CRIS2002
Current Research Information Systems

Wolfgang Adamczak / Annemarie Nase (eds.)

Gaining Insight from Research Information

6th International Conference on Current Research Information Systems

promoted by euroCRIS
Current Research Information Systems

kassel
university
press

Contents

Lectures

Workshops

Poster

DBClear: A Generic System for Clearinghouses

H. Hellweg[1], B. Hermes[1], M. Stempfhuber[1], W. Enderle[2], T. Fischer[2]
[1] Social Science Information Centre (IZ), Bonn
[2] Lower Saxony State and University Library, Göttingen

Summary

Clearinghouses – or subject gateways – are domain-specific collections of links to resources on the Internet. The links are described with metadata and structured according to a domain-specific subject hierarchy. Users access the information by searching in the metadata or by browsing the subject hierarchy.

The standards for metadata vary across existing clearinghouses and different technologies for storing and accessing the metadata are used. This makes it difficult to distribute the editorial or administrative work involved in maintaining a clearinghouse, or to exchange information with other systems.

DBClear is a generic, platform-independent clearinghouse system, whose metadata schema can be adapted to different standards. The data is stored in a relational database. It includes a workflow component to support distributed maintenance and automation modules for link checking and metadata extraction. The presentation of the clearinghouse on the Web can be modified to allow seamless integration into existing web sites.

1 Introduction

Clearinghouses – also called subject gateways – are domain-specific collections of links to high quality resources on the Internet. Experts judge the relevance of the resources, describe them according to a predefined metadata schema and assign them to a subject hierarchy or classification. Clearinghouses try to give users orientation on the fast changing World Wide Web (WWW) and efficient access to relevant online information. Well-known examples are CetusLinks[1], Geo-Guide[2] (Enderle 1999) and SocioGuide[3] (Hellweg 2000).

In the first years of the WWW, clearinghouses were maintained as lists of bookmarks to Internet sites, often collected, maintained and published by a single person and stored as static HTML files on a web server. Understood as vertical (thematically limited) directories, they were relatively small, structured by a single subject hierarchy, and described the resources only by a very limited set of metadata elements. When growing past a certain size, it became difficult and time consuming to manage and maintain these bookmark lists, especially if the gateway's domain was subject to rapid change. Furthermore, file-based bookmark lists do normally not support to structure their content according to different subject hierarchies at the same time and the describing metadata can not efficiently be used for searching. New content was acquired mostly by browsing the hypertext structure of the Internet for relevant resources, by using search engines to find topically related sites, by exchanging links with fellow researchers, or even by chance.

But in most cases, a single person can not have a complete overview on a particular subject or domain. As a consequence, finding new sites, judging their relevance, describing resources and

[1] http://www.cetus-link.org
[2] http://www.geo.giude.de
[3] http://www.gesis.org/SocioGuide/

maintaining them must be spatially and temporally distributed to several persons in order to keep the link collection at the same high level of quality and relevance. Automation techniques are needed to support the editors in keeping the collection as complete and up-to-date as possible. Furthermore, a single shared database must be maintained to ensure the consistency of the link collection and to allow the flexible generation of views on the data with current web technologies.

2 Main features of DBClear

To solve the problems listed above, the Social Science Information Centre (IZ), Bonn develops –together with the Lower Saxony State and University Library Göttingen (SUB Göttingen) – a generic, multilingual clearinghouse system, DBClear[4]. The project is funded by the Deutsche Forschungsgemeinschaft (DFG)[5]. DBClear is based on a relational database management system and allows distributed maintenance and administration of its content. Its main features are:

- Support of different, user-definable metadata schemas and bibliographic standards.
- Storage of the clearinghouse's content in a relational database management system with a JDBC (Java DataBase Connectivity) compliant interface.
- A Workflow system to route the clearinghouse's content from the initial quality check all the way to the release on the Internet.
- Automatic metadata extraction from HTML pages.
- Modules to automate frequently recurring tasks, like checking if links are still reachable or have been updated.
- Support for multiple languages at the metadata and the user interface level.
- Separation of content and presentation, allowing flexible adaptation and branding of content for different usage scenarios.
- Gateways (e.g. Z39.50) to integrate DBClear into larger contexts, like Virtual Libraries.
- Import and export of data with support for mapping DBClear metadata to and from other metadata standards (e.g. Dublin Core).
- Generic and platform-independent Java-based system with plug-in architecture for future extensions (e.g. metadata generated on-the-fly from external systems).

3 Target scenarios for using DBClear

When designing the overall architecture of DBClear, a number of scenarios were evaluated where DBClear and its content could be integrated with existing systems. This includes integration in web presentations of institutes, connecting DBClear with legacy systems, or searching its content as part of a larger Virtual Library. The technologies and standards used in developing the system where determined by their support for these target scenarios.

3.1 DBClear as a standalone portal system

The most common scenario for using DBClear will be as a standalone portal system which is part of an organization's web site. Here it is necessary to adapt the look and feel of DBClear to the surrounding web pages. Through strict separation of content and presentation, it is possible to design and implement different user interfaces. DBClear generates an intermediate XML format from database content and then uses XML/XSL (eXtensible Style sheet Language) transformations to generate views for browsing and searching the clearinghouse and for displaying information. The XSL style sheets allow to include a wide range of Internet technologies in the generated

[4] http://www.gesis.org/research/information_technology/DBClear.htm

[5] DBClear is funded under grant no. Gz: III N – 554 922(1)00

tions to generate views for browsing and searching the clearinghouse and for displaying information. The XSL style sheets allow to include a wide range of Internet technologies in the generated HTML pages (e.g. advanced functions in JavaScript) and to transform data into different formats. Values stored in the system's database in one format (e.g. a numerical value which describes a resource's relevance) can be rendered as text (e.g. "high", "medium", "low") or as graphics (e.g. a number of "stars" representing the relevance).

This flexibility in generating different output formats can also be used in settings where multiple organizations collaborate in creating and maintaining a subject gateway and where every participant "donates" and maintains its existing sub-collection (figure 1). A single instance of DBClear – running at one institute – can be integrated in every institute's own web site and use different style sheets. This allows every organization to adapt DBClear to its corporate layout and to brand content with the contributing organization's logo.

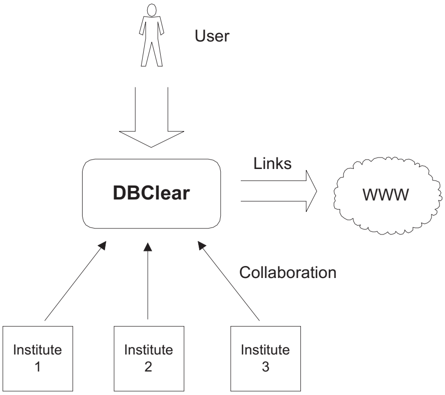

Figure 1: DBClear as a standalone portal

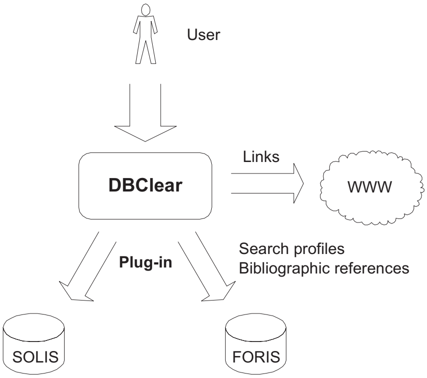

Figure 2: Using external data sources

3.2 Using external data sources

A subject gateway is often maintained as one of a larger number of topically related information services. The Social Science Information Centre (IZ) for example offers a reference database for social science literature (SOLIS), a database with research projects (FORIS) and smaller databases with institutions and conferences. To achieve the maximum benefit for the user, these databases should be integrated with DBClear.

To access external data sources, the DBClear architecture contains a plug-in interface to extend the system with additional features in the form of Java classes, which are dynamically loadad. New features could be "live" attributes, whose values are not stored in the DBClear database, but are retrieved from external data sources and generated on-the-fly (figure 2). By defining an attribute which contains identifiers of bibliographic records in SOLIS or FORIS, Internet resources could be linked with highly relevant external data. In a similar way, search profiles could be stored for single or groups of resources and executed every time a user accesses these resources. This technology is not limited to locally available sources, also external search engines or harvesters – specialized in searching well defined parts of the Internet – could be integrated.

3.3 Integrating DBClear with library catalogues

In other contexts, e.g. ViBSoz, the Social Science Virtual Library[6] (Meier et al. 2000), the resources contained in a clearinghouse could enrich the information contained in other information systems, e.g. library catalogues. DBClear therefore contains an interface which allows third party systems to search its content (figure 3). Currently, the Z39.50 protocol is supported, which is widely used for connecting universities' library catalogues for integrated searches. In ViBSoz, it connects library catalogues with the collection of the Friedrich Ebert Stiftung[7] and the IZ's database SOLIS[8]. The Z39.50 interface of DBClear uses a mapping between the bib-1 attribute set of Z39.50 and the DBClear metadata schema. This mapping can be freely defined and adapted.

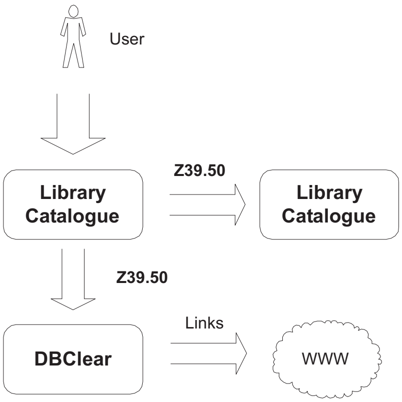

Figure 3: Integration into other systems

4 Storing metadata in DBClear

One of the primary goals of DBClear was to develop a system whose metadata schema can be adapted to nearly every standard. This should not only include official standards (e.g. Dublin Core, see DCMI 1999), but also standards which are currently defined in projects like RENARDUS[9], where the metadata schemas of 12 gateways are mapped onto a common schema suitable for cross searching with a broker architecture.

The following sections describe the fundamentals of the DBClear metadata schema and illustrate its flexibility. In addition, an overview of the overall system architecture is given.

4.1 Metadata

To design a generic, multilingual clearinghouse system, several existing clearinghouses and Internet portals were analyzed. They used different sets of metadata to describe the Internet resources and various subject hierarchies (e.g. classifications) to group and structure them thematically. These differences were not only domain-dependent – clearinghouses on the same subject or domain differed as well.

[6] ViBSoz is funded by the Deutsche Forschungsgemeinschaft (DFG), http://www.vibsoz.de/

[7] http://www.fes.de/

[8] http://www.gesis.org/Information/SOLIS/

[9] http://www.renardus.org/

4.1.1 Metadata types

DBClear has a generic and flexible model for representing metadata, which allows every clearinghouse to define its own set of metadata elements. The metadata elements are divided into *facets* and *attributes* (figure 4). For both, the cardinality (i.e. how many values must be entered at minimum and how many may be entered at maximum) can be defined by the administrator.

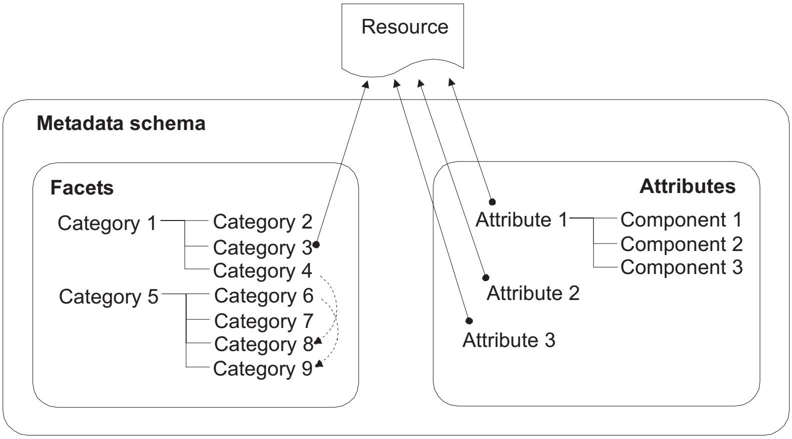

Figure 4: Metadata schema with facets and attributes

Facets are used for metadata elements which consist of a controlled vocabulary whose values – called *categories* – do not change frequently (e.g. a classification, thesaurus, or a list of index terms or country names). The categories may be semantically related to each other (e.g. broader/narrower term relations in a thesaurus) or arranged in a mono-hierarchic list. These relationships can be used to browse the clearinghouse's content.

Attributes, in contrast to facets, are used to model sets of values which are not limited in size and where the single values may not be limited by any formal restriction (e.g. title, author, postal address). Attributes can consist of components (e.g. the attribute *author* may consist of the components *first name*, *last name* and *e-mail address*), which allow more detailed searching and a more flexible definition of output formats. For some metadata elements whose values are naturally limited (e.g. the names of cities), a controlled vocabulary may not always be available. This makes it necessary to use an attribute instead of a facet. In this situation, the reuse of previously entered values is helpful to limit homonyms and to reduce spelling errors. DBClear allows to mark attributes as "reusable" and then presents a list of existing values as needed.

In a clearinghouse with multilingual content, some metadata elements may be language dependant (e.g. keywords or the abstract), whereas others are not (e.g. publisher). In DBClear, a list of languages that defines which languages are optional or mandatory can be assigned to every metadata element. For facets, the vocabulary itself is stored in multiple languages. By associating a category of a facet with a resource, every translation of the category is automatically associated with the resource, too. In contrast, values for language dependant attributes have to be entered separately for every language.

4.1.2 Organizing metadata with stocks

In a clearinghouse, different types of resources may be collected (e.g. links to institutions' home pages, online dissertations or a calendar of events), which can even belong to different domains. The set of metadata elements describing a resource may vary depending on its type. To describe for instance homepages of institutions, the elements *country* and *city* may be useful, which are not applicable to resources like online dissertations.

A *stock* in DBClear defines a group of semantically related resources, together with its set of metadata elements (figure 5) and serves as a blueprint to automatically generate entry forms for the maintenance of the content. Stocks can be used to filter data during searching and browsing, so that only resources of the specified type are visible to the user. Resources are normally assigned to only one stock, taking into account that stocks will have different metadata elements. Cross-searches over multiple stocks are possible, if they share at least one attribute.

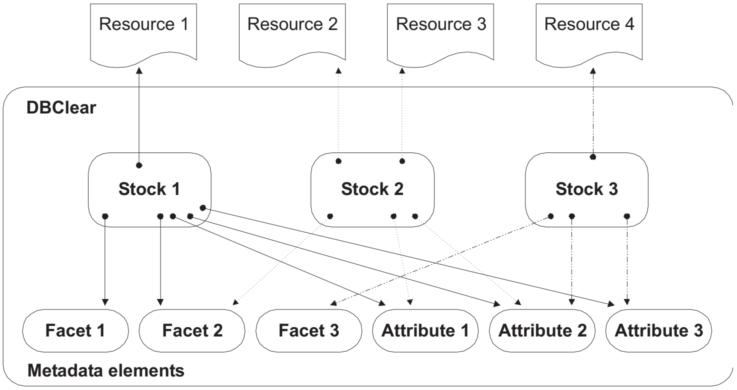

Figure 5: Usage of stocks to organize resources and metadata

4.2 System architecture

DBClear is designed as a multi-tiered information system (figure 6), consisting of:

- A presentation layer, which presents data to the user or editor and permits data manipulation and data entry.
- An application layer, which contains the business logic of the clearinghouse system.
- A data abstraction layer, which stores/retrieves the data in/from a relational database.

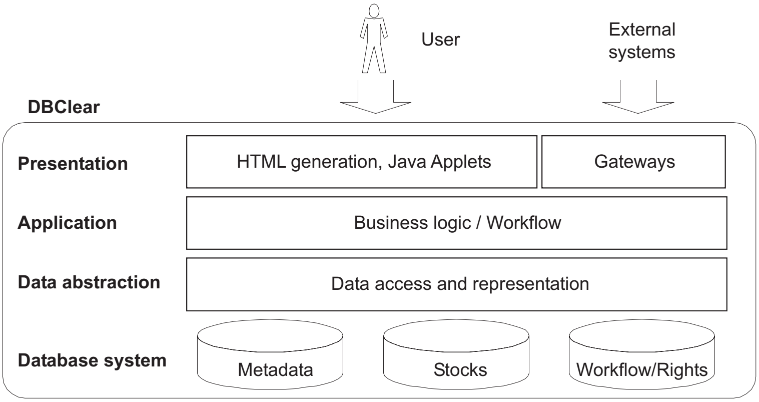

Figure 6: Multi-layered architecture

The application layer receives requests from clients (e.g. Java Servlets), manipulates data according to workflow definitions, and retrieves or stores data using the data abstraction layer. During processing, data is represented in an intermediary XML format. This representation is transformed using XSL style sheets and the resulting output is delivered back to the client.

Using an XML representation of the data and transforming it with XSL into viewable content allows the separation of content (XML) from presentation (e.g. HTML). The presentation can be adapted to different clients according to their display features (e.g. different browsers, Java Applets or WAP-enabled cell phones). It is also possible to tailor the presentation to the corporate design of different institutions collaborating in the same clearing house by simply adjusting the style sheets.

5 Automation and workflow

To facilitate the creation and continuous maintenance of a data collection that exceeds a certain size, tools for automation of tasks are required, as well as support for the distribution of work among several cooperating editors. DBClear provides a number of modules to automate recurring tasks and a workflow system to route information between the people involved in a clearinghouse.

5.1 Automation

Several aspects of clearinghouse maintenance can be automated, the most obvious are regular checks if an Internet resource can still be reached or was modified since the last time an editor had a look at it.

Interviews with people involved in running clearinghouses (e.g. editors and administrators) showed that a number of recurring processes or tasks are not very complex and can either be automated completely – or at least be supported – by the clearinghouse software. This is especially the case with data extraction or classification tasks, which occur mostly in the process of analysing and describing new resources. Here, the system can automatically extract data from the HTML pages by looking for special mark-up elements, e.g. the META elements describing the document's content, or a sequence of elements that suggest a certain document structure. As results from the CARMEN project show (Strötgen 2002), paragraphs of a document can be classified as e.g. abstract, author or keywords with high precision by using heuristics. Additional information can be extracted from the protocol information (HTTP header) a web server sends with each page or from the URL, which can contain country codes or institution names.

The extracted information can be used to produce suggestions for DBClear metadata elements which are then presented to the editors in the normal course of describing a resource and its content. Other values which can be computed automatically are the number of "backlinks" and the document language. The number of backlinks shows how many external sites refer (link) to the resource in question. Its value is usually determined automatically by querying search engines like Google or AltaVista, and therefore can even be updated regularly without user interaction. The document language can be guessed by statistical means, taking the language-specific frequency of certain characters or combinations of characters into account.

DBClear provides a general framework for the implementation of these metadata extraction modules and allows each of its metadata elements to be linked to such a module. Modules for text or header extraction are provided as part of the core system. They can be configured to extract elements like the date or title of web pages and also to retrieve backlinks information.

To maintain the clearinghouse's content and keep it as complete and up-to-date as possible, regular searches for new and relevant content on the Web are necessary. This is usually accomplished by accessing one or more search engines with a predefined query which covers the relevant aspects of the domain, and comparing the results to the resources that are already known (either stored in the clearinghouse's database or in a list of rejected resources).

DBClear allows each clearinghouse editor to define and store any number of queries to external search engines. These queries are executed regularly, and their results are compared to the global list of known resources, as well as to the editor's personal rejection list, which is maintained as part of his personal configuration. New resources are inserted into the editor's personal collection of resources, which he can review and then decide, which resource to reject or to put on his or another editor's work list.

5.2 Workflow

DBClear supports distributed collaboration among clearinghouse editors by allowing to assign a sequence of tasks (workflow), that have to be performed on a resource, to different persons. Each task holds information on the resource and the activities that have to be performed on it. To coordinate the tasks, the system maintains information on the state of a resource together with a work list for each editor. A work list contains the tasks assigned to a person or a group.

Tasks are based on the type of a resource (its metadata) and the (group of) people who carry out the task. Each task consists of activities which can be tied to metadata elements. Rules evaluate the values of those metadata elements and decide which activity has to be performed next, or how the state of a resource changes. This activity can either be associated with an automation module, or with a (group of) person(s). In the latter case, a new item consisting of the resource and information on the activity is entered into the respective person's work list. The person to which a task is assigned may manually change the resource's state or forward the task to someone else in case the rules didn't apply or an exception from the assumed sequence of tasks occurred.

Actions, like checking if a resource is reachable or has been modified, act as initiators of workflow processes, like suggestions of new resources by users of the clearinghouse do. For already catalogued resources, the system can determine the responsible persons and create an entry in the appropriate worklists. For new resources, especially if sufficient information was not provided by the user, the content analysis modules can be used to generate enough context for the workflow rules to determine the initial person responsible for this resource.

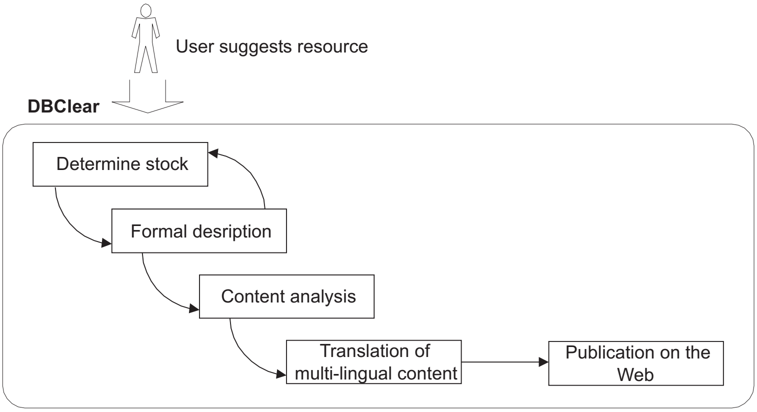

Figure 7: Example workflow sequence in DBClear

Figure 7 shows a simple workflow sequence, in which a user's suggestion is categorized and assigned to a stock, based on the information supplied. By using this initial information, an editor (or a group of editors) is selected for entering the formal description of the resource, like country of origin, language or resource type. If this editor decides that the automatic assignment to the stock was incorrect, he is free to reassign the resource to some other stock. Once the formal information is entered, the resource is forwarded to the next person responsible for content analysis (e.g. writing an abstract) and indexing. The following translation step has to be performed by an editor with knowledge of the required language. The final publication of the resource on the Web is performed by the editor responsible for the consistency of the collection.

6 Conclusion

With the migration of GeoGuide and SocioGuide – two large clearinghouses with different metadata schemes – DBClear has already proven its flexibility as data storage and system features are concerned. Both clearinghouses are currently prepared for release on the Internet. Besides that, additional metadata schemes and workflow requirements are analysed to make sure that DBClear can also be adapted to them.

After project completion in September 2002, the DBClear system will be open for use at other institutes and organizations who want to offer a clearinghouse on the Internet.

7 References

DCMI (1999): Dublin Core Metadata Element Set, Version 1.1: Reference Description. Dublin Core Metadata Initiative, available at .

Enderle, W. et al. (1999): Das Sondersammelgebiets-Fachinformationsprojekt (SSG-FI) der Niedersächsischen Staats- und Universitätsbibliothek Göttingen: GeoGuide, MathGuide, Anglo-American History Guide und Anglo-American Literature Guide. dbi-materialien 185, DBI Berlin.

Hellweg, H. (2000): Der GESIS Socio-Guide: ein kooperatives Link-Verwaltungs-System. In: Ohly, P.; Rahmstorf, G.; Sigel, A. (Eds.). Proceedings der 6. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO), 23.–25. September 1999, Hamburg, S. 291-298.

Meier, W.; Müller, M. N. O.; Winkler, S. (2000): Virtuelle Fachbibliothek Sozialwissenschaften. Problembereich und Konzeption. In: Bibliotheksdienst, 34, 2000, Nr. 7/8, S.1236-1244.

Strötgen, R. (2002): Treatment of Semantic Heterogeneity using MetaData Extraction and Query Translation. In: Proceedings of CRIS 2002: Gaining Insight from Research Information, Kassel, 29. - 31. August 2002 (to appear).

8 Contact Information

Maximilian Stempfhuber
Informationszentrum Sozialwissenschaften
Lennéstr. 30
D-53113 Bonn
Germany

e-mail: st@bonn.iz-soz.de

ISBN 3-933146-844