

A New and Flexible Architecture for the German Environmental Information Network

Thomas Vögele¹, Martin Klenke¹, Fred Kruse¹ and Stefan Groschupf²

Abstract

Since the year 2000, the German Environmental Information Network *gein*[®] provides access to government-owned environmental information of 15 German federal states and the German federal government. In response to the new requirements for public access to environmental information based on EU-directive 2003/4/EC, *gein*[®] currently undergoes a complete technical re-design. The resulting new software (InGrid 1.0) will be used to build *Portal-U*, an improved central information portal which replaces *gein*[®]. In addition, InGrid offers state- and municipal authorities a tool to set up their own information portals and to construct a flexible network of interconnected information nodes. This new technical infrastructure is needed to include, as required by the EU-directive, 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”. 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 P2P communication infrastructure. A number of standardized interfaces, including a web catalog service conforming to OGC standards, enable *Portal-U* and other information nodes in the network to benefit from and contribute to other data infrastructures, like the German national geodata infrastructure (GDI-DE) and INSPIRE.

1. Introduction

The German Environmental Information Network *gein*[®] (Bilo/Streuff 2000) is an online information system specialized on the access to distributed environmental information and data. As of spring 2005, the network connects more than 90 providers of environmental information in Germany, providing access to more than 400,000 webpages and online documents as well as about 500,000 data objects in online data catalogs and databases. The development and maintenance of *gein*[®] is funded through an administrative agreement between the German federal administration and the administrations of 15 German federal states (VwV 2003).

One of the main objectives of *gein*[®] is to provide information seekers with an easy access to high-quality environmental information. To achieve this goal, *gein*[®] applies a number of strategies: Firstly, participation in *gein*[®] is restricted to government agencies and associated organizations. This ensures that only “official” information can be found through *gein*[®], i.e. information that typically has undergone a thorough evaluation before being published. Secondly, *gein*[®] opens a number of parallel, but different paths to access environmental information. This includes search tools and semantic query expansion, direct access to important sites with time-critical news and monitoring data, as well as a structured access to particularly interesting sites through a navigation based on an environmental topic classification. Last but not least, *gein*[®] provides access not only to web-pages and online documents, but also to parts of the “hid-

¹ Coordination Center UDK/GEIN, Ministry for the Environment of Lower Saxony, Archivstr. 2, D-30169 Hannover, Tel.: +49-511-48313447, E-Mail: kug@numis.niedersachsen.de

² media style GmbH, Mansfelder Straße 13, D-06108 Halle, Tel.: +49-345-5222702, E-Mail: info@media-style.com

den web”, i.e. to information and data stored in online databases and data catalogs. For this purpose, the system implements a standardized interface on the basis of XML/SOAP (Kruse et al. 2003).

Considerable efforts are necessary to maintain a network with such a high quality and complex data connections. For this reason, *gein*[®] is managed and administered by the *Coordination Center UDK/GEIN (KUG)* in cooperation with representatives of the participating federal states and the federal government. Each state-representative is the *gein*[®] contact officer for all information providers in the respective state. They control who is allowed to participate in the network, and with what content. The KUG’s responsibilities are to overview the technical functioning of the system and the correct representation of the associated content, as well as to guide the continuing improvement and technical development of the software.

gein[®] is supported by the *Umweltdatenkatalog (UDK)*, a catalog of environmental data sources (Swoboda et al. 1999, 2000). The UDK is a metainformation system that is used since the early 1990’s by most federal states and a number of federal environmental agencies to catalog and manage environmental data. The UDK metadata model which is based on international standards like DC, ISO 19115 and ISO 19119, has evolved to a de-facto standard within the environmental community in Germany. All UDK data catalogs can be searched through a central online-portal (www.umweltdatenkatalog.de). This portal is connected to *gein*[®], so that the UDK data catalogs are searchable through *gein*[®] as well.

2. New requirements – The EU Directive on Public Access to Environmental Information

The organizational structure of *gein*[®] works well in the current situation, i.e. with a relatively small number of information providers. To adjust to the new requirements defined by EU Directive 2003/4/EC on “public access to environmental information” (EU 2003), this structure will have to be changed. The directive considerably extends the right of European citizens to free access to environmental data and information and re-defines the “terms of trade” for environmental information in both an organizational and a technical sense:

- The directive extends the circle of organizations that are required to publish environmental information. Not only environment agencies on the higher administrative levels, but all administrative organizations on all levels of the hierarchy have to make environmentally relevant information available. In addition, all private companies and semi-private organizations that fulfill environmentally relevant public obligations are included as well. For *gein*[®] this means that the number of participating information providers may increase dramatically. For example, there are more than 14000 municipalities in Germany that are information providers according to the definition used by the directive.
- The directive explicitly calls for active distribution of data and information. Thus, not only references to data and information (i.e. metadata), but in many cases the data and information objects themselves should be online. For *gein*[®] this means that more and better tools to access and visualize databases, documents and digital maps have to be developed.

In order to be able to meet these requirements in the future, *gein*[®] currently undergoes a major re-design and re-implementation. The new software, InGrid³ 1.0, integrates not only the functionality of *gein*[®], but also the main features of the UDK. Using InGrid, a new central environmental information portal, *Portal-U*⁴, will be build. The new portal is expected to go online in the first half of 2006. *Portal-U* will represent the top node of a new environmental information network that can incorporate a large number of data

³ Derived from “Information Grid”.

⁴ „U“ is relating to „Umwelt“, the german word for environment.

sources and UDK databases as well as new customized information nodes. In the following, we will provide a brief overview of the software architecture needed to support this new network structure.

3. System architecture

At the current stage, the organizational layout of the information network which will be implemented to meet the challenges described above is not yet finalized. A number of different organizational scenarios are possible. Which one will be chosen depends on political and organizational decisions that cannot be described within the scope of this paper. However, it is obvious that the basis for any choice of network organization has to be a flexible and scalable technical infrastructure. For this reason, the system architecture implemented in InGrid 1.0 follows a highly modular approach which does not adhere to a pre-set network topology, but makes it possible to build the network in a flexible way as needed. The main building blocks of the network are information nodes and data sources.

3.1 A flexible network of information nodes

Each information node in the *Portal-U* network represents a customized set of InGrid 1.0 software modules and functions. The top-node in the *Portal-U* network is formed by *Portal-U* itself, i.e. the central information portal managed by the Coordination Center UDK/GEIN (KUG) in Hannover (Info Node A in Figure 1). Other information nodes may be added as needed.

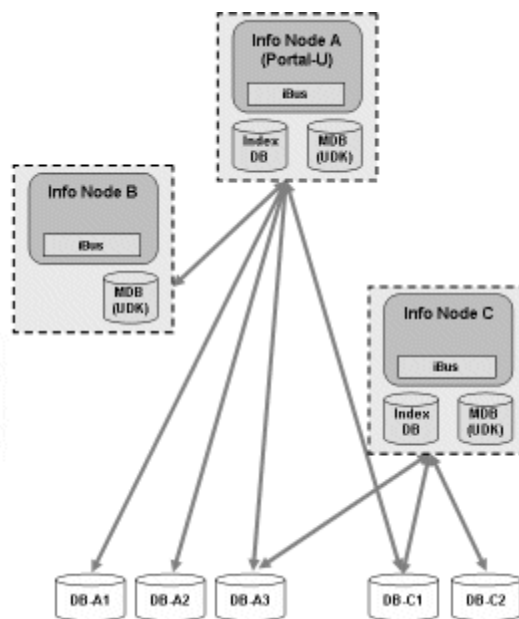


Figure 1
A flexible network of information nodes

These information nodes can be located on any level of the administrative hierarchy. Which InGrid modules a specific information node uses depends on its purpose and the functionality needed. For example, one information node may want to use all InGrid components, including portal functions, search tools, visualization services, access to databases, web-pages and metadata. In this case, the installation would of-

fer the same capabilities as the central Portal-U node and could for example be used by a state to build a state-specific environmental portal, or by a community to build an environmental information portal on the community level (Info Node C in Figure 1). Another information node may be configured to provide only the functionality of the UDK, i.e. metadata storage and a web-client with search tools. Such a node can be used in place of a traditional WWW-UDK installation (Info Node B in Figure 1). It is possible to customize an InGrid portal using portal-specific design patterns (colors, logos, etc.).

3.2 Peer-to-Peer Communication and Direct Access to Data Sources

Data sources in the *Portal-U* network remain distributed. There are no centralized data storage devices except for a metadata component and an index database as integrated part of each InGrid information node. The metadata component implements the UDK metadata model. It can therefore be seen as a UDK database integrated in an InGrid portal. The index database holds a fulltext index of all web-sites that are directly connected to the portal. To create this index, the Open Source components *Lucene* and *Nutch*⁵ are used.

Other data sources are connected to an information node through interface-adapters called “iPlugs”. The information node implements a standard interface (the so-called “iBus”) in which iPlugs can be “plugged-in” as needed. There are different types of iPlugs, each specialized for a specific type of data source. There are iPlugs that implement legacy interfaces inherited from the *gein*[®] and UDK software (e.g., the *g2k*-iPlug and the UDK-iPlug). Others, like the Catalog Service Web-iPlug, provide new interfaces that are not supported by the current system.

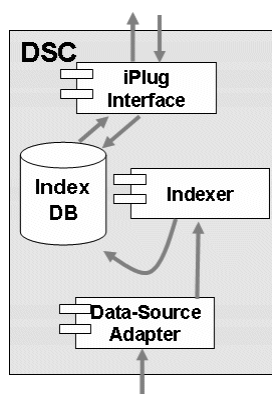


Figure 2
Datasource client (schematic)

Technically an iPlug is a small (< 10 MB) software device that is preferably installed directly at the site of the data source. Installation and configuration of an iPlug is simple and supported by a user-friendly GUI tool. When an iPlug is installed, it connects automatically to the network. To be able to do so, all iPlugs are embedded in a JXTA⁶-based peer to peer communication infrastructure. This infrastructure en-

⁵ <http://lucene.apache.org/>, <http://lucene.apache.org/nutch/>

⁶ <http://www.jxta.org/>

sures that information about the iPlug's connection parameters are available at each information node in the network.

All data sources registered in the network through an iPlug can be accessed by each information node, i.e. from each iBus in the network, a direct connection to each known iPlug in the system can be established. The central *Portal-U* information node, for example, may have direct access to its own data sources (i.e., DB-A1, DB-A2, and DB-A3 in Figure 1) as well as to data sources that are primarily registered with another information node (i.e., DB-C1 and DB-C2 in Figure 1). A horizontal architecture of distributed data sources and direct connections was chosen mainly because it has a significant performance advantage over a hierarchical architecture of cascading information nodes.

3.3 Adaptable Datasource-Clients

The most important new interface-adapter, or iPlug, is the so-called "Datasource Client (DSC)". It is used to make the content of databases searchable by *Portal-U*. Online databases, as well as databases that do not have a web-interface, can be connected to an InGrid portal through the DSC-iPlug. The core components of a DSC are a database-specific data-source-interface, an indexer, and an index database (Figure 2). The DSC data-source-interface will support JDBC for the connection to standard relational databases and the Tamino API to connect to XML-based data pools.

Through the data-source-interface, the schema of a data source is mapped to a generic InGrid schema. Once connected to the data source, the DSC's built-in indexer scans the data source and creates a full-text index of its content. The resulting index database is part of the DSC and stored at the same physical location as the DSC. An information request routed to the DSC through the iBus interface is matched directly against the DSC index database.

Overall, the Datasource Client offers three major benefits:

1. The DSC is able to reply to information requests from the InGrid portal very effectively and without the need to access a data source in real-time.
2. Because all DSC index databases follow the same logic and general setup, the results of a query answered by multiple DSCs (i.e., by multiple distributed data sources) can be integrated into a homogeneous and uniformly ranked result set.
3. As a small (< 10MB) software tool implemented in platform-independent JAVA code, the DSC can be installed on almost any computing platform. All included third-party modules (e.g. the Lucene indexer) are Open Source products and the installation, configuration, and long-term administration of the tool is supported through a comfortable user-interface. Consequently, the DSC offers the means to link databases to an information portal with much less effort and improved functionality, for instance ranking capabilities, compared to the interfaces that are used in the current *gein*[®] system.

4. Portal-U metadata as part of the German national geodata infrastructure

The InGrid 1.0 software features a catalog interface that is compatible with the OGC CS-W 2.0 specification. Through this interface, each InGrid information node can function as a geodata catalog.

Portal-U metadata of the types "Geographic Information – Metadata" (ISO 19115) and "Geographic Information – Services" (ISO 19119) can be queried through this interface. The German national geodata infrastructure (GDI-DE) will integrate the respective German environmental metadata residing under the roof of Portal-U via the CS-W 2.0 compliant interface. In general, the Portal-U CS-W interface ensures an easy integration of German environmental metadata into other data infrastructures, for example relating to the European INSPIRE initiative.

5. Summary

Two of the most important online information systems within the German environmental administration, the Umweltdatenkatalog (UDK) and the German Environmental Information Network (*gein*[®]), are currently in a process of re-design. One reason for this process is that new European environmental regulations call for a better and more comprehensive access to government-held environmental information.

The software presented in this paper (InGrid 1.0) will offer information providers from all levels of the administrative hierarchy a tool to contribute to meeting these requirements. It will integrate the actual UDK and *gein*[®] systems, improve and extend their capabilities, and constitute a basis for building a highly flexible and scalable information network.

Portal-U, the central environmental portal of the German federal and state governments, will be the top node of this new network. Additional information nodes may be added as needed.

6. References

- Bilo, M. and H. Streuff (2000): Das Umweltinformationsnetz Deutschland - GEIN2000 - Fachliche Anforderungen an ein Forschungs- und Entwicklungsvorhaben. In: Proceedings of the 3rd Workshop "Hypermedia im Umweltschutz", Ulm.
- 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, http://europa.eu.int/eur-lex/pri/en/oj/dat/2003/l_041/l_04120030214en00260032.pdf.
- Kruse, F. et al. (2003): *gein*[®] - Planning the next generation. In: Proceedings of the 17th International Conference "Informatics for Environmental Protection", Cottbus.
- 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>.
- Swoboda, W. et al. (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.
- Swoboda, W. et al. (2000): Harmonisierter Zugang zu Umweltinformationen für Öffentlichkeit, Politik und Planung: Der Umweltdatenkatalog UDK im Einsatz. In: K. Tochtermann, W.-F. Riekert (Eds), Proceedings of the 14th International Symposium „Computer Science for Environmental Protection“, Bonn.