Erschienen in: International Journal on Digital Libraries ; 11 (2010), 2. - S. 111-123

A visual digital library approach for time-oriented scientific primary data

Jürgen Bernard · Jan Brase · Dieter Fellner · Oliver Koepler · Jörn Kohlhammer · Tobias Ruppert · Tobias Schreck · Irina Sens

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 fully accommodated by existing Digital Library technology, have come into focus. Scientific data, as produced for example, by scientific experimentation, simulation or observation, is such a document type. In this article we report on a concept and first implementation of Digital Library functionality for supporting visual retrieval and exploration in a specific important class of scientific primary data, namely, time-oriented research 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 to discussing relevant challenges, we present results from a first implementation of our approach as applied on a real-world scientific primary data set. We also report from initial user feedback obtained during discussions with domain experts from the earth observation sciences, indicating the usefulness of our approach.

Keywords Visual search · Content-based retrieval · Time series · Scientific research data · Visual cluster analysis

This paper is a substantially revised and extended version of a paper with the same title originally appeared in the Proceedings of the 14th European Conference on Digital Libraries (ECDL 2010).

J. Bernard (✉)
Technische Universität Darmstadt, Darmstadt, Germany
e-mail: juergen.bernard@gris.informatik.tu-darmstadt.de

J. Brase · O. Koepler · I. Sens
German National Library of Science and Technology, Hannover, Germany
e-mail: jan.brase@tib.uni-hannover.de

O. Koepler
e-mail: oliver.koepler@tib.uni-hannover.de

I. Sens
e-mail: irina.sens@tib.uni-hannover.de

D. Fellner · J. Kohlhammer · T. Ruppert
Fraunhofer IGD, Darmstadt, Germany
e-mail: dieter.fellner@igd.fraunhofer.de

J. Kohlhammer
e-mail: joern.kohlhammer@igd.fraunhofer.de

T. Ruppert
e-mail: tobias.ruppert@igd.fraunhofer.de

T. Schreck
University of Konstanz, Konstanz, Germany
e-mail: tobias.schreck@uni-konstanz.de

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 mostly restricted to *textual* documents. While text is of high importance, increasingly, *non-textual* document types arise in many application areas and treating these with advanced digital 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 these results 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 [16,29], *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 of research 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 that will provide new insight about historic data. Consider *climate data* for an example, which is expensive to obtain, as it typically involves large scale and distributed observation facilities. 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 a devised visual Digital Library system for research data. 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 identified time series of interest.

Devising and implementing Digital Library support for tasks like these examples is a complex challenge that involves finding solutions on many technical and organizational 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 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 it is prototypically implemented, as a basis for future evaluation with domain experts. 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, as well as 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 [20], Greenstone [32], and DLib [10]. 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. This can be attributed to high variability between and within collections of non-textual documents, which is typically observed in practice and which makes standardization difficult. Prototypical systems exist for a number of multimedia document types, including music [14] or image and other multimedia documents [1]. These systems offer advanced support for indexing and visual retrieval of certain content. For example, the PROBADO system [8] supports searching in digital music and 3D architectural model data by means of content-based search, allowing for visual query specification.

Scientific primary data may also be regarded as a non-textual document type. It often comprises numeric data on continuous or discrete scales (e.g., time and space). They can stem from a variety of different origins, such as observation, experimentation, or simulation. The primary data is usually also associated with textual metadata including data description, author and origin information, or references to corresponding publications. In case of the earth observation data discussed later in our case study in Sect. 5, georeferences

(coverage) are an important metadata aspect. While the necessity of treating scientific primary data by library services is generally recognized, significant challenges exist to this end including [16] 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 [24], which is a psychological research data repository, and Dryad [13], which is a repository for generic data underlying publications in the natural sciences. The degree of *harmonization* among the contained data sets, offered by the respective repository, is decisive for implementation of advanced search and access methods. Simple approaches may treat individual data files as uninterpreted data containers. More advanced repositories may provide harmonized data, allowing, e.g., to implement global search algorithms over different data sets, owing to normalized data. The Publishing Network for Geoscientific and Environmental Data (PANGAEA, [23]) is an encompassing digital library for earth observation research data. It provides carefully curated and harmonized primary and textual metadata, which is persistently identified and freely available online. Full text search is supported in the respective textual metadata records. Currently, in PANGAEA there is no search functionality available for content-based search in the measurement data itself, apart from searching for raw numbers. Development of content-based search in time-oriented measurement data is the precise goal of this work. PANGAEA, owing to its highly harmonized data sets, is an ideal candidate for our research purposes to this end. In Sect. 5, we will come back to PANGAEA, where a subset of this repository will be used in a case study. But as we will discuss deeper in Sect. 6 the challenge of visual search in research data is not only a technical one. While users are familiar with the usability of different types of digital libraries in the context of text document, we have recognized that the approach of a content based visual search is an innovative concept not considered by a major number of scientists yet.

To date, several research projects address conceptual challenges and implications in digital library support for research data. The KoLaWiss initiative [29] 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" [9]. Establishing the European infrastructure for biological information is aimed at by the ELIXIR [15] coordination research initiative. Approaches towards service-oriented infrastructure in the Arts and Humanities are considered in the project BAMBOO [6].

2.2 Similarity search in time series

As denoted by the example in Sect. 1, content-based access to time series data requires the definition of similarity measures, which is important for search and visual clustering purposes. Liao [21] 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 [3]. 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 [2], on Discrete Wavelet Transform [11], or on aggregation or discretization approaches [18] [22]. Descriptor approaches usually are robust, amenable to database indexing, and simple to implement. Depending on the chosen descriptor approach, several data preprocessing steps are performed, like outlier replacement, data normalization, or time series quantization. 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.

2.3 Visualization and query specification

Visualization of time-oriented data is important in many application fields. Depending on the analysis task, a variety of visualization techniques have been introduced to date, with a survey presented in [5]. Van Wijk et al. [30] propose visual time series clustering using a calendar-based approach. Searching in time series data can effectively be supported by visual interactive query specification and result visualization. Shneiderman uses so-called Dynamic Query Filters [4] to reduce the number of data elements that are shown on the display. The user can customize the query specification with interactive control elements like buttons and sliders. The Time Searcher system [17] 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.

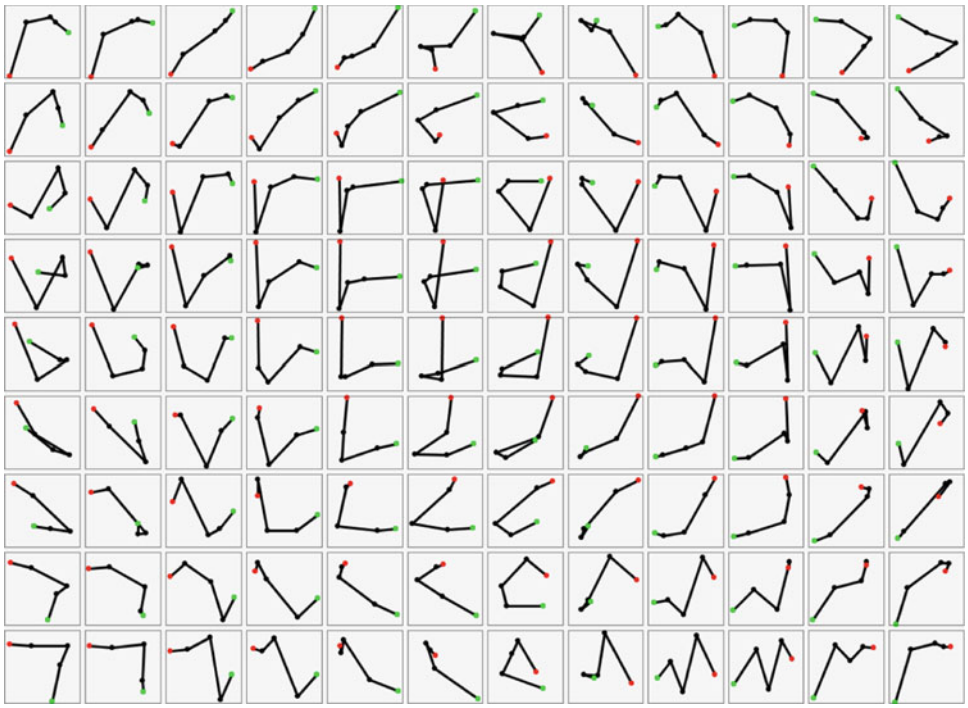

Fig. 1 Self-Organizing Map computed for trajectory-oriented data [25]

Wattenberg [31] introduced the QuerySketch tool that enables the user to interactively draw time series shapes and search for similar patterns in a database. With WireVis, Chang et al. [12] introduce a visualization tool for analyzing financial transactions. There, a search by example technique to search for clusters similar to a selected example cluster is introduced. The explorative analysis of large financial time series data sets is considered in [34]. There, the authors compare and cluster time series by means of perceptual interest points detected in the raw data.

In previous work, we implemented a system for visual exploration of 2D time-dependent scatter data interpreted as trajectory data [25]. Based on a simple geometric descriptor, the system clusters large sets of trajectory data by means of the Self-Organizing Map (SOM) algorithm [19] (cf. Fig. 1). An early application of SOM method to visualization of stock market chart data was explored in [28]. The SOM approach is a popular method for visual cluster analysis due to producing similarity-preserving layouts. 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 Sect. 4.2.1).

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 [33]. 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.

3.1 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 two key 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. For example, the German National Library of Science and Technology (TIB) develops 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 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.

3.2 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 article is one step toward extending the access with respect to visual and content-based methods for these data sets.

4 Approach

In this section, we describe our concept for visual retrieval in time-dependent scientific primary data. In Sect. 5 we will apply this concept to a selected research 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. Sect. 6).

4.1 Feature-based decriptor extraction of time series data

As an initial step of the feature extraction pipeline (see Fig. 2), the primary data is read from provided data files, or from a data repository. We are currently developing a generic data structure in our system, which enables the import of heterogeneous sequence-based data formats. For case studies, we focus on parsing data files from the PANGAEA platform. However, importing time series data from other sources is in principle 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 also implemented a baseline time series segmentation techniques for supporting local similarity search. After preprocessing, the feature extraction step can take place. To this end, we can rely on a variety of time series descriptors explored in the literature [21]. 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 have applied 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. We also integrated further feature extraction approaches, including Discrete Fourier Transformation (DFT), Discrete Haar Wavelet Transformation (DWT), Piecewise Aggregate Approximation (PAA), and Symbolic Time Series Representation (SAX).

Our approach requires accessing the data of each data repository to be included in the search at least once. Each data file is read, preprocessing is done, and descriptors are extracted and stored in our index. Note that typically, descriptor indexes are much smaller than the represented original data. Therefore, we expect to be able to consolidate the indexes of different repositories in a central database. We do not aim at providing a cache for the primary data. Consequentially, detail-on-demand views, e.g., for accessing original and metadata for individual search results, require another access to the original data in their respective repository.

The final step in our preprocessing pipeline concerns the definition of a metric to calculate the similarity between two time series descriptors. We currently focus on the Euclidean distance, but in the future we will expand our repertoire by robust similarity measures such as the Earth Mover Distance and Dynamic Time Warping. Considering that the implementation of the descriptor and the similarity metric is of high importance for the supported similarity concept, the question arises which descriptor and preprocessing options to chose for a given search. This is an important research problem relating to the semantic gap, which we plan to address by user evaluation. Our goal is to let the user flexibly select the used descriptors and processing options, finding the best

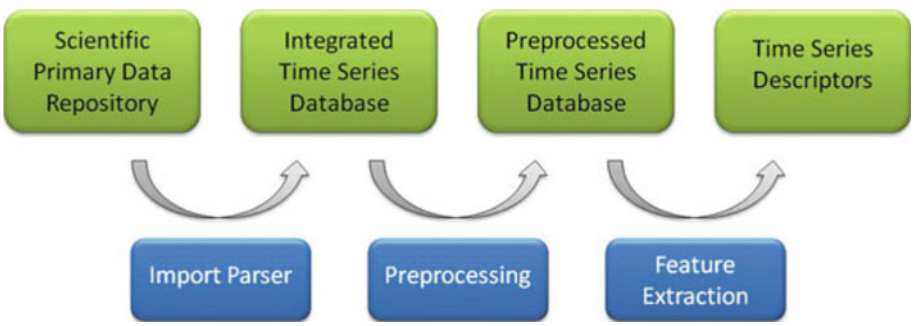

Fig. 2 Feature extraction pipeline

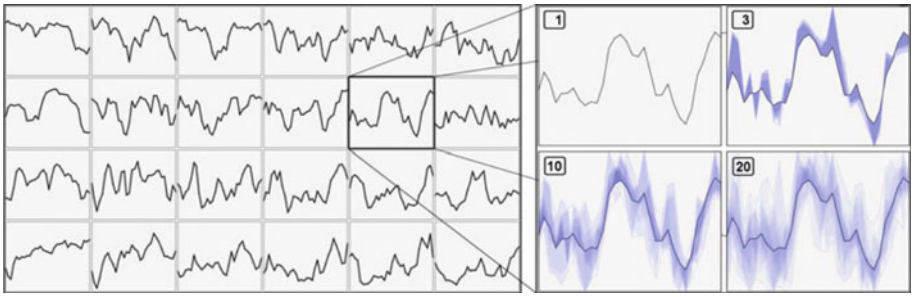

Fig. 3 *Left* Visual time series catalog, provided by SOM clustering. Each cell shows one data cluster by a representative time series. *Right* A detailed view of a selected cluster is shown by an opacity-based overlaying view

settings for conducting the visual search. Also, techniques based on relevance feedback 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.

4.2 Visual search and exploration in time series data

In the following we describe the user interface that enables scientists to interactively explore and search in the time series data content. The search itself consists of two major components: (a) a visual catalog of time series data for data exploration and (b) a visual query specification editor for defining content-based queries on the time series data. Both components can be used iteratively. The described system forms the baseline for subsequent refinement of search and navigation functionality to be developed in collaboration with scientific users (cf. Sect. 6).

4.2.1 Visual catalog of time series data

As our approach suggests an explorative content-based search, we adhere to Shneiderman's *Information Visualization Mantra* [26] ("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 time 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 in the visual display as intuitively as possible. A global ordering of the displayed time series patterns is desirable. Regarding (2), an appropriate clustering algorithm needs to be applied, which supports (1) and is compatible to the available data descriptors.

After consideration, and based on good experience on other data domains, we decided to apply the SOM algorithm [19], which addresses the aforementioned requirements. The algorithm is widely used in the explorative data analysis domain and is very suited as a basis for overviewing displays. The method is able to reduce a large data set to a user-settable number of clusters that are arranged in a low-dimensional grid in an approximately topology-preserving way. For algorithmic details, we refer to [19]. As an example, we apply the SOM approach on a subset of the PANGAEA content, described by PAA features. Figure 3 shows a SOM display representing a number of clusters of time series curves given in the data set. Applying the example from the introduction, it can be seen in Fig. 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 an appropriate method for a time series visual catalog display. Based on the overview provided by SOM, search interfaces and detail visualization displays are implemented to support drill-down by the user. With this approach, three main functionalities are realized. Firstly, it gives a global overview of the data repository by considering every data set in the database. Secondly, it provides example patterns for the visual query specification editor for the definition of the search query (see Sect. 4.2.2). And finally, it is used as a possible result visualization only considering a subset of data sets, that are retrieved by the query specification.

4.2.2 Visual query specification editor

As a possible first step of the visual search process in time-dependent scientific primary data, the user can get an overview of the whole data repository with the visual catalog described in the previous section. To get content-based access to this data, the user must be enabled to specify a content-based query on the data repository. This is realized via a visual query specification editor (see Fig. 4). This editor, in our design, consists of four components. The filter panel, the sketch panel, the example pattern panel, and the zoom panel. In the filter panel, the search space can be restricted by defining constraints on the meta-information (e.g., by only considering a special physical unit, or a specific author or

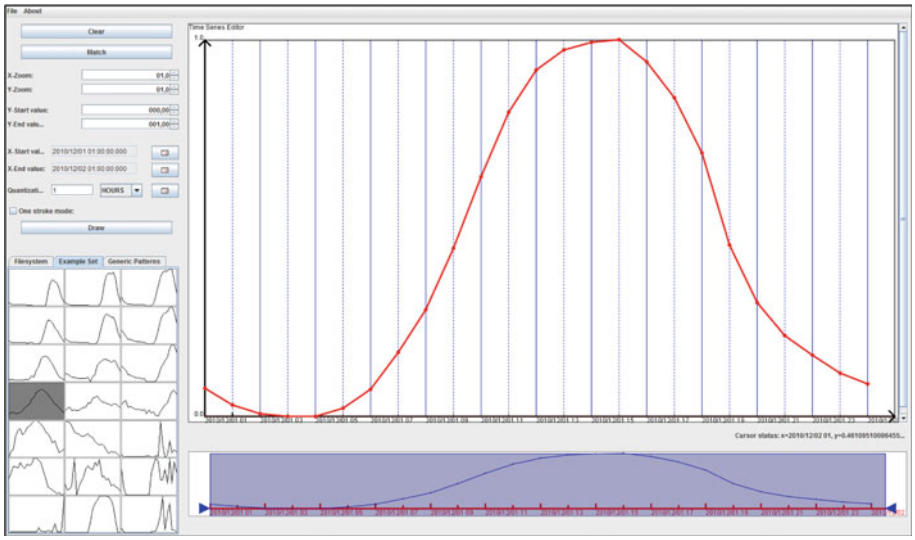

Fig. 4 Visual query specification editor consisting of filter panel (*upper left*), sketch panel (*central*), example pattern panel (*lower left*), and zoom panel (*lower right*)

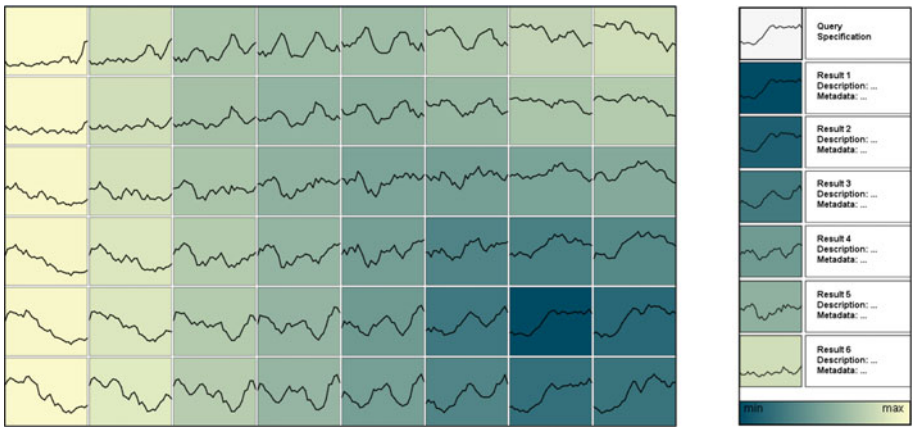

Fig. 5 Result visualization based on the visual catalog (*left*) and list (*right*). Color coding is used to show the distance of each curve to the query specification (*dark color values* denote low distances and high similarity, accordingly)

project name). Also, time intervals and time quantization can be specified.

In the sketch panel the query curve is defined. Based on the time series descriptors (see Sect. 4.1), distances to this curve are calculated and a search ranking is obtained defining the query results. There exist several possibilities to define the query curve. The simplest way is to draw the sketch with the pointing device (mouse), realizing the principle of query-by-sketch. This is an intuitive way for researchers that have a special curve progression in mind. A second possibility is to add example patterns to the sketch panel, enabling query-by-example. This is realized via the example pattern panel. With this panel the provided patterns can be chosen and inserted in the sketch panel by drag-and-drop. There exist three ways of choosing such an example pattern. Either the user can specify an example pattern via a specific data set searchable via its DOI and attribute description, or she can choose one of a set of prototypical functional patterns provided (e.g., sine, linear, etc.) Finally, also a pattern taken from the visual catalogue can be selected. These patterns are inserted in the sketch panel in whole, or in a user-specified region of the curve. After adding the example pattern, the user can refine the shape of the given curve with the cursor again and so on. To support the curvature definition process an additional zoom panel is provided, enabling the user to zoom into special regions and refine the curve.

Once the query definition process is completed, the query can be executed. The results of the query are visualized as a color coding overlay in the visual catalog. Also, a SOM only of the retrieved search results can be produced. As a default, also a sorted list view is implemented (see Fig. 5).

4.2.3 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. As a first concept, uninterpreted

Table 1 Excerpt of meta-information in PANGAEA data files

Field name	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

full text search in the metadata fields was provided. In recent work, we prototypically extended the metadata search functionality to the most important PANGAEA-specific metadata tags, as listed in Table 1. 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, full text search in uninterpreted field data is a pragmatic approach. In our implementation, simple text input fields enable the user to search in the meta-information of the data sets and filter meaningful time series. 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 metadata search field. As a result, the data sets authored by the special researcher are highlighted in the visual catalog (cf. Fig. 6 for an example).

Since 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 [27], which visualizes given time series in publication-ready quality, is possible (cf. Fig. 7 for an example).

5 Case study

In this section, we demonstrate the applicability of our system for explorative and content-based search in time-oriented research data repositories. We will detail the considered experimental data set, and present two plausible use case scenarios. They illustrate how our approach can help to get an overview over a previously unknown data repository, allow exploration of interesting patterns including searching, and support the analysis process in general.

5.1 Considered data set and case study scenario

As a starting point, we consider research data from the scientific data repository PANGAEA [23] operated by the Alfred-Wegener Institute for Polar and Marine Research in Bremerhaven, and the Center for Marine Environmental Sciences in Bremen. PANGAEA archives, publishes, and distributes georeferenced scientific earth observation data (cf. also Sect. 2.1). Data in PANGAEA comprises observations from four main areas of study, namely water (e.g., temperature, salinity,

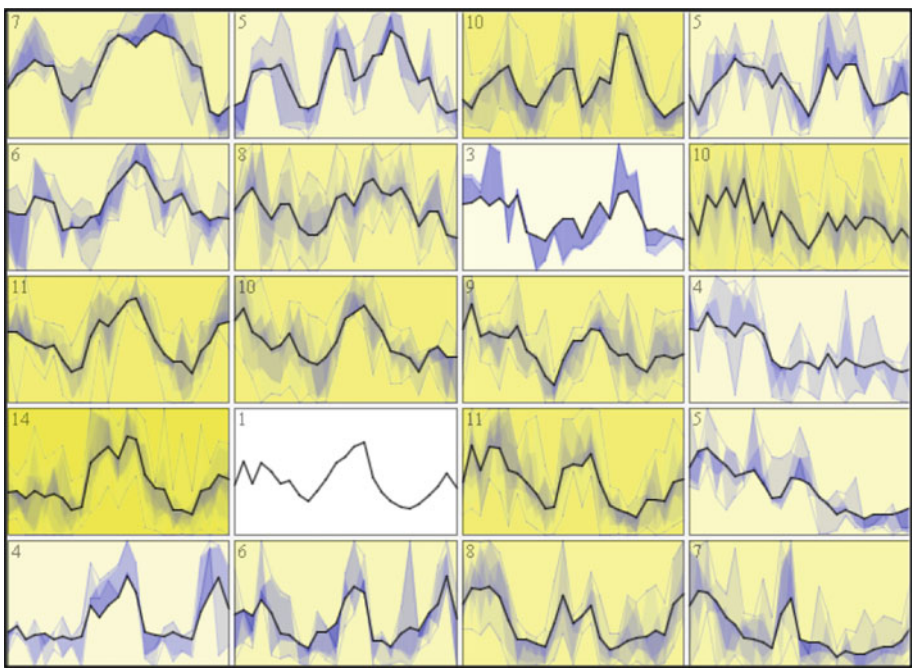

Fig. 6 Highlighting the frequency of occurring keywords from a metadata search

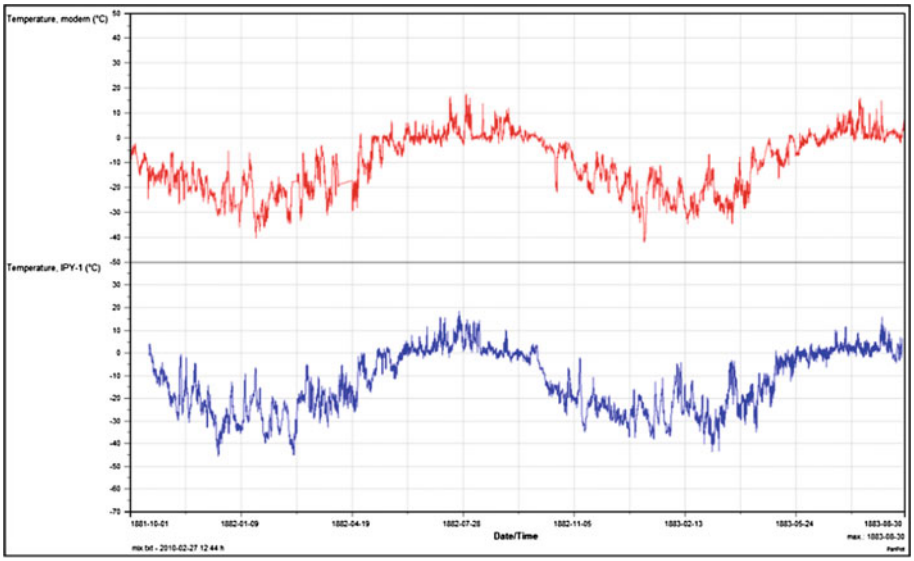

Fig. 7 Time series search result exported to a specialized analysis tool (PanPlot)

oxygen), sediment (e.g., total organic carbon (TOC)), ice (e.g., chemical composition, dust concentration), and atmosphere (e.g., temperature, humidity). PANGAEA supports data export, e.g., for post-processing and analysis purposes, in form of plain text files. The export covers metadata information on citations, originating project name, spatial and temporal conditions, parameter description, etc. The export also covers raw data in ASCII table format, containing discrete measurements at respective points in time, thereby, forming time series data. The rich available metadata can also be considered for filtering purposes, for combined content-based and metadata search, and for detail-on-demand views in general. Our sample data pool consists of a PANGAEA subset of 14,331 data files from the years 1992 to 2009, as provided by the Baseline Surface Radiation Network (BSRN, [7]) PANGAEA compartment. The data tables have up to 100 columns (time series), and up tot 50,000 rows (number of observations per series). Data provided by BSRN is dominated by measurements of radiation (short-wave, long-wave, diffuse, direct), temperature, humidity and wind (speed, direction). Each data file contains time series content over the period of 1 month, thus, longer measurements are monthly separated, following a granularity as defined by the data providers.

We choose to experiment with BSRN data, for it is (a) relevant to a large research community, (b) provides excellent quality standards in terms of raw data and metadata, and (c) features a substantial amount of data. Therefore, it is well suited for our development efforts. Specifically, presence of metadata allows us to test for combined content-based and metadata-based search and result set visualization functionality. Currently, our focus is on developing content-based search and exploration facilities. We also do recognize that metadata is important for filtering and refinement, and also, leveraging found results in subsequent scientific analysis of the accessed data. As an overview, Table 1 lists the key metadata fields given in the BSRN data set.

Based on this input data set, we discuss an example scenario, illustrating how a researcher can use our visual search and exploration system in leveraging the data from the repository. We assume the researcher is interested in daily patterns of radiation observation series data. For our experiments, we import approximately 2,500 radiation observation time series into our system by sampling from the BSRN repository. Our sampling strategy is based on randomly picking individual data elements to get a subset with mixed measurement types. Preprocessing of the data includes replacing missing values by interpolation, and aligning the time series to a global synchronized time scale. We segment all time series into 24 h intervals, spanning whole days. Segmentation leads to more than 70,000 daily time series observations. We used the Piecewise Aggregate Approximation (PAA) descriptor [22] to get equidistant time series representations. We assume a resolution of one value per hour as appropriate. Therefore, the PAA descriptor provides a feature vector with 24 bins for each time series. Furthermore, we choose a local min-max normalization procedure, as we are interested in the overall shape of the time series patterns, occurring in the considered data set. As a final preprocessing step, we use the SOM algorithm for visual cluster analysis (see Sect. 4.2.1), to obtain a visual catalog as a starting point.

5.2 Case study 1: explorative global search

Figure 8 shows the visual catalog of all daily time series curves that appear in our considered data set, calculated with

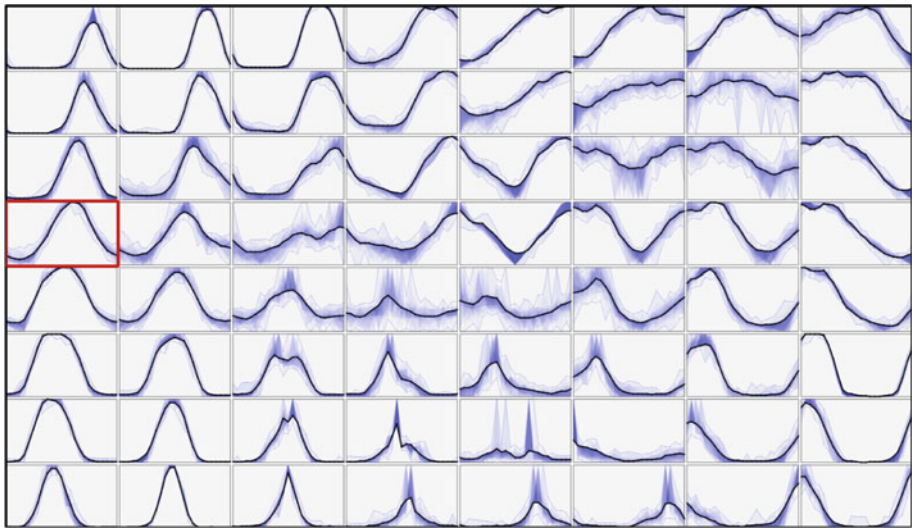

Fig. 8 Visual catalog of more than 70,000 daily time series, sampled from the BSRN data set. The patterns represent the results of a variety of measurements at different locations, times, and physical units

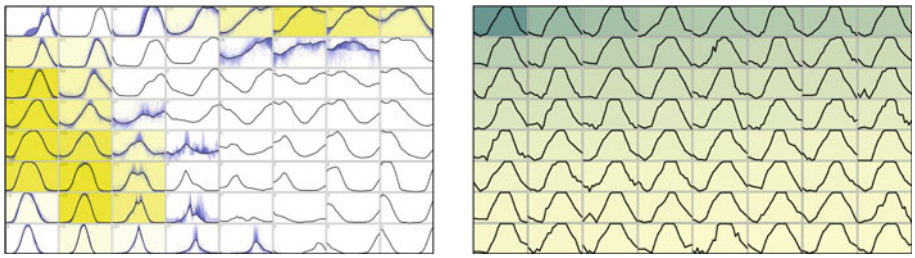

Fig. 9 Visualization of the visual search specified according to Fig. 4. *Left* Highlighting the most similar prototypes in the global visual catalog. *Right* Linear list view, showing the most similar curves from the data set. Note that color coding is used to redundantly indicate the similarity scores, that lead to the ranking

the SOM visual cluster algorithm. From this catalog, the analyst can make some first observations. Prominent patterns in the catalog include (1) a peak near noontime (approximately located in the middle of the diagrams), (2) a peak at midnight (left and right hand side of the diagrams), (3) constant behavior, (4) linear increasing and decreasing behaviors, and (5) oscillating behavior. The data elements best fitting to the SOM cells are visualized with blue opacity bands, indicating the spread of matched curves around representative pattern, in each cell of the visual catalog.

Having a global overview and keeping in mind that the visual catalog shows patterns of the complete data set implying shapes with different physical units, we want to deeper explore the content of some interesting clusters. Considering the previously identified five patterns, the visual catalog offers a series of interesting clusters to explore. For example, the red marked pattern in the visual catalog in Fig. 8 can be seen interesting, as it may resemble very prototypical temperature profiles. It starts with a low value, rises to a maximum in the afternoon, and then decreases in the evening. Inspired by this finding, the analyst decides to explore more occurrences of this behavior. Therefore, the visual time series query editor (see Sect. 4.2.2) is used to formulate an appropriate visual query. First, the curve prototype of the visual catalog is imported into the query editor as a reference example. After that, this shape is manually refined to correspond exactly to our hypothesis about a possible temperature progression. Figure 4 illustrates the visual query.

In Fig. 9, we visualize the results of the execution of this visual query. The left image shows via highlighting the best matching prototype curves from the visual catalog. We can also take a look using a list-based view of the curves best matching our sketched query (right hand side in Fig. 9). The background color used in the image denotes, redundantly to position, the decreasing similarity of the sorted list elements. Dark color values mean a strong similarity, while bright color values denote decreasing similarity to our sketch, drawn in the time series editor. Further investigation of the metadata of the found curves confirms the assumption of the analysis. Namely, the explored pattern group contains large quantities of temperature observations. Additionally, many radiation measurements are included. This finding might warrant further analysis of the relationships between temperature and radiation levels. Inspired by this visually obtained assumptions, the analysts might want to explore, possibly with other analysis tools, the detailed relations of the found data sets.

5.3 Case study 2: domain-specific search

The previous example demonstrated searching in all data sets and parameters simultaneously, which is useful for assumption-free data exploration, and for becoming acquainted with

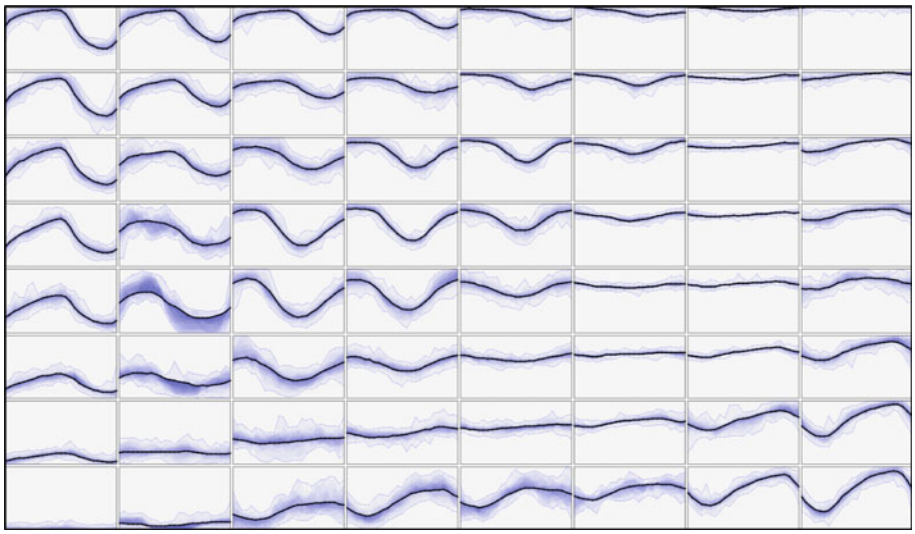

Fig. 10 Visual catalog of a restriction of our BSRN sample to series of humidity observations

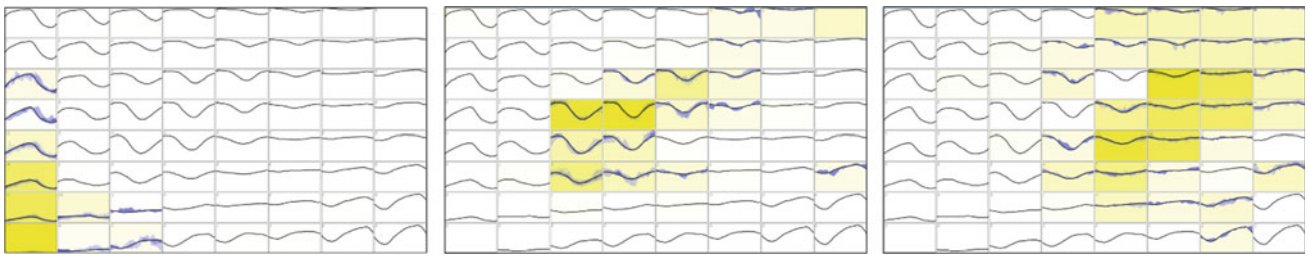

Fig. 11 Humidity pattern distribution highlighting of the radiation stations in Desert Rock (*left*), Payerne (*center*) and Spitsbergen (*right*). It can be seen that these three areas on earth with their different humidity properties are well separated in the visual catalog

new, previously unseen data sets. The search can also be restricted to specific data sets or parameter domains of interest. Specifically, the physical unit of a time series can be regarded as an important filter criterion. Here, our hypothetical analyst chooses the physical unit "humidity, relative" as a filter criterion. Not restricting other parameters, specifically also not the geo-location, the explorative search can be interesting in context of climate observations from all over the world.

First, we calculate a new visual catalog based only on observations of humidity (see Fig. 10). In this scenario, we choose a global min-max normalization, as we are thereby able to discover progressions described by shape and value simultaneously. For example, patterns with nearly constant curve progressions can be found with different value offsets, ranging from minimal to maximal humidity (lower-left and upper-right areas in the catalog in Fig. 10). A detailed exploration of the constant low patterns (lower left in the visual catalog), reveals that these measurements are provided by five predominant observation stations, located in rather dry regions (e.g., Desert Rock, Nevada, and Boulder, Colorado USA; Solar Village, Saudi Arabia; Ilorin, Nigeria; Izana, Teneriffe Spain). At this point, the analyst draws the hypothesis that these areas exhibit similar climate behavior. Further investigations, facing (possibly hot) temperatures can be conducted by the specialist to draw additional hypotheses. The upper right pattern in our visual catalog in Fig. 10 shows contrasting behavior. Measurements of this cluster occur in Spitsbergen, Norway; Goodwin Creek, Mississippi, USA; and Bondville, Illinois, USA, and stand out for high constant humidity.

The analyst finds a further interesting pattern in the center of the visual catalog in Fig. 10. There, the curve shapes appear with much higher variability than in the lower-left and upper-right areas described before. What stands out, is a decrease around noontime in these patterns. An exploration of this cluster brings the finding that it is strongly dominated by measurements taken in central Europe (Carpentras, France; Lindenberg, Germany; Payerne; Switzerland) and again gives causes for hypotheses and further investigations to domain specialists. The found taxonomy of these humidity observations can be compared in more detail by contrasting highlighted views of the catalog. Figure 11 compares the occurrences of measurements from Desert Rock, from Payerne, and from Spitsbergen, from left to tight.

5.4 Summarization

These two case studies were conducted to illustrate the potential of our visual and content-based search system for explorative search and analysis in time series research data repositories. The application examples were compiled to the best of our domain knowledge. Please note, however, that our findings in these studies were not confirmed with earth observation researchers. Future work will include conducting such exploration together with domain experts, to further the understanding of practical requirements of researchers in visual search in time series data.

6 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 catalogs and on visual 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 development will take place in close collaboration with scientific users.

To this end, a first application workshop was conducted with approximately 20 researchers of Alfred-Wegener Institute in Bremerhaven, Germany. We introduced the basic idea and first results to researchers from different domains such as meteorology, climate research, and oceanography. In the preparation of the workshop initial discussions with researchers revealed the severe need of thoroughly introducing the concept of visual search in research data. While scientists use visualization tools for data analysis, the idea of a visual search approach has not been taken into consideration by the majority of the researchers. A classical approach of a user-centered design proved to be challenging, as the user workflows and tasks did neither include the concept of a visual search, nor did the user have any expectations to a visual search system yet. However, once closely introduced to the basic ideas, researchers started to adapt their working problems to our presented search concept, discussing new ways of knowledge explorations. All researchers expressed high interest in the idea of applying content-based and visual search as a tool to facilitate their daily research. It is expected that by visually searching for specific temporal measurement patterns, researchers are able to quickly search for evidence for specific hypotheses. During discussion with researchers it became clear, that in the different research disciplines, knowledge about data characteristics and assumptions about the natural processes underlying the observation data is implicitly available. This knowledge is expected to require different technical consequences, regarding, e.g., data normalization and descriptor extraction. Therefore, we plan to externalize this knowledge for a few selected scientific domains, and devise appropriate system configurations. In the long run, we hope to consolidate our findings in a search interface that can be applied as generically as possible across a wide variety of research data domains.

On a more conceptual level, we expect that the most useful search functionalities will not consist of only a single modality (e.g., curve shape), but rather a combination of different modalities (e.g., shape, metadata, parameter descriptions, and domain intervals). Additional modalities may involve correlation-based comparison of time series at different scales and localities. We further expect that metadata will play an important role, either for filtering of search results or as input to adaptive search algorithms. The outcome of the workshop advices the following steps next. The design of a real user problem with given real data and evaluation by a user group with expertise in the given domain will provide feedback concerning search modalities, descriptors etc., as mentioned above. Results of this evaluation will be a starting point for a more general usability evaluation with a broader user group invited from PANGAEA users. 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.

Finally, taking a long-term perspective, research should be investigated on how visual search in primary research data can be combined with searching in secondary publications that refer to primary data. References to publications are part of the metadata header in certain data sets. The automatic processing of secondary publications, and the integration of visualization and search regarding both, publications and underlying primary data, is a seperate open challenge that most likely requires consideration of semantic text analysis protocols.

7 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 We thank the Alfred-Wegener Institute (AWI) in Bremerhaven, Germany for kindly supporting this research effort. Rainer Sieger and Hannes Grobe helped us in developing an initial understanding of the data domain, and in selecting an appropriate PANGAEA data subset to experiment with. Gert König-Langlo was so kind as to organize a first user workshop for us. During this workshop, numerous AWI researchers provided helpful comments and suggestions for research and development in our effort. We thank the reviewers of the initial ECDL2010 submission for their helpful comments and suggestions. 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. 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, pp. 36–45 (2007)
2. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Lecture Notes in Computer Science, pp. 69–69 (1993)
3. 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, pp. 490–501 (1995)
4. Ahlberg, C., Shneiderman, B.: Visual information seeking: tight coupling of dynamic query filters with starfield displays. In: Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence, pp. 313–317 (1994)
5. Aigner, W., Miksch, S., Muller, W., Schumann, H., Tominski, C.: Visualizing time-oriented data—a systematic view. Comput. Graphics **31**(3), 401–409 (2007)
6. Bamboo Research Initiative: http://projectbamboo.org/. Accessed 20 May 2011
7. Baseline Surface Radiation Network (BSRN): http://www.bsrn.awi.de/. Accessed 20 May 2011
8. Berndt, R., Blümel, I., Clausen, M., Damm, D., Diet, J., Fellner, D., Fremerey, C., Klein, R., Krahl, F., Scherer, M., Schreck, T., Sens, I., Thomas, V., Wessel, R.: The PROBADO project—approach and lessons learned in building a digital library system for heterogeneous non-textual documents. In: European Conference on Digital Libraries, Lecture Notes in Computer Science, vol. 6273, pp. 376–383 (2010)
9. Brase, J.: Using digital library techniques-Registration of scientific primary data. In: Lecture Notes in Computer Science, pp. 488–494 (2004)
10. Castelli, D., Pagano, P.: Opendlib: a digital library service system. In: ECDL, pp. 292–308 (2002)
11. Chan, K., Fu, A.: Efficient time series matching by wavelets. In: Proceedings of the 15th IEEE International Conference on Data Engineering, 1999, pp. 126–133 (2002)
12. Chang, R., Charlotte, U., Ghoniem, M., Kosara, R., Ribarsky, W., Yang, J., Suma, E., Kern, D., Sudjianto, A.: Wirevis: visualization of categorical, time-varying data from financial transactions. In: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (2007)
13. Dryad Digital Repository for Data Underlying Published Works: http://www.datadryad.org/. Accessed 20 May 2011
14. 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, ACM, New York, NY, USA, pp. 12–19 (1999)
15. ELIXIR European Life Sciences Infrastructure for Biological Information.: http://www.elixir-europe.org/. Accessed 20 May 2011
16. German Research Foundation (DFG).: Report on round table meeting of research data (in German). Whitepaper (2008). http://www.dfg.de/download/pdf/foerde-rung/programme/lis/forschungsprimaerdaten_0108.pdf. Accessed 20 May 20 2011
17. Hochheiser, H., Shneiderman, B.: Dynamic query tools for time series data sets: timebox widgets for interactive exploration. Inf. Vis. **3**(1), 1–18 (2004)
18. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Knowl. Inform. Syst. **3**(3), 263–286 (2001)
19. Kohonen, T.: Self-Organizing Maps. 3rd edn. Springer, New York (2001)
20. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. Int. J. Digit. Libr. **6**(2), 124–138 (2006)
21. Liao, T.W.: Clustering of time series data—a survey. Pattern Recognit. **38**, 1857–1874 (2005)
22. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2003)
23. PANGAEA Publishing Network for Geoscientific & Environmental Data: http://www.pangaea.de/. Accessed 20 May 2011
24. PsychData National Repository for Psychological Research Data: http://psychdata.zpid.de/ (in German). Accessed 20 May 2011
25. Schreck, T., Bernard, J., Von Landesberger, T., Kohlhammer, J.: Visual cluster analysis of trajectory data with interactive kohonen maps. Inform. Vis. **8**(1), 14–29 (2009)
26. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, IEEE Computer Society, Washington, DC, pp. 336–343 (1996)
27. Sieger, R., Grobe, H., Diepenbroek, M.: Panplot—software to visualize profiles and core logs. Alfred Wegener Institute for Polar and Marine Research, Bremerhaven (2005). doi:10.1594/PANGAEA.330147
28. Šimunić, K.: Visualization of stock market charts. In: Proceedings of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (2003)
29. Society for Scientific Data Processing Goettingen: Cooperative long-term preservation for research centers (in German). Project Report (2009)
30. Van Wijk, J., Van Selow, E.: Cluster and calendar based visualization of time series data. In: IEEE Symposium on Information Visualization 1999 (Info Vis' 99), pp. 4–9 (1999)
31. Wattenberg, M.: Sketching a graph to query a time-series database. In: CHI '01 extended abstracts on Human factors in computing systems, CHI '01, pp. 381–382 (2001)
32. 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)
33. World Data Center System: http://www.ngdc.noaa.gov/wdc/. Accessed 20 May 2011
34. Ziegler, H., Jenny, M., Gruse, T., Keim, D.: Visual market sector analysis for financial time series data. In: IEEE Symposium on Visual Analytics Science and Technology, pp. 83–90 (2010)