence, the ISO19115 and ISO19119 compliant model builds the basis of the metadata component of PortalU.

The step from gein[®] and the Virtual-UDK to PortalU is accompanied by a complete technical redesign which will enhance the functionality of the PortalU predecessors in many aspects. This new set-up is needed to include, as required by the new EU-directive 2003/4/EC on public access to environmental information (EU 2003), a much larger number of providers of environmental information, including agencies on the local level. It is also needed to be able to provide better access to environmental information that is not yet online or belongs to the so-called “hidden web”.

Our paper describes the concepts and architecture of the software behind PortalU: InGrid 1.0. The name InGrid relates to “Information Grid”, pinpointing to the distributed approach of the new system. To support direct access to environmental data and metadata, the new system architecture features distributed indices based on highly adaptable database interfaces and a self-administrating peer-to-peer communication infrastructure. A number of standardized interfaces, including a web catalog service conforming to ISO and OGC standards and recommendations enable PortalU and other InGrid information nodes in the network to benefit from and contribute to other data infrastructures, for example the German national geodata infrastructure GDI-DE and the EC INSPIRE initiative (RDM Working Group 2002).

1. Introduction

The challenges that had to be met with the new system were defined by the requirements of the data providers and the user community. From the provider’s point of view, up-to-dateness and singularity are of prime importance. That is, data should not be duplicated in a central data warehouse to avoid problems arising from redundant data storage, for example (meta)data copyright, timeliness, necessary “harvest” cycles, mirror traffic, bandwidth, required disk space, and increased administration effort. Therefore, InGrid strictly follows an approach of querying distributed data providers directly.

Another important issue for data providers is how to get information online that is not yet accessible through the internet. For example the huge amounts of data, which are “buried” in data bases, but none the

¹ Coordination Center PortalU, Ministry for the Environment of Lower Saxony, Archivstr. 2, D-30169 Hannover, Tel.: +49 (0)511 1203407, E-Mail: kst@portalu.de

less relevant with respect to the EU-directive 2003/4/EC. A tool is needed to easily make available selective parts of data stored in RDBMSs and other sources. By means of such a tool, data providers can save the costs of implementing a unique interface and web application for every single data base. The InGrid-solution, that is the “Data Source Client”, is described later in this paper.

Two main requirements for the PortalU web search have been assumed on the user side. First of all, fast response times are crucial for user acceptance in times of Google and other efficient search engines. The only practical approach to meet this demand is to try to get as much data indexed as possible, so that the index can be queried to construct the first level result set. Naturally, the answer of an index, or of multiple distributed indices, is fast and can keep the users first attention while adding results from slower, non-indexed data sources bit by bit. E.g. responses from systems communicating via a Web Catalog Service.

The second important user requirement is an effective ranking mechanism. PortalU will integrate huge amounts of data from many sources including web pages, metadata repositories and generic data bases. Depending on the search string, the result sets will get too large to be displayed on a single result page, or even a reviewable set of result pages. Therefore, content indexed by InGrid will be supplied with a weighting factor to enable a classified, integrated result set. The weighting of content will be estimated by means of the Tf-idf algorithm.

2. InGrid Architecture

The architecture of InGrid 1.0 is shown in figure 1. Core component is a broker accepting requests from its own overlying user interface, that is the portal component, as well as from external systems: the so-called “iBus”. A request can reach the iBus over one of three interfaces: (i) the portal interface, for example a PortalU search query, (ii) the CSW interface which accepts requests conforming to the DE-profile build upon the OGC CSW 2.0 application profile (OGC 2005) and (iii) the ECS (Environmental Catalog Service) interface which basically is another SOAP-based interface covering the complete InGrid metadata model. The ECS evolved out of the semantic XML interface of the actual Windows-UDK 5.0 systems (Karschnick et al. 2003).

The portal as user interface is realized with the Apache Jetspeed² framework featuring Velocity³ as template engine. It allows a high degree of personalization of interface, content and query characteristics.

A Web Map Server conforming to the OGC WMS 1.1.1 specification and a stand-alone client supporting local and remote WMS data integration are part of the distribution. Within the portal, the client is used for geographic query extension being an important part of the PortalU detailed search. Software components integrated include the UMN Mapserver⁴, Mapbender⁵ and Maplab⁶.

On the bottom side, to connect the iBus to a data source, a generic plug has to be installed to negotiate between iBus and respective source. In the Ingrid context, we call these plugs “Data Source Clients” (DSC’s) and their technical implementations “iPlug’s”. Besides its role as a translator between iBus and data source, the DSC supplies the functions for indexing and ranking query results. The final integration of the ranked results from different sources then happens in the iBus.

Currently, the DSC has JDBC support for MySQL, Oracle and Microsoft’s SQL Server. Moreover, a Tamino DSC can be used to connect Tamino-based XML data stores, for example the “Umweltobjektkatalog” (UOK) from Bavaria. The PortalU installation of InGrid 1.0 will use MySQL as RDBMS to store local metadata and configuration information.

² <http://portals.apache.org/jetspeed-2/>

³ <http://jakarta.apache.org/velocity/>

⁴ <http://www.umn-mapserver.de/>

⁵ <http://www.mapbender.org/>

⁶ <http://www.maptools.org/maplab/>

A special kind of DSC is the Search Engine iPlug which controls the crawling process over the defined web space (start-URLs) and 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, will be supported. The search engine iPlug has been built on the Apache Nutch⁷ framework featuring the Apache Lucene indexer⁸.

To support semantic and geographic query extension and automatic, thesaurus-based key word generation of web content, the Semantic Network Service (SNS) of the German Umweltbundesamt (Bandholtz 2003) is being connected by means of another iPlug. Moreover, the SNS supports the PortalU user by offering similar terms for refining an actual search query.

Communication between iBus and iPlugs is canalized by the Java JuXTApose⁹ (JXTA) P2P platform.

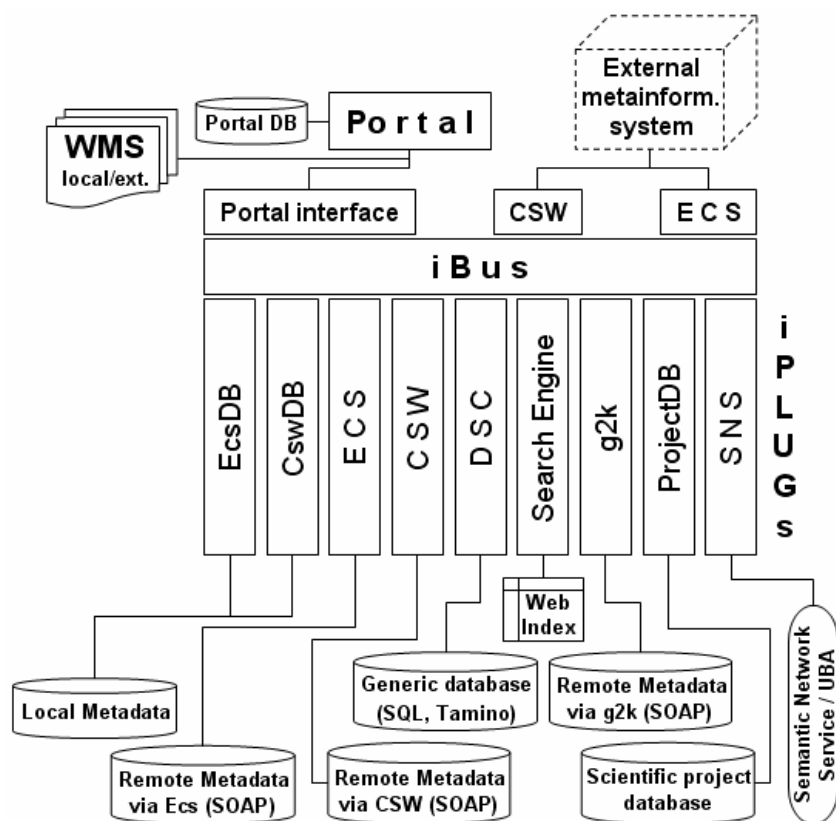


Figure 1: InGrid 1.0 architecture

JXTA specifies message syntax without being bonded to a specific transmission protocol. In distributed systems without a unique definition of the network characteristics, intelligent and flexible broking is of prime importance for the overall performance of the system. Firewall-friendly “protocols” like SOAP can be a performance killer. However, more efficient protocols can be blocked by firewalls. Hence, JXTA communication can be established over an efficient protocol, for example TCP/IP sockets, but also, if two information nodes are separated by a firewall, http over port 80 can be used to exchange information. Moreover, JXTA enables intelligent routing between two subsystems, if

necessary even over another intermediate system which acts as protocol translator. By means of JXTA it is possible to physically distribute iBus and iPlug’s. An iPlug connecting a remote data source can be installed locally with the iBus, or remotely on top of the data source itself. However, there are some good reasons to physically store the index close to the respective data source, so that the data provider has full

⁷ <http://lucene.apache.org/nutch/>

⁸ <http://lucene.apache.org/>

⁹ <http://www.jxta.org/>

control over the index generation process (what, when, etc.) and the indexing does not have to deal with too many network bottlenecks.

3. Use cases and outlook

PortalU will be the central InGrid 1.0 installation substituting gein[®] as well as the Virtual-UDK and acting as new top-level broker for government-owned environmental information in Germany. All InGrid software components are being built on the basis of “free” software. Hence, the application can be used without any costs by institutions and organizations under the roof of the administrative UDK/gein[®] cooperation (VwV 2003), signed by the German federal government and the German federal states.

The highly modularized application is suited to be installed in many application scenarios ranging from a full-featured environmental or geo-information internet portal, down to a fast and slim ISO-compliant metadata retrieval system. A DSC can be connected to any number of iBus's to share its content. Vice versa, an iBus can query an unlimited number of DSC's. By this architecture, multiple physical data source cascades with their manifold negative implications on performance and content are avoided. Part of the PortalU technical infrastructure will be a “rendevous-server” based on JXTA P2P technology as central repository of all existing iBus's and iPlug's in the network. Registration of iBus's and iPlug's at the rendevous-server is fully automated during the set-up process.

Currently, the Windows-UDK 5.0 (Karschnick et al. 2003) is used for metadata registration and maintenance. With InGrid 1.1, which is planned to be launched in spring 2007, the Windows-UDK will be replaced by another InGrid module allowing the registration and maintenance of metadata over a fully integrated web application with a sophisticated role- and quality insurance system. With InGrid 1.1, the full integration of gein[®], WWW-UDK and Windows-UDK will be realized, finally fulfilling the strategic goal formulated by the UDK/gein[®] cooperation in 2003.

4. References

- T. Bandholtz (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.
- O. Karschnick, F. Kruse, S. Töpker, T. Riegel, M. Eichler, and S. Behrens (2003): The UDK and ISO19115 standard. In: Proceedings of the EnviroInfo 2003, Cottbus.
- 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.
- RDM working group (2002): INSPIRE Infrastructure for Spatial Information in Europe. INSPIRE RDM PP v4-3 en, EUROSTAT.
- W. Swoboda, F. Kruse, R. Nikolai, W. Kazakos, D. Nyhuis, and H. Rousselle (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.
- T. Vögele, F. Kruse, and O. Karschnick (2004a): The gein[®] 2.0 Information Broker for Environmental and Geospatial Data. In: Proceedings of the 10th EC GI & GIS workshop, Warsaw.
- VwV (2003): Verwaltungsvereinbarung zwischen Bund und Ländern über die gemeinsame Entwicklung und Pflege des Metainformationssystems Umwelt-Datenkatalog UDK und des Umweltinformationsnetzes Deutschland GEIN. <http://www.udk-gein.de/allgem/vwv.shtml>.