A Visual Digital Library Approach for Time-Oriented Scientific Primary Data

Jürgen Bernard[1], Jan Brase[2], Dieter Fellner[1,3], Oliver Koepler[2], Jörn Kohlhammer[1,3], Tobias Ruppert[3], Tobias Schreck[1], Irina Sens[2]

[1] Technische Universität Darmstadt, Germany
[2] German National Library of Science and Technology, Hannover, Germany
[3] Fraunhofer Institute for Computer Graphics Research, Darmstadt, Germany

Abstract. Digital Library support for textual and certain types of non-textual documents has significantly advanced over the last years. While Digital Library support implies many aspects along the whole library workflow model, interactive and visual retrieval allowing effective query formulation and result presentation are important functions. Recently, new kinds of non-textual documents which merit Digital Library support, but yet cannot be accommodated by existing Digital Library technology, have come into focus. Scientific primary data, as produced for example, by scientific experimentation, earth observation, or simulation, is such a data type. We report on a concept and first implementation of Digital Library functionality, supporting visual retrieval and exploration in a specific important class of scientific primary data, namely, time-oriented data. The approach is developed in an interdisciplinary effort by experts from the library, natural sciences, and visual analytics communities. In addition to presenting the concept and discussing relevant challenges, we present results from a first implementation of our approach as applied on a real-world scientific primary data set.

Keywords. Visual Analysis, Visual Search, Content-Based Search, Scientific Primary Data, Visual Cluster Analysis

1 Introduction

Digital Library systems are indispensable elements of an effective information infrastructure. Modern acquisition, processing, storage, and delivery technologies have improved existing and created totally new ways by which libraries can serve users. For example, Web technologies enable distributed user access; full text processing allows issuing specific, on-target queries and services may be enhanced by recommendation and personalization functionality. While much of this functionality is available in existing Digital Library systems, it is most often restricted to *textual* documents. While text is of high importance, increasingly, *non-textual* document types arise in many application areas and treating these with library services is desirable. This is quite obvious for popular non-textual document types such as digital image, video, and audio content. In these cases,

results from Multimedia Processing and Retrieval apply and can be used to realize content-based search and presentation for such content.

While ubiquitous and relevant, such multimedia document types are not the only, nor the per se most important document types. In recent discussion among research institutions and research funding agencies [1, 2], *scientific primary data* has been identified as a document type worth considering strategically. Consequently, development of infrastructure to support indexing, storage, accessing, delivery, and archival of scientific primary data is identified a necessity. Let two out of many relevant observations motivate this point. (a) *Re-usage* of scientific data is desirable to increase transparency of research and research results, and to lower the cost by sharing of data; and (b) *archival* of scientific primary data is useful for possible re-examination of that data in the future, when new analysis methods may become available. Consider *climate data* for an example, which is expensive to obtain (involving large scale and distributed observation). In the future, novel climate analysis programs may become available, where historic data can support calculation of more accurate climate models. Library support for such data clearly would benefit science and society.

For illustration purposes we describe a possible application scenario for our system. Here, a natural scientist detects an interesting *curve progression* in her collected measurements. According to her hypothesis, this exemplary time series pattern might indicate a future event that is relevant to her research. To verify the hypothesis that there is a connection between her measurements and the event, she wants to examine similar curve progressions in related data sets. A requirement for this task is a visual overview of the most similar data sets grouped by their similarity to the chosen reference example. Furthermore, measurements in the same category (e.g. global radiation) are a matter of particular interest. This is obtained by offering filtering options that operate on the meta-information appended to the data. Besides defining a search pattern by choosing a curve progression example from the existing data ("query-by-example"), a scientist wants to search for an artificial curve sketched manually ("query-by-sketch"). This can be realized in a visual-interactive graphical interface. Finally, the results of the scientist's query are displayed in the same time scale to analyze correlations between the detected time series.

Devising and implementing Digital Library support for tasks like the above mentioned is a complex challenge that involves finding solutions on many levels, ranging from acquisition to standardization over to retrieval, delivery, and archiving. In this work, we focus on the specific problem of visual retrieval and exploration in large sets of *time-oriented* scientific primary data, as an important subtype of scientific primary data in general. We present a concept devised as well as early results developed in the course of a joint research project carried out by librarians, computer scientists, and natural scientists. Our approach adapts and combines techniques from time series analysis, multimedia retrieval, and information visualization, and will be prototypically implemented and evaluated in practice. The results presented are one step towards advanced Digital Library support for this kind of data.

2 Background and Related Work

We review related work in Digital Libraries, scientific primary data initiatives, and retrieval and visualization in time series data.

2.1 Scientific Primary Data in the Digital Library Context

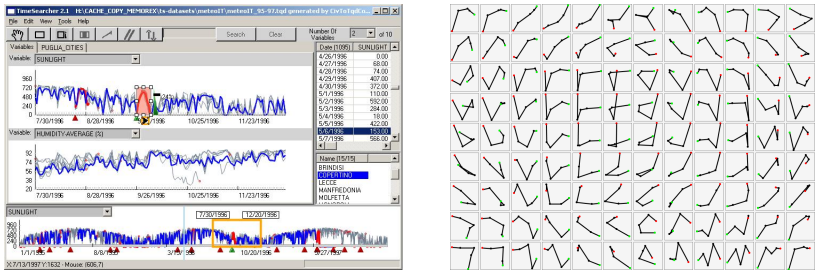
Digital Library systems have evolved over time from purely academic and pioneering works, to standardized and established systems, which are available for practical usage. Popular example systems include Fedora [3], Greenstone [4], and DLib [5]. These systems typically are oriented towards textual documents, considering non-textual documents as uninterpreted digital content for which no native system support is provided. Digital Library systems for non-textual documents which allow content-based search are relatively scarce in practice, owing to the high variability between and within collections of non-textual documents making standardization difficult. Prototypical systems exist for a number of multimedia document types, including music [6] or image and other multimedia documents [7]. These systems offer advanced support for indexing and visual retrieval of certain content. For example, the PROBADO3D system [8] supports searching in architectural model data by means of global shape and room structure, and allows for visual queries specification.

Scientific primary data may also be regarded as a non-textual document type. It often comprises numeric data or georeferenced data on continuous or discrete scales and stem from many different sources including earth observation, experimentation, or simulation. The primary data is usually also associated with textual metadata including data description, author and origin information, and even references to corresponding publications. While the necessity of treating scientific primary data by library services is generally recognized, significant challenges exist to this end including [1] but not limited to (a) persistent storage of massive volumes of data; (b) standardization of data formats and encoding; (c) quality control, peer review, and citability of data sets; and (d) clarification of legal aspects regarding ownership, access, and re-usage.

To date, a number of operational Digital Library systems for scientific primary data already exist. Examples include PsychData [9] (psychological data), PANGAEA [10] (geoscientific and environmental data), or Drayd [11] (generic data underlying natural sciences publications). Several research projects address conceptual challenges and implications in this area. The KoLaWiss initiative [2] identified organizational, technical, economic and data type-oriented challenges for establishing a collaborative scientific primary data infrastructure. Citability and publication of this data has been devised by the project "Publication and Citation of Scientific Primary Data" [12]. Establishing the European infrastructure for biological information is aimed at by the ELIXIR [13] coordination research initiative. Approaches towards service-oriented infrastructure in the Arts and Humanities are considered in the project BAMBOO [14].

2.2 Search and Visualization in Time-Oriented Data

As denoted by the example in the introduction chapter, content-based access to time series data requires the definition of similarity measures, which is important for search and visual clustering purposes. Liao [15] surveys many measures for time series similarity estimation, distinguishing three groups of time series similarity calculation approaches: raw data-based, model-based, and feature-based. Raw data-based (or transformation) approaches directly compare time series raw data, usually by measuring the cost of transforming one series to match another [16]. Model-based approaches work by calculating the degree to which two time series to be compared share the same underlying statistical model. In the feature vector (or descriptor) approach, descriptor metadata is automatically extracted from the time series data. Then, the similarity between two time series is estimated by the distance calculated between their respective descriptors. Consequently, the definition of the descriptor extraction algorithm determines the similarity concept. Examples of time series feature extractors rely e.g., on Fourier analysis [17], or on aggregation or discretization approaches [18]. Descriptor approaches usually are robust, amenable to database indexing, and simple to implement. An important conceptional distinction in time series similarity search is between global and partial search. While in global search whole time series are compared, partial search identifies similar subsequences. Techniques for partial similarity search are typically based on Sliding-Windows approaches, or on segmentation approaches such as top-down or bottom-up analysis.



(a) Time Searcher (b) Self-Organizing Trajectory Map

Fig. 1. (a) Time Searcher system [19]. (b) Self-Organizing Map computed for trajectory-oriented data [20].

Searching in time series data can effectively be supported by visual interactive query specification and result visualization. The Time Searcher System [19] (cf. Figure 1(a)) enables interactive query specification via visual filters called Timebox Widgets. These filters define ranges in the time and parameter axis. Similar time series within these ranges are found and highlighted, giving immediate feedback upon query specification. In previous work, we implemented a system for visual exploration of 2D time-dependent scatter data interpreted as

trajectory data [20]. Based on a simple geometric descriptor, the system clusters large sets of trajectory data by means of the Self-Organizing Map algorithm [21]. An early application of SOM to visualization of stock market chart data was explored in [22]. The Self-Organizing Map approach is a popular method for visual cluster analysis due to producing similarity-preserving layouts (cf. Figure 1(b) for an illustration). The SOM approach is well-suited to support visual search as a sort of *visual catalog*. Our proposed approach will rely on this algorithm (cf. also Section 4.3).

3 Library-Oriented Treatment of Scientific Primary Data

Recognizing the need for data sharing, several scientific communities have already organized data collection, archiving and access, to serve their community demands. For example, earth and environmental studies data are collected and shared on a worldwide level through the World Data Center System [23]. Data publication is an essential component of every large scientific instrument project (e.g., the CERN Large Hadron Collider). These trends induce development of new library services. DOI-based data set registration and portal-based access are two practical developments in current library support for primary data.

Data set Registration. Data set identification is a key element for citation and long term integration of data sets into text as well as supporting a variety of data management activities. To achieve the rank of a publication, a data publication needs to meet the two main criteria, *persistence* and *quality*. Quality is a rather difficult concept typically addressed by data curators building on domain-dependent guidelines and best practices. Data persistence is a rather technical problem, and addressed by the data hosting infrastructure. Technical infrastructure for data set identification is already practically provided. E.g., the German National Library of Science and Technology (TIB) developed and promotes the use of Digital Object Identifiers (DOI) for data sets. DOI names are already widely used in scientific publishing to cite journal articles. Since 2005, TIB is an official DOI registration agency with a focus on the registration of scientific primary data. In cooperation with several World Data Centers, data collected from various scientific disciplines amounting to over 700,000 data sets have been registered by TIB with DOI names as persistent identifiers.

Portal-Based Access to Remotely Stored Data. Having a DOI-based index of scientific primary data in principle allows the creation of user-friendly portal solutions to browse and access the data, based on textual metatada. An example is the *GetInfo* portal operated by TIB. It bundles access to subject databases, publishing house offerings and library catalogs with integrated full text delivery. The aim is to include all sorts of non-textual information into GetInfo. Primary research data sets are already integrated into GetInfo, and can currently be accessed by metadata queries. The concept presented in this paper is one step toward extending the access to visual and content-based methods for this data sets.

4 Approach and First Results

We describe our concept for visual retrieval in time-dependent scientific primary data and apply it to a concrete scientific data set. The described system forms the baseline for subsequent refinement of search and navigation functionality to be developed in collaboration with scientific users (cf. Section 5).

4.1 Considered Data Set

For initial development we use data from the scientific data information system PANGAEA [10] hosted by the Alfred-Wegener-Institute for Polar and Marine Research, Bremerhaven, Germany and the Center for Marine Environmental Sciences, Bremen, Germany. PANGAEA archives, publishes, and distributes geo-referenced scientific observation data. The data is organized by categories comprising observations on water (e.g., temperature, salinity, oxygen), sediment (e.g., total organic carbon (TOC)), ice (e.g., chemical composition, dust concentration), and atmosphere (e.g., temperature, humidity). Most data sets can be downloaded as text files including the measurement data and accompanying metadata. The latter covers information on citations, originating project name, spatial and temporal conditions, parameter description, etc. The raw data is provided in ASCII table format, containing time stamp and respective measurement data. We are currently considering the raw data, while we will include also metadata (if available) for filtering and query refinement. Our sample data pool consists of over 12,000 data files from the years 1981 to 2009, provided by the project BSRN [24]. The data tables have up to 100 columns corresponding to the number of time series, and a maximum of 50,000 rows, regarding to the number of measured time samples.

The data set is chosen for initial test phases, since it is enriched by a structured meta-information block (see Table 1), which is currently neglected. Future work will address the integration of content-based and meta-information search.

Meta-Information	**Description**
Citation	Data set citation (name of author, name of data set, institution, publication year, DOI-Code)
Project	Project name, link to project website
Coverage	Spatial and temporal conditions (time start and end, longitudinal and lateral coordinates, height above sea level)
Event	Description of measurement event (e.g. measurement setup)
Other version	Link to related measurements
Comment	Additional comments
Parameter	Description of parameters, unit, methodology, investigator
Size	Number of rows

Table 1. Excerpt of meta-information in PANGAEA data files.

4.2 Feature-Based Descriptor Extraction

In our baseline system, we currently support descriptor-based global similarity search in time series, based on the notion of geometric similarity of respective curves. We compute descriptors by application of a work-in-progress modular descriptor calculation pipeline described next (cf. Figure 2).

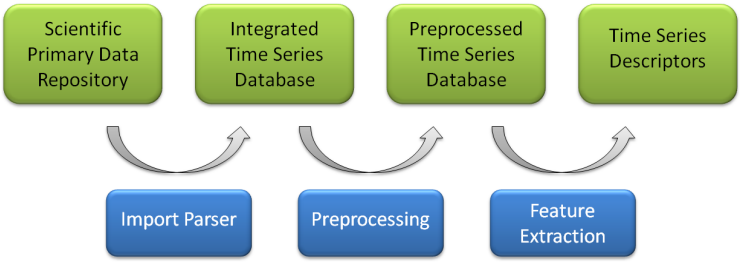


Fig. 2. Feature extraction pipeline.

The initial step reads the primary data from provided data files, or from a data repository. Currently we focus on parsing data files from the PANGAEA platform. However, importing time series data from other sources is possible by using dedicated data parsers. After data import, time series preprocessing may be applied as required by the descriptor extraction approach, the application need, and/or condition of the primary data. Several standard normalization techniques including data discretization, transformation, interpolation, and outlier and missing value treatment are implemented and can be applied prior to feature extraction. We currently consider whole time series. However, work is ongoing to implement time series segmentation to support local similarity search as well. After preprocessing, the feature extraction step can basically rely on any appropriate feature proposed so far [15]. Features based on Fourier transform or on discrete approximation have shown to be effective in the literature, and should be supported as baseline similarity functions in our system. For our first experiments, we apply a simple aggregation-based descriptor to reduce each series to a comparable, discretized representation of constant length, which will be used for subsequent clustering and retrieval steps. Considering that the implementation of the descriptor is of utmost importance for the supported similarity concept, the question arises which descriptor and preprocessing options should be chosen for a given search. This is an important research problem relating to the semantic gap, which can be tackled by user evaluation. Our goal is to let the user flexibly select the used descriptors and processing options, finding the best settings for conducting the visual search. Also, techniques based on relevance feedback (RF) are in principle applicable to mediate the semantic gap problem. Addressing interactive and visual descriptor choice is an important aspect of future work in our project.

4.3 A Visual Catalog of Time Series Data for Data Exploration

As our approach suggests an explorative content-based search, we adhere to Shneiderman's *Information Visualization Mantra* [25] ("overview first, zoom and filter, then details on demand"). To create a useful overview for thousands of time series, we propose to offer a "visual catalog" supporting effective data exploration. Two properties of such a catalog we deem useful include (1) reflectance of similarity relations between series data elements for intuitive navigation, and (2) reduction of the data cardinality while identifying the most prominent patterns in the data set. Regarding (1), the patterns should be arranged on the visual display as intuitively as possible. A global ordering of the displayed time series patterns is desirable. The more samples are presented in a sorted way, the better will be the applicator's comprehension. Regarding (2), an appropriate clustering algorithm needs to be applied, which supports (1) and is compatible to the available data descriptors.

After careful consideration, and based on good experience on other data domains, we decided to apply the Self-Organizing Map (SOM) algorithm [21], which addresses the aforementioned requirements. The algorithm is widely used in the clustering domain and has beneficial visualization properties. It is able to reduce a large data set to user-settable number of clusters that are arranged in a low-dimensional grid in an approximately topology-preserving way. For details, we refer to [21]. We apply the SOM approach on a subset of the PANGAEA content, based on our first descriptor implementation. Figure 3 gives an illustration of a SOM map showing a number of clusters of time series patterns from the data set. Applying the example from the introduction, it can be seen in Figure 3 that the natural scientist can obtain an effective overview of the curve shapes of the scientific primary data pool (left image). Furthermore she can pick an example pattern and search the data set for details, which can be displayed on demand (right image).

We consider the SOM approach in combination with an appropriate descriptor as a good candidate for a visual catalog of time series. Based on the overview provided by SOM, search interfaces and detail visualization displays can be implemented to support drill-down by the user.

4.4 Visual Query Specification

The visual time series catalog gives an overview of the whole data library. Based on this, content-based user queries may be executed via visual query specification. We initially support two search modalities. A query curve can be specified either by selection from the visual catalog, or by drawing a curve sketch, as described in the use case in Section 1. Based on the time series descriptors, distances are calculated, a ranking is obtained, and the results are highlighted by color coding on the catalog itself, or displayed in a separate list view. Figure 4 illustrates.

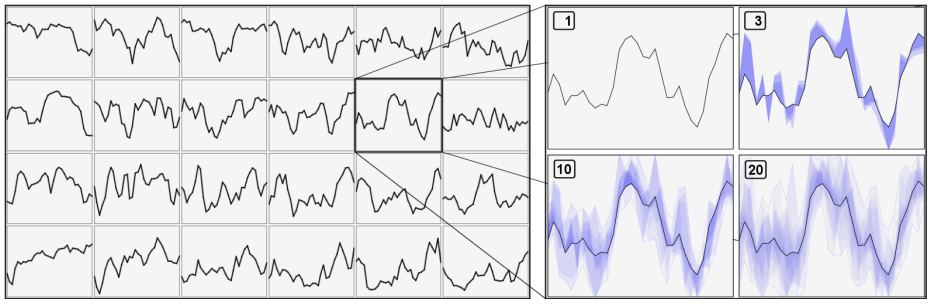


Fig. 3. Left: Visual time series catalog, provided by SOM clustering. Each cell shows one data cluster by a representative time series. Right: A detail view of a selected cluster is shown by an opacity-based overlaying view.

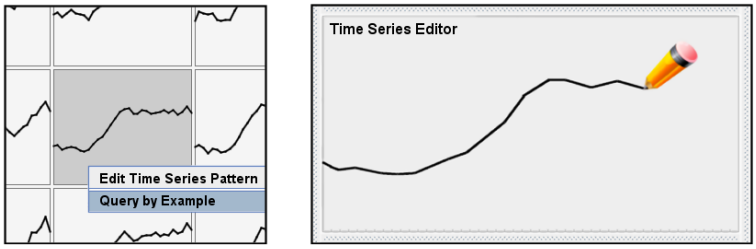
4.5 Metadata and Export to User Tools

As already indicated, scientific primary data sets are often enriched by meta-information regarding author, originating project, measurement specifics and so on. Of course, such information (if available) must not be neglected in the visual search. We currently support a light-weight approach to include metadata search. Specifically, uninterpreted full text search in the metadata fields is provided. We point out that metadata integration over heterogeneous data sources is a difficult and expensive process. As we aim to search over heterogeneous data sources, this is a pragmatic approach. In our implementation, a simple text input field enables the user to search in the meta-information of the data sets and filter meaningful time series plots. For example, if the user only wants to consider measurements of a certain researcher she is able to specify her search by typing the researcher's name in a projected meta-information search window. As a result, the data sets authored by the special researcher will be highlighted in the visual catalog (cf. Figure 5(a) for an example).

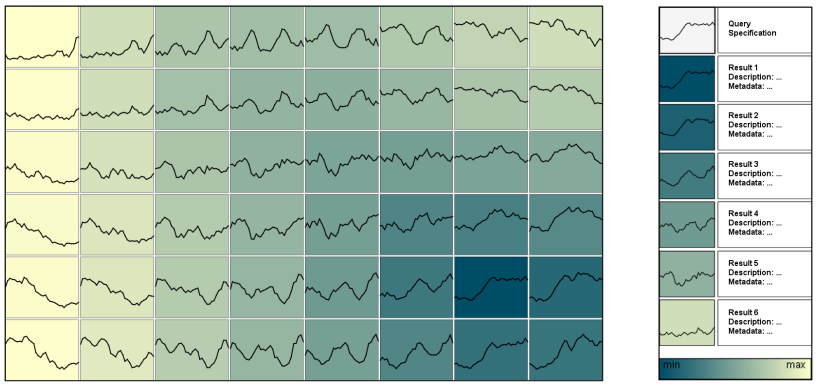
As our system is intended to support the data-oriented scientific research process, it is important to support domain-dependent tools for export of search results. As a starting point, export of found time series to PanPlot [26], which visualizes small amounts of time series in publication-ready quality, is possible (cf. Figure 5(b) for an example).

5 Discussion and Next Steps

Our first step towards visual search in a Digital Library system for time-oriented data is based on the concepts of visual catalog and on content-based queries. Our implemented descriptor supports the similarity notion of global curve shape and is only a starting point. Technically, a wealth of further functionality to explore exists, including design of additional curve shape descriptors, partial similarity, and time- and scale invariant search modalities. We recognize that for the prototype to be successful, it needs to solve real user problems and therefore, further

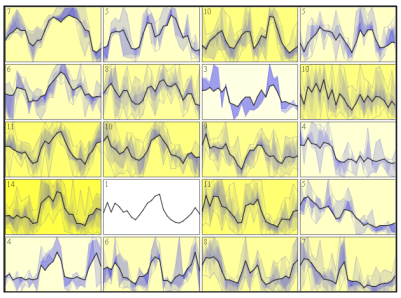


(a) Query modalities

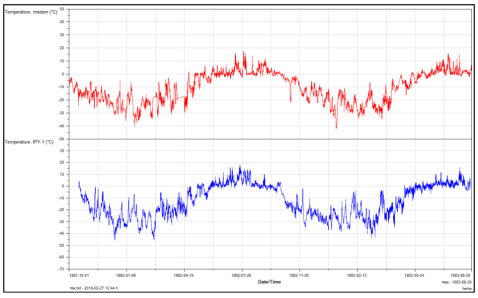


(b) Result visualization

Fig. 4. (a) Query-by-example based on selection and sketching. (b) Result visualization based on the catalog and list.



(a) Keyword frequency highlighting



(b) Search result export

Fig. 5. (a) Highlighting the frequency of occurring keywords from a metadata search. (b) shows a time series search result exported to a specialized analysis tool (PanPlot).

development will take place in close collaboration with scientific users. During an evaluation workshop, we will demonstrate the currently existing prototype to scientists, expecting that relevant use-cases will be defined that can in another iteration be supported in the prototype.

We expect that the most useful search functionalities will not consist of only a single modality (e.g., curve shape), but rather a combination thereof. Additional modalities may involve correlation-based comparison of time series at different scales and possibly applying on a partial level. We further expect that metadata will play an important role, either for filtering of search results or as input to adaptive search algorithms. Conceptually, we are interested in more closely combining browsing and searching. Tight coupling of browsing and searching is expected to yield effective search results. Also, implications regarding scientific data infrastructure are given. For our methods to be broadly applicable, our system needs to interface with many data providers, raising the question of interoperability.

6 Conclusions

We introduced the problem of Visual Digital Library support for scientific primary data. We argued that this data is requiring library support, and that a user-interface based on visual search is desirable. Specifically, content-based visual search should complement purely metadata based search to be effective. A design and development methodology based on visual cataloging and content-based searching in time-oriented data was presented. A first implementation was applied on real data. Options for future work and a user-in-the-loop development model were presented.

Acknowledgments

Rainer Sieger and Hannes Grobe of the Alfred Wegener Institute kindly provided data and initial expert feedback. Tatiana von Landesberger and Sebastian Bremm of Interactive-Graphics Systems Group at TU Darmstadt provided helpful discussion and suggestions. This work was supported by a grant from the Leibniz Association as part of the "Joint Initiative for Research and Innovation" program.

References

1. German Research Foundation (DFG): Report on round table meeting of research data (in German). Whitepaper (2008) http://www.dfg.de/download/pdf/foerderung/programme/lis/forschungsprimaerdaten_0108.pdf.
2. Society for Scientific Data Processing Goettingen: Cooperative long-term preservation for research centers (in German). Project Report (2009)
3. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. Int. J. Digit. Libr. **6**(2) (2006) 124–138
4. Witten, I.H., Mcnab, R.J., Boddie, S.J., Bainbridge, D.: Greenstone: A comprehensive open-source digital library software system. In: Proceedings of the Fifth ACM International Conference on Digital Libraries. (2000)

5. Castelli, D., Pagano, P.: Opendlib: A digital library service system. In: ECDL. (2002) 292–308
6. Dunn, J.W., Mayer, C.A.: Variations: a digital music library system at indiana university. In: DL '99: Proceedings of the fourth ACM conference on Digital libraries, New York, NY, USA, ACM (1999) 12–19
7. Agosti, M., Berretti, S., Brettlecker, G., Bimbo, A.D., Ferro, N., Fuhr, N., Keim, D.A., Klas, C.P., Lidy, T., Milano, D., Norrie, M.C., Ranaldi, P., Rauber, A., Schek, H.J., Schreck, T., Schuldt, H., Signer, B., Springmann, M.: Delosdlms - the integrated delos digital library management system. In: DELOS Conference. (2007) 36–45
8. Berndt, R., Blmel, I., Krottmaier, H., Wessel, R., Schreck, T.: Demonstration of user interfaces for querying in 3d architectural content in PROBADO3D. In: 13th European Conference on Digital Libraries. (2009) Demonstration Paper.
9. PsychData National Repository for Psychological Research Data. (http://psychdata.zpid.de/ (in German))
10. PANGAEA Publishing Network for Geoscientific & Environmental Data. (http://www.pangaea.de/)
11. Dryad Digital Repository for Data Underlying Published Works. (http://www.datadryad.org/)
12. Brase, J.: Using digital library techniques-Registration of scientific primary data. Lecture Notes in Computer Science (2004) 488–494
13. ELIXIR European Life Sciences Infrastructure for Biological Information. (http://www.elixir-europe.org/)
14. Bamboo Research Initiative. (http://projectbamboo.org/)
15. Liao, T.W.: Clustering of time series data-a survey. Pattern Recognition **38** (2005) 1857–1874
16. Agrawal, R., Lin, K., Sawhney, H., Shim, K.: Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In: Proceedings of the International Conference on Very Large Data Bases, Citeseer (1995) 490–501
17. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. Lecture Notes in Computer Science (1993) 69–69
18. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. (In: Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery)
19. Hochheiser, H., Shneiderman, B.: Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. Information Visualization **3**(1) (2004) 1–18
20. Schreck, T., Bernard, J., Von Landesberger, T., Kohlhammer, J.: Visual cluster analysis of trajectory data with interactive kohonen maps. Information Visualization **8**(1) (2009) 14–29
21. Kohonen, T.: Self-Organizing Maps. 3rd edn. Springer (2001)
22. Šimunić, K.: Visualization of stock market charts. In: Proc. Int. Conference in Central Europe on Computer Graphics, Visualization and Computer Vision. (2003)
23. World Data Center System. (http://www.ngdc.noaa.gov/wdc/)
24. Baseline Surface Radiation Network (BSRN). (http://www.bsrn.awi.de/)
25. Ben, S.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proc. Of the 1996 IEEE Symposium on Visual Languages, IEEE Computer Society, Washington, DC. (1996) 336–343
26. PANGAEA PanPlot Tool. (http://doi.pangaea.de/10.1594/PANGAEA.330147)