In: Eibl, Maximilian; Wolff, Christian; Womser-Hacker, Christa (Hg.): Designing Information Systems. Festschrift für Jürgen Krause. Konstanz: UVK Verlagsgesellschaft mbH, 2005. S. 198 – 208.

Towards Expressive and User Friendly Interfaces for Digital Libraries Containing Heterogeneous Data

Maximilian Stempfhuber

Abstract
Digital libraries in Germany are currently making a shift from more self-contained projects or institutional driven activities to concerted actions coordinated by government and funding agencies. The "library" aspect of collecting things and cataloguing them in a standardized format and with a standardized indexing language is no longer the primary concern here. The goal is to combine heterogeneous information from different, often distributed sources and to make them available in an integrated way. As the user interface is the intermediate between the user's information needs and the digital library's complex inner structure, new ways of designing the user interface are needed to adapt to the growing complexity while ensuring a consistent look and feel.

1 The Changing Face of Digital and Virtual Libraries

1.1 Defining Digital and Virtual Libraries

Digital or virtual libraries, often understood as access points to collections of electronic documents or digitized artifacts on the Internet, play a more and more important role for the access to scientific information. The basic idea is often accredited to Vannevar Bush, who develops the idea of the "memex", an automated personal library which can not only store a user's individual information but can be filled with purchased information (e.g. books or newspapers) on microfilm: "Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, 'memex' will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory" [Bush 1945]. Many other concepts of today's digital libraries and hypertext systems, with the World Wide Web being the biggest in existence, were already sketched out by Bush in great detail, just at about the time the first computer had been built by Konrad Zuse.

Though often used as synonyms, both concepts can be distinguished by the result of a user's search for information. [Kochtanek et al. 2001] list different definitions of digital libraries and emphasize the importance of electronically available documents in contrast to virtual libraries, which hold only metadata and references to documents. [ARL 1995] takes a broader view and states some common properties of digital libraries [cf. Drabenstott 1994]:

- The digital library is not a single entity;
- The digital library requires technology to link the resources of many;
- The linkages between the many digital libraries and information services are transparent to the end users;
- Universal access to digital libraries and information services is a goal;
- Digital library collections are not limited to document surrogates: they extend to digital artifacts that cannot be represented or distributed in printed formats.

For the rest of this article both are used synonymously, focusing on the commonalities in the information retrieval process, which essentially is the same for digital and virtual libraries. The possibilities of e.g. full text or image retrieval will not be discussed here.

1.2 Recent Developments in Germany's Digital Libraries

In Germany currently two funding programs run in parallel but are coordinated at the political level: Four scientific information networks, funded by the Federal Ministry for Education and Research (BMBF), and more than 20 virtual libraries[1], funded by the Deutsche Forschungsgemeinschaft (DFG). All projects defined their information services according to the needs of their scientific discipline, relevant user groups, and to the information resources at hand. While the virtual libraries mostly offer quality controlled subject gateways or clearinghouses for resources on the Internet together with library catalogues, the basis of the scientific information networks are either high quality reference databases with costs or newly created, free databases for the same purpose, both linked to electronic full text documents and electronic document delivery services. Apart from these types of information, some virtual libraries offer information on research projects, book reviews, maps and other materials (e.g. digitized historical documents), but the offers vary very much between the disciplines.

[1] Die Virtuelle Fachbibliothek Vifanet (2004). Die Virtuelle Fachbibliothek Homepage. http://www.virtuellefachbibliothek.de [Access September 2004].

At the level of content representation and analysis, like metadata schemas or indexing vocabularies, currently no single common standard exists across all information networks and virtual libraries. Depending on the institutional context, domain-specific thesauri or classification systems are used in the information networks, while the virtual libraries mostly use the thesaurus (Schlagwortnormdatei, SWD) or classification (Basisklassifikation) of the German National Library for literature references or the Dewey Decimal Classification[2] (DDC) for internet resources. The metadata schemas used in the virtual libraries often conform to the standard structure of library catalogs or the Dublin Core Metadata Element Set[3] (DC) for literature references, and in addition a proposal for a common metadata schema for cross-searching virtual libraries as a DC application profile (VLIB[4]) exists. The metadata structures used for literature references throughout the information networks are normally much more detailed than in the library context and contain additional elements (e.g. abstracts), but are not necessarily based on common standards.

Since 2002, the information networks and virtual library are more tightly integrated, aiming at a new scientific information portal *vascoda*[5] which integrates all relevant information sources from all disciplines at a single point of access. With this decision, the heterogeneity in the information collections connected to *vascoda* becomes an issue, as now a user's query is sent to many databases with specific content, content analysis and metadata structures. The differences in semantics of search terms between disciplines and databases suddenly become obvious, and especially users interested in cross-domain or cross-database searches face difficulties in formulating precise queries or interpreting the results. Without better support for the users of such large and interdisciplinary portals it has to be doubted whether the results achievable will – despite the controlled and highly relevant content provided – be rated higher by users than what is provided by current search engines on the Internet.

---

[2] Online Computer Library Center (2004). Dewey Services – Dewey Decimal Classification. http://www.oclc.org/dewey/ [Access September 2004].

[3] Dublin Core Metadata Initiative (2004). Dublin Core Metadata Element Set, Version 1.1: Reference Description. http://www.dublincore.org/documents/dces/ [Access September 2004].

[4] SUB Göttingen (2002). VLib Empfehlungen: Metadaten-Core-Set. http://www2. sub.uni-goettingen.de/metacore/empfehlungen/index.html [Access September 2004].

[5] Technische Informationsbibliothek Hannover (2004). Vascoda Homepage. http://www.vascoda.de [Access September 2004].

1.3 The Challenges of Digital Libraries

Looking at the number of players involved in portals like *vascoda* and the amount of resources needed to standardize content analysis and metadata schemas in existing databases, it seems not feasible that the transition to a small number of common standards is feasible in the near future. While it is advisable to build new information collections on existing standards, other solutions have to be developed for dealing with the heterogeneity of existing collections – a position which recently also has been acquired by the German standardization body [DIN 2003].

The challenges arising from the integration of heterogeneous content are manifold and complex, and manifest at all levels of an information system. On the level of metadata and communication, structural heterogeneity can be found. Here, formal methods exist for schema evolution, schema integration or mediation (e.g. using software agents) in distributed, heterogeneous contexts, but a broadly accepted model or architecture which covers all relevant varieties of information sources is currently not available. Activities based on open technologies and standards, like Web services or metadata registries are promising and might enable users in the future to "plug" together information sources as needed – without paying attention to structural or technical issues.

For dealing with semantic heterogeneity between databases, which is caused by different indexing vocabularies, cross-concordances and statistical transfer modules can be used. Both try to map between the specific languages used in different thesauri or classifications, or even the uncontrolled keywords assigned by authors. While cross-concordances directly map between a single word from one thesaurus to either a single word or a combination of two or more words from the second thesaurus – and therefore ignore how the thesaurus entries are actually used for indexing documents – statistical transfer modules are based on documents which are indexed with two different thesauri at the same time (parallel corpora) [Hellweg et al. 2001]. They take into account how entries from different thesauri are actually used for describing the content of the same document in a large document collection and calculate the likelihood of groups of entries from different thesauri to appear as descriptors for the same document. In cases where parallel corpora do not exist, they can be simulated by using the first thesaurus for a free-text search in a document collection and use the documents with the highest ranks in the result set for calculating the likelihood mentioned above [Stempfhuber et al. 2002]. First tests yield promising results for cross-concordances and statistical transfer modules, but a more in-depth evaluation is needed to further optimize them to specific domains and to identify good combinations of both. Of concern are also the resources needed to create and maintain the underlying

knowledge structures. This is especially the case for the intellectually built cross-concordances, for which the maintenance effort increases with each new indexing vocabulary. It seems reasonable that only co-operations between institutes and the coordinated creation and management of the knowledge structures can ensure a sustainable service, for which a conceptual model still has to be developed.

2 A User Interface for Handling Heterogeneity

Heterogeneity in information systems is also reflected at the user interface level, because this is where an information seeking user expresses his information needs by the means of a formal or graphical query language. As soon as the heterogeneity can not be handled automatically, adequate means have to be provided for the user to explore and actively use information about structural and semantic differences between the data collections. The challenge here is to take into account the different skill levels and user requirements concerning interaction with the information system and domain knowledge, which influences the search strategy and the complexity of a query. But also the changes a single user goes through, either during a longer period of regularly using a system or during a single session, where its information needs and search strategies might change several times, have to be considered and different, seamlessly connected modes of interaction have to be provided.

An example where heterogeneity is reflected at the user interface level is the information network for pedagogic, social sciences and psychology, *infoconnex*[6]. In *infoconnex*, three domain-specific reference databases for literature are available, each with its own thesaurus and metadata schema. The user can choose to search in only one database at a time by using the matching thesaurus and query form, or he may perform a cluster search over all three databases simultaneously if its information needs require. For mapping the user's search terms to each database, cross-concordances are used which pair wise map between the three thesauri. This allows the user to resort to the specific language of its domain (e.g. the social sciences) for formulating his query and at the same time to receive very precise results from the other domains (e.g. pedagogic and psychology). Depending on the knowledge and experience a user has it is a requirement to make the mapping of search terms visible and let the user also change the mapping if alternatives exist.

[6] Informationszentrum Sozialwissenschaften (2004). infoconnex Homepage. http://www.infoconnex.de [Access September 2004].

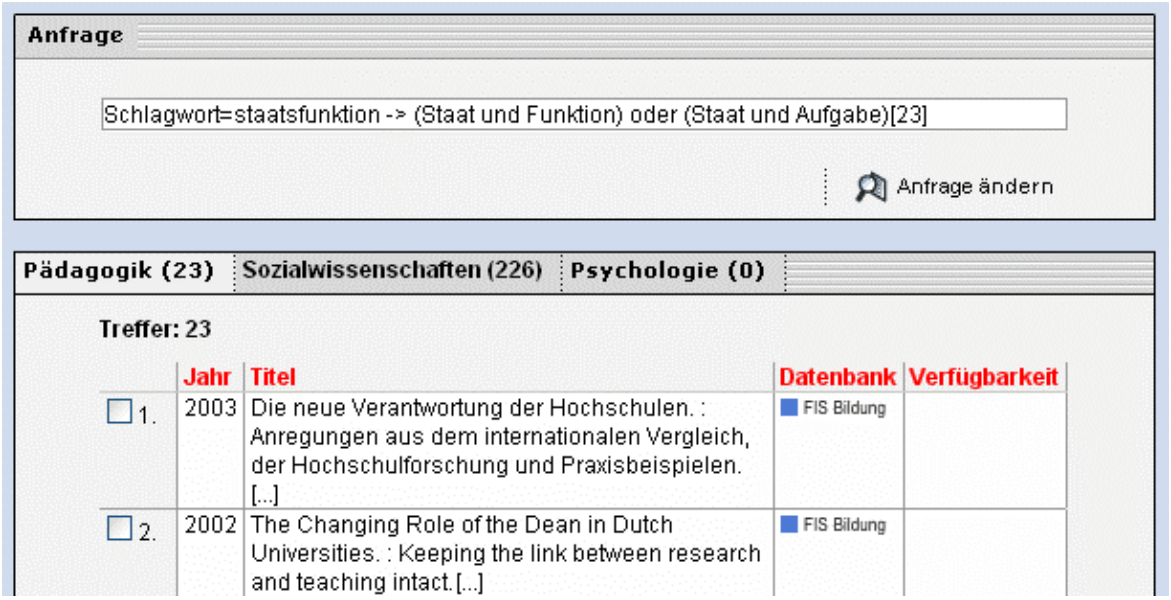

Fig. 1: Displaying information about transformation of search terms

Figure 1 shows an example where the descriptor “staatsfunktion” (governmental function) is automatically transformed into “staat UND function” (government AND function) and “staat UND aufgabe” (government AND duty). The status display explains the transformations, helps the user to build up knowledge about the domains (i.e. the different use of indexing vocabularies) and lets him compare the query with the result set produced. At the time of writing, only term transformations with high relevance are used. Users of *infoconnex* will in a future version be able to use broader term and narrower term relationships in addition which will make it necessary to provide means for selecting the best alternatives for a given search strategy.

The status display currently used in *infoconnex* serves only the purpose of reducing the user’s short term memory load during the exploration of query results but fails short of meeting the requirements stated in the WOB model [Krause 1995] for designing object-oriented user interfaces based on the “tool metaphor”. One of the main principles here is that information entered by the user in one screen should not only be available in every following step of the sequence of screens necessary to complete a given task, in fact the user should be able to modify the information whenever and wherever needed. This is similar to the “output-is-input” principle [Ahlberg & Shneiderman 1992] which states that the output of a system should at the same time serve as input, like it is in hypertext systems where the output (the page displayed) is also input (the hypertext links embedded in the page) and the user doesn’t have to revert to a different mode for interacting with the system.

A related requirement for interactive and responsive information systems is direct, immediate and dynamic feedback to the user in a way that helps him to

understand the consequences of changes to the query for the result set. Here, a tight coupling between the user interface widgets for query formulation and the visualization of the query result – together with a preview of the expected result in cases where the calculation of the exact result set would be too time consuming – lets the user narrow down his search step by step, showing immediately the effect of changes in the query and avoiding unexpected empty result sets [Plaisant et al 2001].

2.1 Dealing with Complexity

While dynamic query interfaces and query previews provide ways for making information systems more interactive and responsive, they are by themselves no means for visualizing complex semantic relationships within the data of an information system. Here, generic types of visualizations are needed which can be adapted to specific use cases and levels of complexity. ODIN, a framework for object-oriented, dynamic user interfaces [Stempfhuber 2003], provides solutions for this problem which were developed in the context of information systems for time-series data [Stempfhuber et al. 2002] and text documents. The features of ODIN and its elements are based on the needs of information specialists and end users, and cope with the broad range of requirements by task-based self-adaptation of screen layout and information density, and by letting the user directly adapt the complexity of the user interface elements.

The user interface elements of ODIN are based on visual formalisms (Nardi & Zarmer 1993), dynamically change size and content, and are adaptable to the user's specific needs. The basic visual formalism is an interactive table, whose cells are either empty or occupied by a two-state checkbox. The checkbox reflects an existing semantic relationship between both dimensions of the table and can be activated or deactivated by the user. It can be used for query formulation and exploration of result sets at the same time (output-as-input principle).

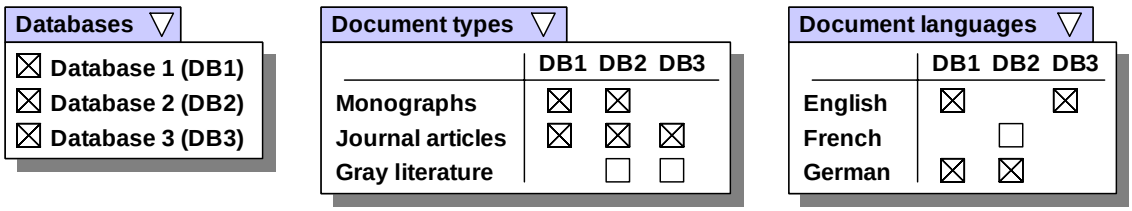

Fig. 2: Primary and secondary filters

Figure 2 shows a primary ("databases") and two secondary filters ("document types", "document languages") in ODIN. Filters can be opened and closed to save space and are used for settings which remain unchanged over a longer

period of time. It may be seen as a limitation to have only two dimensions for combining attributes because many more attributes are involved in query formulation, but user interviews showed that there is a natural hierarchy which determines a primary attribute (e.g. the databases to be searched) which is then combined pair wise with additional attributes. A combination of document type and language for example is not relevant in practice.

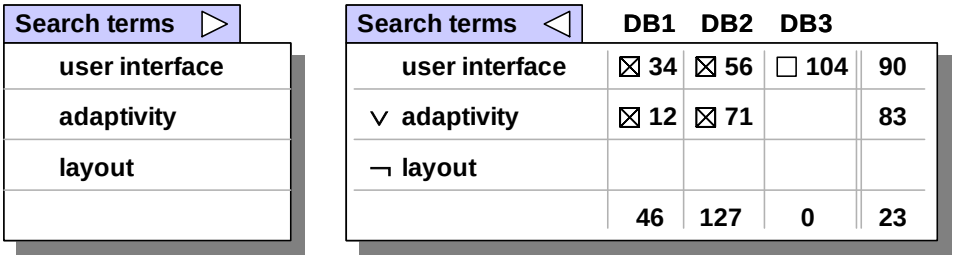

Figure 3: Two states of the user interface with reduced and maximum complexity

The tabular display of filter attributes in ODIN is also used for query formulation, so the user faces consistent visualization and interaction principles. Figure 3 shows a control for entering search terms in its compressed (left) and extended form (right). The cells of the table again visualize the semantic relationships in the data, i.e. whether a search term is valid for a specific database. For searches with Boolean logic, the Boolean operators, the number of hits per search term and database, the total of hits per search term, and the total of hits per database can additionally be displayed.

ODIN lets the user configure the complexity of the controls at a very fine-grained level so that novice users can focus on searching information while experienced users may explore in detail the effect a search term has in the different databases. In addition, context-sensitive fly-over information can be presented for each table cell which includes mapping information from the search term entered by the user to the search terms actually used for searching in the specific database. This reflects the maximum level of detail for using and exploring the knowledge structures available for dealing with the heterogeneity of the databases, i.e. cross-concordances and statistical mappings.

2.2 Evaluation of the ODIN user interface design

The design principles of ODIN have been implemented in the context of geo-referenced data, MURBANDY [Hermes et al. 2003], and reference databases for literature. Both implementations have been the basis for heuristic usability testing with the goal to detect general problems of the abstract visual formalism and cross-cultural differences [Stempfhuber et al. 2003]. The tests have been carried out with 10 subjects at the University of Koblenz-Landau and 6 subjects at the Pai Chai University in Daejon (South Korea) with the German

subjects carrying out tasks with both systems and the Korean subjects using only MURBANDY. The results showed that the principle behind the visual formalism was well comprehended without any introduction by nearly all subjects and that problems arose exclusively from deficits in the implementation of the user interfaces (e.g. too small check boxes or double clicks for editing cells). As predicted, there were no significant differences in the cross-cultural comparison.

3 Conclusion

Heterogeneity in information systems can not in all cases be handled automatically by the system. Certain requirements or retrieval strategies make it necessary to visualize the complex semantic relationships within the data and make them accessible for in the retrieval process. ODIN represents a way for bringing this heterogeneity to the user interface level in a very flexible and consistent way, allowing users with different demands to seamlessly switch between simple and complex search strategies by adapting the user interface elements.

Building on the positive results from user tests, ODIN is currently re-designed in cooperation with Bauhaus University, Weimar, and implemented for the *infoconnex* information network. This will allow further evaluation of usability and effects on retrieval quality in a context with heterogeneous databases.

4 References

Ahlberg, Christopher; Shneiderman, Ben (1994). “Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays.” In: Proceedings of CHI’94 Human Factors in Computing Systems. Boston, Massachusetts, April 24-28, 1994, 313-17.

ARL (1995). Definition and Purposes of a Digital Library. Association of Research Libraries, Washington, DC, October 1995. http://www.arl.org/sunsite/definition.html [Zugriff September 2004].

Bush, Vannevar (1945). “As We May Think.” In: Atlantic Monthly 176 (1), 101-8. http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush-all.shtml [Zugriff September 2004].

DIN (2003): Standardization in Information and Communication Technology (ICT). German Positions. DIN Deutsches Institut für Normung e. V. Strategy Committee on Standardization in Information and Communication Technology (SICT). http://www.ni.din.de/sixcms_upload/media/1436/sict_artikel_engl.pdf [Zugriff September 2004].

Drabenstott, Karen M. (1994). Analytical review of the library of the future, Washington, DC: Council Library Resources, 1994. http://www.eff.org/ Infrastructure/Regional_rural_edu/library_future_review.ps.gz [Zugriff September 2004].

Hellweg, Heiko; Krause, Jürgen; Mandl, Thomas; et al. (2001). “Treatment of Semantic Heterogeneity.” In: Information Retrieval. IZ Working paper 23. Bonn: IZ Sozialwissenschaften. http://www.gesis.org/en/publications/reports/ iz_working_papers/ [Zugriff September 2004].

Hermes, Bernd; Stempfhuber, Maximilian; Demicheli, Luca; et al. (2003). “MURBANDY: the (so far) Missing Link; User-Friendly Retrieval and Visualization of Geographic Information.” In: Schader, Martin; Gaul, Wolfgang; Vichi, Maurizio (eds.) (2003). Between Data Science and Applied Data Analysis: Proceedings of the 26th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Mannheim, July 22-24, 2002, 438-46.

Kochtanek, Thomas R.; Hein, Karen K.; Kassim, Ahmad Rafee Che (2001). "A digital library resource Web site: Project DL.” In: Online Information Review 25 (1) 29-40. http://www.emeraldinsight.com/pdfs/oir251.pdf [Zugriff September 2004].

Krause, Jürgen (1995). Das WOB-Modell. IZ Working paper 1. Bonn: IZ Sozialwissenschaften. http://www.gesis.org/en/publications/reports/ iz_working_papers/ [Zugriff September 2004].

Nardi, Bonnie A., Zarmer, Craig L. (1993). ”Beyond Models and Metaphors: Visual Formalisms in User Interface Design.” In: Journal of Visual Languages and Computing 4, 5-33.

Plaisant, Catherine; Shneiderman, Ben; Tanin, Egemen; et al. (2001). Dynamic Queries and Query Previews for networked information systems: the case of NASA EOSDIS. http://www.cs.umd.edu/hcil/eosdis [Zugriff September 2004].

Stempfhuber, Maximilian (2003). Objektorientierte Dynamische Benutzungsoberflächen – ODIN. Behandlung semantischer und struktureller Heterogenität in Informationssystemen mit den Mitteln der Softwareergonomie. Forschungsberichte Band 6. Bonn: IZ Sozialwissenschaften.

Stempfhuber, Maximilian; Hellweg, Heiko; Schaefer, André (2002). "ELVIRA: User Friendly Retrieval of Heterogeneous Data in Market Research.” In Callaos, N., Hernández-Encinas, L., & Yetim, F. (eds.) (2002). Proceedings of SCI 2002, 6th World Multiconference on Systemics, Cybernetics and Informatics; Orlando, USA, July 14-18, 2002. Vol. I: Information Systems Development, 299-304.

Stempfhuber, Maximilian; Kim, Do-Wan; Petrick, Marco (2003). “Cross-Cultural Issues of Visual Formalisms in User Interface Design.” In: Callaos, Nagib et al.(eds.) (2003). Proceedings of SCI 2003, 7th World Multiconference on Systemics, Cybernetics and Informatics. Orlando, USA, July 27-30, 2003. Vol. 1: Information Systems, Technologies and Applications, 479-c08384.