

VisInfo: a digital library system for time series research data based on exploratory search—a user-centered design approach

Jürgen Bernard · Debora Daberkow · Dieter Fellner ·
Katrin Fischer · Oliver Koepler · Jörn Kohlhammer ·
Mila Runnwerth · Tobias Ruppert · Tobias Schreck · Irina Sens

Abstract To this day, data-driven science is a widely accepted concept in the digital library (DL) context (Hey et al. in *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, 2009). In the same way, domain knowledge from information visualization, visual analytics, and exploratory search has found its way into the DL workflow. This trend is expected to continue, considering future DL challenges such as content-based access to new document types, visual search, and exploration for information landscapes, or big data in general. To cope with these challenges, DL actors need to collaborate with external specialists from different domains to complement each other and succeed in

given tasks such as making research data publicly available. Through these interdisciplinary approaches, the DL ecosystem may contribute to applications focused on data-driven science and digital scholarship. In this work, we present *VisInfo* (2014), a web-based digital library system (DLS) with the goal to provide *visual access to time series research data*. Based on an exploratory search (ES) concept (White and Roth in *Synth Lect Inf Concepts Retr Serv* 1(1):1–98, 2009), VisInfo at first provides a content-based overview visualization of large amounts of time series research data. Further, the system enables the user to define visual queries by example or by sketch. Finally, VisInfo presents visual-interactive capability for the exploration of search results. The development process of VisInfo was based on the user-centered design principle. Experts from computer science, a scientific digital library, usability engineering, and scientists from the earth, and environmental sciences were involved in an interdisciplinary approach. We report on comprehensive user studies in the requirement analysis phase based on paper prototyping, user interviews, screen casts, and user questionnaires. Heuristic evaluations and two usability testing rounds were applied during the system implementation and the deployment phase and certify measurable improvements for our DLS. Based on the lessons learned in VisInfo, we suggest a generalized project workflow that may be applied in related, prospective approaches.

J. Bernard (✉) · J. Kohlhammer · T. Ruppert
Fraunhofer IGD, Darmstadt, Germany
e-mail: juergen.bernard@igd.fraunhofer.de

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

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

D. Daberkow · K. Fischer · O. Koepler · M. Runnwerth · I. Sens
German National Library of Science and Technology, Hannover, Germany
e-mail: debora.daberkow@tib.uni-hannover.de

K. Fischer
e-mail: katrin.fischer@me.com

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

M. Runnwerth
e-mail: mila.runnwerth@tib.uni-hannover.de

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

D. Fellner
Technische Universität Darmstadt, Darmstadt, Germany
email: dieter.fellner@igd.fraunhofer.de

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

Keywords Information visualization · Visual analytics · Exploratory search · Research data · Time series analysis · Digital library system

1 Introduction

Today, scientific information and knowledge are no longer solely encoded in text format, such as journal articles or book chapters. Increasingly, non-textual formats such as images, audio-visual material, computational models, and other numerical data sets are considered. For scientists, this body of non-textual data and information represents an invaluable source of possibly undiscovered knowledge if meaningful data subsets can be retrieved from typically large and heterogeneous research data repositories. The value of such research data and its (potential) benefit to society are widely accepted [17]. Time series data is an important research data type originating, e.g., in climate research, medical treatment, etc. However, besides the size and the heterogeneity, the time-varying behavior adds another level of *complexity* to this type of research data [34]. Today, a variety of research data repositories exist. DLs can support the access to and re-use of this valuable type of research data.

Historic scientific discoveries based on experimental, theoretical, and computational science paradigms were always the subject of library service support. In times of *data-intensive science* (also called the fourth paradigm in scientific discovery [26]), the role of DLs may be more important than ever. Part of the mission of a library of science and technology is to support scientists with methods that allow them to effectively use the available body of knowledge. This includes search and retrieval methods as well as exploration methods that also consider the non-textual data content. Further, it includes providing indexing and citation methods for future reference. To support scientists effectively, these DL functionalities should be adopted to their specific workflows (see Deelman et al. [19] for a characterization of scientific workflows). For example, the definition of similarity for the underlying data content, the calculation of features for retrieval, or the incorporation of aggregation techniques for large data collections have to meet the specific requirements of scientists.

While query-response technologies (as used in classical web search engines) typically act as lookup tools for fact retrieval and known-item search, ES goes beyond that: it aims to support scientific investigation and discovery, e.g., by revealing interesting facets of knowledge that scientists had not seen and considered before. Learning and decision-making are aspects that need to be involved in the information seeking process, especially for the content of complex data types such as time series research data. The success of ES tools depends on their visual-interactive capabilities.

Important components include the visual overview of the data content, the visual query definition, and the visual representation of retrieved items [68]. Visual overviews of the content can help scientists to explore large collections of time series research data. Visual-interactive query definitions based on examples or sketches of time series curves can help to make the search process more intuitive. Enhanced visual result representations can serve as an exploration space for applying facets or for gaining new and potentially unexpected knowledge. However, many scientists still perform at least part of their work using general purpose tools—most notably, Excel [29,64]. User-centered design approaches may help to raise the trust in new innovative technologies that address the discussed challenges.

We introduce VisInfo, a web-based, exploratory search system for time series research data. Our contribution is as follows: (1) we present our user-centered design approach, which was executed following a distinct design study methodology. We chose a user-centered approach to support the scientific workflow of our users in the best possible way, and raise the users' trust [29,64]. We included the users from the beginning of the design process that started with a domain and problem characterization, followed by an iterative development phase, and ended with the presentation of the web-based VisInfo system. The process was conducted as a collaborative effort between data collectors, data curators, digital librarians, digital library users, and computer scientists. (2) Furthermore, we present the result of this process, VisInfo DLS, that enables the content-based access to time series research data collected by the Baseline Surface Radiation Network (BSRN). Scientists can explore large data collections with a content-based overview visualization (Visual Catalog). Moreover, they are enabled to define queries of the time series content visual-interactively by example and by sketch. The Result View allows for the exploration of retrieved data subsets in detail. Different perspectives thus allow for an analysis of interesting relations hidden in (a) the time series data content, (b) the geo-location of respective measurements on earth, and (c) measurement period in a calendar-based view. (3) In addition, we present the results of different evaluation strategies iteratively conducted with digital library users. A use case and the results of the conducted evaluations reveal that the final VisInfo prototype is both usable and useful. (4) Finally, we review the lessons learned within the project and contribute a generalized project workflow description, that may help related visualization research efforts within the DL context.

2 Related work

In the following, we review initiatives in the field of DL to support scientific work with research data. Further-

more, we highlight ES approaches for time oriented data. Finally, we report on user-centered design practices from the perspectives of DL, information visualization, and visual analytics.

2.1 Scholarly support for research data

In this work, we treat research data as a special document type to be provided by a DL. Examples for data-driven research domains are physics, chemistry, medicine, biology, or earth observation, just to name a few. Facilitating the access to research data leads to both special benefits and challenges for scholarship and DL use. This especially applies for the type of data content considered in this approach (time series data), e.g., arising from measurements in a variety of research domains. Additionally, metadata attached to the research data documents play an important role. Examples of metadata include secondary data about test conditions, or geo-locations of scientific measurements. Many research domains differ in the nature of their research data, their conventions about data use and re-use, and their applied methods [17]. A classification of different types of research data is presented by Kehrer and Hauser [33], with an emphasis on visualization and visual data analysis. The process of passing research data through different phases is often referred to as the *data-life-cycle* [3, 13, 17]. The typical phases are data creation, data processing, data analysis, data preservation, data access, and data re-use. The VisInfo approach is primarily targeted towards the phases of data analysis, data access, and data re-use, and aims at tackling respective challenges related to search [13, 24, 40], and exploration [3, 25, 68] tasks (see Sect. 2.2 for more details).

Focusing on the user- and task-centered perspective, we highlight the relation of VisInfo to scientific workflows and scientific workflow systems. A (scientific) workflow is a high-level specification of steps (and dependencies in between) to accomplish a specific goal in a data-centered working environment [19, 39]. Each step represents the execution of a combinational unit, such as running a program, submitting a query to a database, submitting a job to a compute cloud or grid, or invoking a service over the web to use a remote source [26, 29]. One of the most challenging goals in the workflow construction phase is the user-centered development of a data processing pipeline. Typical challenges are based on the complexity of (time series) data [34], the involvement of different stakeholders (including data scientists) [18], and the construction of data-centered workflows with appropriate algorithmic routines in the correct order and with the correct parameter values. Regarding the latter, we refer to the survey of Knowledge Discovery in Databases (KDD) [21], or to basic research in visual analytics [34].

Other challenges related to analysis workflows for research data concern, data management, and data transformation [39], the visualization [23], and advanced search interfaces [25].

As a DLS prototype, VisInfo incorporates research data warehouses (data repositories) as the targeted data source to facilitate data access and data re-use. While not being limited to a particular repository, VisInfo currently accesses thousands of (time series) data sets of the PANGAEA repository [48], a publisher for earth and environmental science data. A variety of research data repositories exists, such as, for example, the Sloan Digital Sky Survey (SDSS) [57] for physics and astronomy research data. Surveys of scientific publications, (open) data publications (e.g., Elsevier)¹, and scientific data repositories are presented by Marcial et al. [41], and Costas et al. [17]. Finally, we highlight the beneficial means of open data initiatives [63] and metadata standards [14] for this work, increasing the demand for (visual) accessibility functions by scientists working with such data.

2.2 Exploratory search in time series data

While a variety of DL approaches for textual access to research data exist [20, 47, 48, 61], the number of DL approaches for non-textual access to research data is comparatively small. However, research fields like information visualization [16], visual analytics [34] and ES [40, 68] concentrate on challenges like enhanced exploration and search interfaces for research data. Kehrer and Hauser compare approaches from an information visualization perspective [33]. A review on works for visualizing climate change data is presented in [46]. In exploratory analysis scenarios, clustering approaches are often applied to group large research data. Based on such aggregation techniques, global overviews of the complete data set can be provided, also known as content summaries. Promising examples exist for geo-science [1], earth observation [8, 55], energy consumption [66], human motion analysis [11], or cancer research [10]. A metadata-based clustering approach is presented in [65]. In [8, 9], and [10], clustering was used to gain insight into interesting relationships between the data content and metadata associated to the research data content. Visual query definition concepts for time series data are reviewed in [5], a comprehensive survey for the visualization of time series is presented in [2]. Research data-based approaches on multivariate data changing over time are presented in [12, 62] considering event data or climate data. From this review of the related work, to the best of our knowledge, combined approaches combining research data with content-based visual search in the DL context are scarce.

¹ <http://www.elsevier.de/>.

2.3 User-centered design in digital libraries

In the early 1990s, usability concepts and methods made huge progress and also found their way into web-based applications. Shackel’s work from 1991 can be seen as one of the cornerstones of this time [59]. Concepts and models such as iterative test and development cycles [43], heuristic evaluations [45], or cognitive-walkthroughs [37] were introduced. A comprehensive overview of these approaches is provided by John [32]. In 2000, Saracevic reviewed usability evaluation efforts in the DL context [54]. He confirmed a lack of evaluation methodologies in DL, and suggested a conceptual framework for DL evaluation. An early formal evaluation during the DL development process was done by Hill in the Alexandria DL Project (ADL) [27]. In the following, usability models and concepts were examined and adapted to the DL context [30,31]. In 2005, Reeves published a comprehensive guide on evaluation of DL [50]. Some of the latest outcomes of usability evaluation for DL include the Greenstone User and Developer Survey [60]. In 2010, Yuan et al. carried out a usability evaluation to examine differences in users’ experiences between an information visualization system and a text information retrieval system [69]. Their final results indicate that visualization techniques help to improve the representation and organization of information in retrieval systems. The discussion and adoption of usability evaluation in the information visualization context can be found in [4,49]. Hoerber [28] surveyed user evaluation methods for visual web search interfaces. Their model already includes several development iterations followed by distinct evaluation methods. A generalization of the design and validation of visualization techniques is introduced by Munzner [42]. In her nested model, she clearly defines four nested layers: domain problem and data characterization; operation and data type abstraction; visual encoding and interaction design; and the algorithm design. In our approach, we adapted this method to the DL context. Finally, Sedlmair et al. fit these concepts into a process describing design study methodologies in general [58]. The authors highlight the importance of characterizing the relevant stakeholders and collaborating closely with these during the entire design process.

3 The VisInfo concept

A wide-spread method, which scientists across many fields employ, is to create visual representations of the data they have generated. Often, it is effective for humans to intuitively assess the essence of a data set that is plotted [16]. Graphical methods of analysis, optimization, problem solving, and design have been used for centuries. Further, there is the notion of designing displays that deliberately show

chart- and visual-based representations of numerical values, rather than tables of numbers. That is because for a human, it is easier to intuitively gauge the position of pointers than to actively read digits and have to mentally process them, given the limitations of the short-term working memory [16]. In a similar way, the essence of a curve is often grasped more easily than a series of numbers. Or to quote Henry Hubbard: “There is a magic in graphs. The profile of a curve reveals in a flash a whole situation—the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces.”

This notion is at the heart of the VisInfo approach. For this reason, the focus lies on visual analysis of curve shapes. At the Alfred Wegener Institute (AWI) in Bremerhaven, Germany, scientists have developed and employed tools to collect and plot data, enabling them to intuitively process and visually analyze the data set, which shows the importance of graphic representations. For this reason, a related data collection of the Baseline Surface Radiation Network (BSRN) [6] has been chosen for a first prototypical application.

In the following, we recall the basic concepts of our previous work, which formed the basis for the VisInfo prototype.

3.1 Focus and adopted similarity notion

We first started to map out how the goal of visual and content-based search in research data could be operationalized. The domain of research data is huge, and, together with our project partners, we decided to restrict ourselves to *time series* research data. An initial workshop held together with domain experts from the PANGAEA data library [48] operated at AWI revealed that among the many different research interests in earth observation science, time series are a ubiquitous data type, the support of which is expected to benefit a larger user community. As a core DL functionality, we decided to implement content-based visual search support to help scientists explore and retrieve data of interest. Content-based similarity is a multifaceted problem and appropriate similarity notions depend highly on the type of application, including partial similarity, and similarity across different levels of resolution, including agreement of measurement parameters [22,38].

In our initially proposed concept [5], we chose to start with a simple, robust similarity measure. We partitioned all time series of a PANGAEA test repository to segments of one day length each. The resulting curve shapes were named *curve patterns* in our retrieval system. Finally, the Euclidean distance function was applied, to define the similarity of the time series patterns.

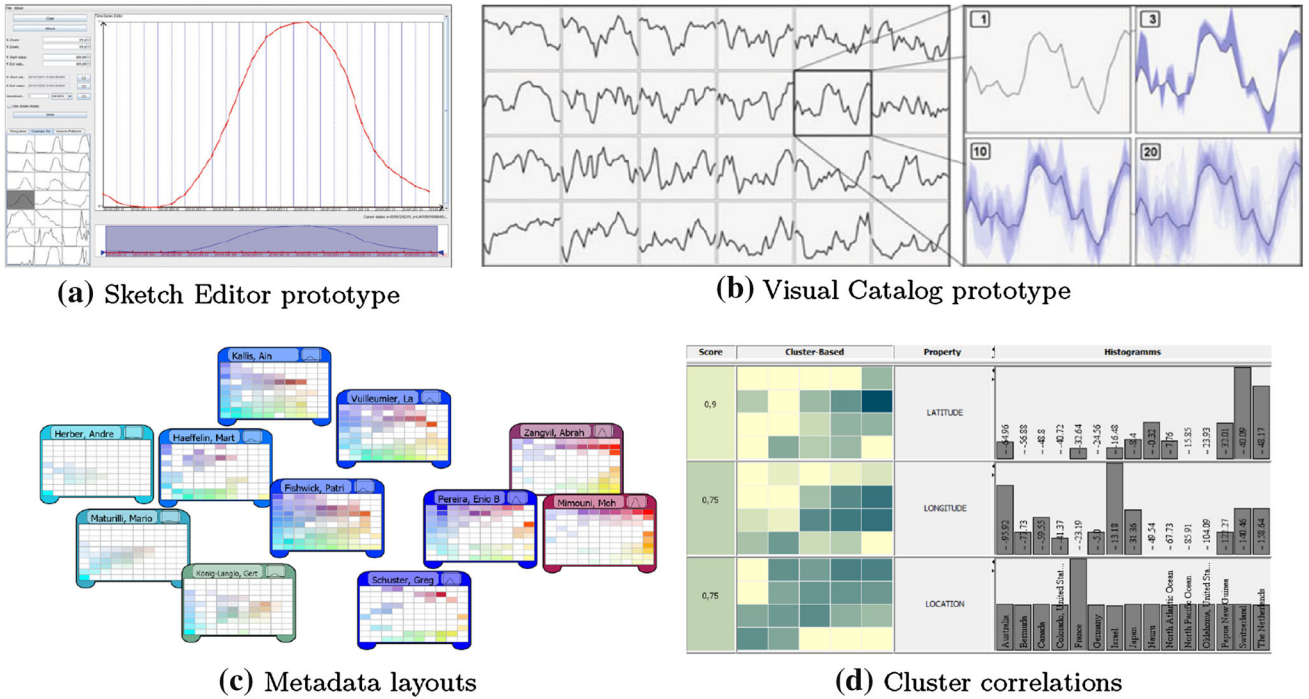


Fig. 1 Key functionalities of our initial concept for visual search and analysis in time series data included (a) sketch-based querying and (b) a cluster-based overview. Further analytical support can be pro-

vided based on these key elements, including (c) combined content- and metadata-based layouts, or (d) highlighting of interesting correlations of clusters with metadata

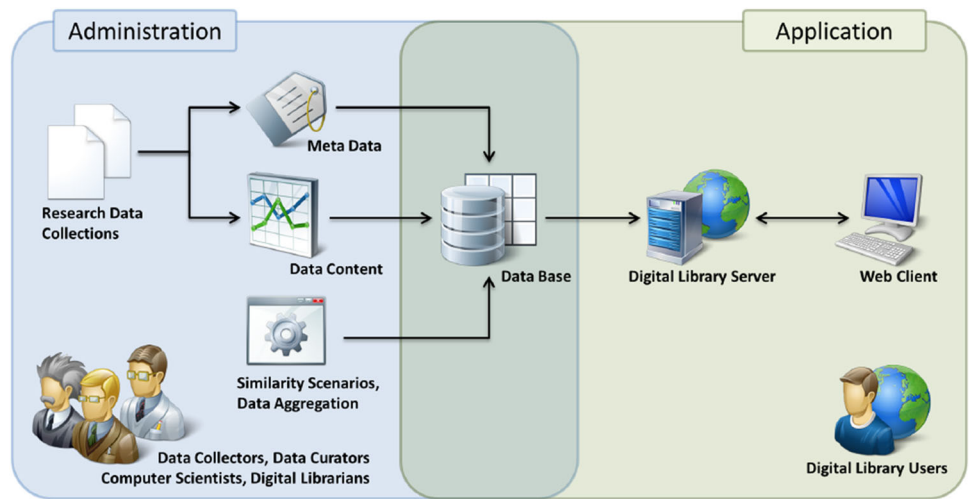
3.2 Visual search and exploration facilities

Our initial concept supported two data access methods: content-based *visual querying* and *visual browsing*. Visual search is based on the query-by-sketch principle. By means of an interactive sketch interface, users can define a curve pattern in which they are interested (Fig. 1a illustrates this concept). The system computes a set of top n results of the most similar time series and displays these in a sorted list. Visual browsing allows a bird’s-eye perspective on the entire set of time series. The Self-Organizing Map method (SOM) [36] is applied on the set of curve patterns and produces a 2D grid of cluster prototypes, representing distinct time series patterns in the given data. As a specific feature of the SOM, the output can be directly visualized as a content summary solution, which we call ‘Visual Catalog’ (see Fig. 1b). By the nature of SOMs, the patterns in the Visual Catalog are sorted based on their similarity. This ensures topology preservation of the provided map metaphor. With the Visual Catalog, users can browse and drill down for specific clusters of interest. Bundles of data elements can be visualized with blue opacity bands [56]. It is also possible to select a curve pattern in the Visual Catalog for search (query-by-example). Further basic functionality includes the filtering of metadata items and the highlighting of the correspondence to the clusters or content-based search results.

3.3 Analytical facilities

In our work, we focused on the question as to which extent digital data repositories can or should support analytical tasks of the end users. To most scientific users, retrieving time series data of interest is not the end, but the start of further research. In previous work, we therefore also explored which analytical services could be added to the basic search and exploration approach described above. Ideally, such analysis functionality should a) be efficient to implement on top of the basic system and b) be of value to a large number of users. We earlier considered two such analytical services. For one, we devised a schema for *combined content-based and metadata-based visualization* [8]. This schema allows the user to select a metadata attribute. As an example, consider the selection of the *creator* attribute. For each instance of this attribute, a Visual Catalog is enriched with a highlighting of the data items provided by the given creator. Glyphs of the catalog for each creator are then shown in a layout, which represents the similarity between each creator by distance in the layout (illustrated in Fig. 1c). Thereby, users can compare the similarity between data level and metadata level, potentially finding interesting cross-relationships between both data aspects. A second analytical service relates to the identification of strong correlations between clusters of time series and frequent metadata items [9]. For example, consider the set of

Fig. 2 The VisInfo architecture



locations at which the member time series within a given cluster of time series have been measured. Then, the presence of a high number of identical location attributes within a cluster may be of interest, as it relates to a robust cluster of measurements. Technically, we compute the degree of unimodality of the histogram over metadata for a given cluster, and identified the most unimodal metadata elements (see Fig. 1d).

4 Domain and problem characterization

The conceptual focus of this research effort lies on the similarity notion of research data and the visual search and exploration facilities realizing visual access to this data. As is typical in design studies, the design process started with a domain and problem characterization, with respect to the VisInfo concept presented in the previous section. First, we identified the stakeholders relevant for the application domain, and accordingly, for our user-centered design process (Sect. 4.1). Then, we identified domain challenges in collaboration with the identified stakeholders (Sect. 4.2).

4.1 Relevant stakeholders in the design process

Inspired by the stakeholder characterization presented in the DELOS digital library reference model [15], we identified five different stakeholder types: data collector, data curator, digital librarian, computer scientist, and digital library user. Representatives of these stakeholders were included in the design process. In the following, the stakeholders are characterized, and their input for the design process is presented.

The *data collectors* are gathering the data, which is later used to create and validate scientific hypotheses. Thereby, the *data type* is defined. In our approach, the data collectors are earth observation scientists of the Baseline Surface

Radiation Network (BSRN). The data consists of time series measurements, e.g., several physical parameters of radiation, temperature, etc. An excerpt of the BSRN data set [6] comprising all measurements of the BSRN is used as the basis of our search and retrieval system. The size of the data set is 25 Gigabytes in ASCII format. For more information about the data set, we refer to our concept paper [5], where a subset is used and described in detail.

The *data curators* have to ensure that the data provided by the data collectors are persistently stored in a data repository in a standardized *format*. Moreover, they enable the citability of data sets through digital object identifiers (DOIs) [14]. In our approach, the hosts of the PANGAEA portal [48] play the role of the data curators. They provide access to most of the data collected within the BSRN. The data is stored in a specific format—“ISO-8859-1: ISO Western - PANGAEA default”—that consists of tab-separated text files comprising time series measurements of one month and associated metadata. Each file has an own DOI (Fig. 2).

The *digital librarians* in our approach are representatives of the German National Library of Science and Technology (TIB). Their main goal is to provide their users a content-based search and retrieval system for time series research data. Their main focus is on the *ease of use* of the system, since due to its innovative nature caused by the implied paradigm shift [26], any additional barrier could potentially cause the user turn away from the system before even fully assessing its possibilities. Moreover, the digital librarians requested to work from early on with *real data* to discuss early demonstrators with real users. To support this, a quick import of new data sets into the system was also requested by the digital librarians.

The *computer scientists* are the developers of the DLS. To enable the visual access to content-based information, experts from the field of data analysis and information visu-

Fig. 3 Visual-interactive query definition

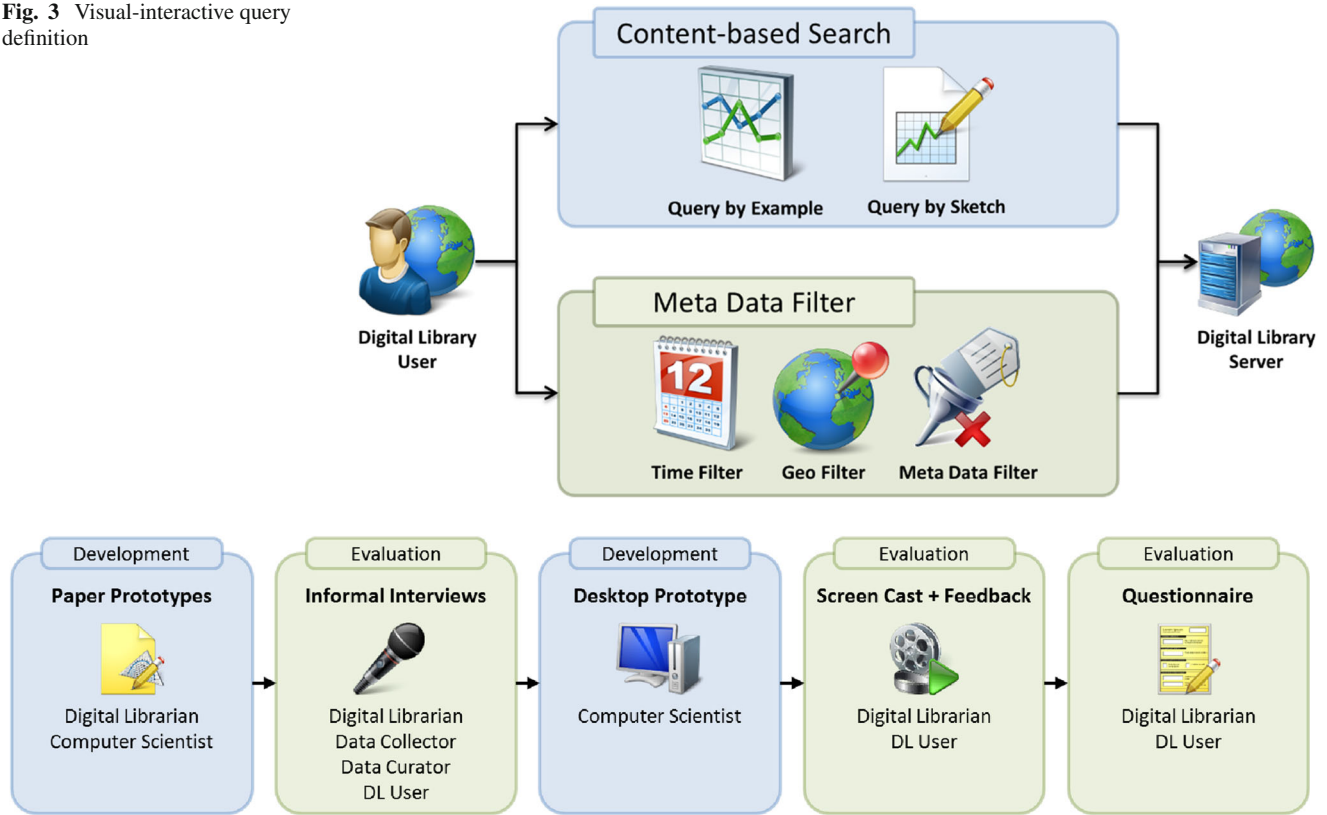


Fig. 4 User-centered design process: domain and problem characterization

alization are needed to cover this role. In our approach, computer scientists from the Technical University Darmstadt and the Fraunhofer Institute for Computer Graphics Research represent this stakeholder.

The *digital library user* is the end user of our DLS. In our approach, earth observation scientists represent this stakeholder. They basically cover two roles in our approach: data collectors and digital library users. Thus, digital library users fulfill the role of the *domain expert* known from design study methodology [58].

In addition to these implicit requirements on the VisInfo system given through the addressed domain, the stakeholders involved, and the selected data set, we also identified challenges in the domain in collaboration with potential users of our system in a distinct process. From these challenges, explicit requirements on our system may be extracted. In the following section, this process is described in detail.

4.2 Relevant challenges in the application domain

To understand and characterize the problems and challenges in our application domain, we worked in close collaboration with our potential users, the earth observation scientists, from the very beginning. The definition of explicit challenges to be tackled by the VisInfo system was supported through a 5-step process (Fig. 4).

As a first step, we developed paper prototypes and showed them to the users to demonstrate our general idea of exploratory content-based search. We then gathered first feedback from the users through informal interviews. We learned that for all users the general idea was inspiring. Still, it was difficult for them to adapt the concept of searching in the data content, since this possibility was not given before. As a second step, a desktop prototype was developed. The core functionality of this system has been described in Sect. 3. Basically, the users could define queries by sketch or example, they could activate filters on the metadata, and finally, they could look up details of the searched patterns in the result list. The prototype was introduced to the users via a screen-cast. Finally, we evaluated the feedback of 19 earth observation scientists from the BSRN for our interpretation of the domain and their challenges. In the following, we describe the main challenges identified in the described process.

(1) *Varying similarity notions*: To support content-based search in time series research data, a definition of similarity has to be provided. Depending on the tasks and the users, the interpretation of similarity may vary. In some cases, absolute values of the measurements have to be compared (e.g., to distinguish arctic from desert regions). In some cases, similarity has to be defined via the relative shapes of curves (e.g., to identify similar trends). Summarizing, the similarity of

absolute values as well as curve shapes need to be considered.

(2) *Different time intervals*: For the earth observation scientists within our stakeholder group, the time series to be compared should have the same length (e.g., one day, etc.). However, different lengths may be of importance for their work. The most important time intervals for a curve shape identified were years, days, and hours. Years and days reflect periodicities of natural phenomena. In reconciliation with an expert at the Alfred-Wegener Institute, we decided to focus on daily curve shapes as fixed time interval called patterns in our retrieval system.

(3) *Relevant Parameters*: For our specific user group, the most relevant physical parameter in the BSRN environment is the shortwave downward radiation (SWD). This measure is relevant for climate research, especially for giving statements about cloud occurrences. Although the parameter temperature is not of the highest relevance for the earth observation scientists, it was considered to be important for demonstration purposes to scientists from other domains.

(4) *Visual overview*: To provide a visual overview of the entire data set is a very important challenge to be addressed in order to support data exploration. The concept of a Visual Catalog showing the underlying data in one view was presented to the users via the desktop prototype. They reported the lack and the importance of such an instrument.

(5) *Different query modalities*: Another challenge described by the users was the support of different modalities to formulate a query. Both concepts, query-by-example and query-by-sketch, were considered relevant for their work. Moreover, including the metadata in the query formulation was important for the users.

(6) *Flexibility of scientific process*: As another outcome of our domain characterization, most of the users emphasized the need for flexibility in their work with time series data. As mentioned before, possible usage scenarios always vary with respect to time interval lengths, physical parameters, similarity measures, etc., based on the task and user at hand.

In summary, the concept presented via the initial desktop prototype, integrating content-based and metadata-based search, addressed many challenges expressed by the earth observation scientists.

5 The web-based VisInfo prototype

VisInfo [67] is a web-based DLS for the ES of time series research data collections based on visual access. Interested readers are invited to test VisInfo by following this link.²

² <http://demo.vis-info.info/>.

Now, we first describe the results of the data abstraction phase and highlight the non-visual functionality. Second, we present the visual-interactive capability of VisInfo, as a result of the iterative development phase. In three subsections, we show different visual-interactive views for the content summary, the visual query definition, and the search result analysis. Furthermore, we illustrate a use case of a real-world ES scenario conducted together with domain experts within the development phase. The aim of the use case is to showcase the hypothesis generation and validation process, and thus, to assess the usefulness of VisInfo. Finally, performance aspects of the system are discussed.

5.1 Data abstraction and non-visual capabilities

The system architecture of VisInfo is structured in an administration and an application section (see Fig. 2). One data-centered goal of the administrative section is the construction of scientific workflows in a collaborative effort between data collectors, data curators, digital librarians and computer scientists. To cope with the heterogeneity of different time series research data sources, we provide a general internal time series data structure. The varieties of different time series properties are well characterized in the book of Aigner et al. [2], which served as a guideline for the development of the time series data structure used here. An extension of the DataCite metadata kernel [14] was used for the definition and storage of associated metadata. We analyzed the metadata corpus to find the most relevant metadata properties to be applied for faceted search (see Sect. 5.4). On the one hand, data collectors and librarians were asked for relevant metadata properties. Qualified metadata attributes include the contributing scientist, the measurement location on earth, and the climate and surface type (also see Fig. 12). On the other hand, the computer scientists analyzed the relevance of available metadata properties in two research efforts [8,9]. They identified interesting relations between the daily patterns and the season attribute, which was added to the facet list, respectively..

The data-centered workflow of VisInfo is shown in Fig. 7. On the basis of the general time series data structure containing the raw input data, a variety of transformations need to be applied on the time series content. The provided routines help, e.g., to (a) establish consistent data quality, (b) make the time series equidistant, (c) remove outliers, or (d) normalize the value domain. Time series descriptors transform the time series data into the feature space. For the VisInfo prototype, the Piecewise Aggregation Approximation (PAA) [35] and the Perceptual Important Points (PIP) [70] descriptor were finally applied for the workflow execution [7]. The resulting features are utilized to facilitate both the content-based retrieval and the data aggregation. Thus, a meaningful definition of time series similarity on the basis of the domain

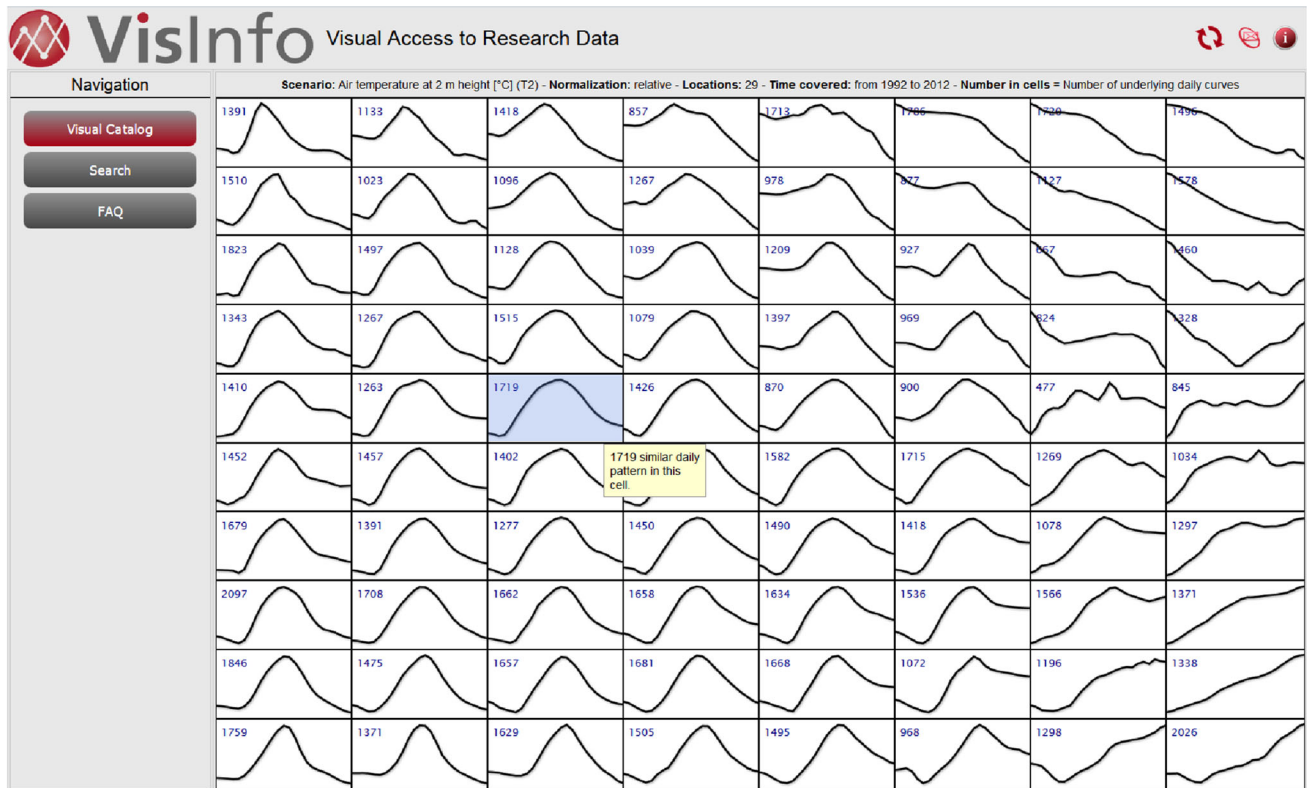


Fig. 5 Visual Catalog. An overview of 125,000 daily temperature curves. Each of the 80 cells represents about 1,500 daily curve patterns of the data set (see small blue numbers)

experts was of key importance for the usefulness of the VisInfo system. One of the most important insights gained in the domain characterization phase was the variety of different notions of time series similarity depending on the interest of the domain experts (cf. Sect. 4.2). To cope with this challenge, we developed the data preprocessing workflow in a modular way. A visual-interactive toolkit served as a platform for the domain experts and the data scientists for the definition of meaningful workflows, both for time series preprocessing and time series similarity. See [7] for the technical details of the four similarity scenarios provided in VisInfo. With the VisInfo web prototype, users can choose between the content of daily curve patterns in shortwave downward radiation or temperature measurements. Moreover, based on an inquiry of the domain experts, the curve patterns of the measurements can be compared by absolute values, or by relative curve progression (based on an additional normalization). Alternatively discussed, similarity scenarios considered variations, e.g., in the time duration of a pattern (yearly instead of daily), in the choice of alternative measurements (CO_2 , hPa , etc.), or in the choice of distance measures for feature comparison (dynamic time warping [53] instead of the Euclidean distance).

An important functional component of VisInfo is the data aggregation strategy. The overall goal of the data aggrega-

tion is to provide a content-based overview (content summary) of potentially large research data repositories. Visual content summaries, in turn, facilitate exploration capability, and enable querying by example (see the visual-interactive sections). We chose the Self-organizing Maps clustering algorithm (SOM) [36] (see Sect. 3.2) on the basis of the PAA feature vector to provide a content-based overview. For the construction of the SOM clustering step in the workflow, we applied visual analytics techniques presented by [56].

In the following, we demonstrate the SOM-based content summary and other visual-interactive capabilities of VisInfo to show the applicability as a DLS and as a research support tool for time series research data. The provided figures of the different aspects of the VisInfo system are all based on a temperature measurements scenario.

5.2 Visual overview of large research data collections

The initial step in the ES process is getting an overview of the important aspects of the provided data collection. This helps to get the ‘big picture’ of the data set. Further, it enables the user to define appropriate queries [40,68]. The SOM-based Visual Catalog in Fig. 5 provides a global overview of the complete data content by showing the most prominent daily

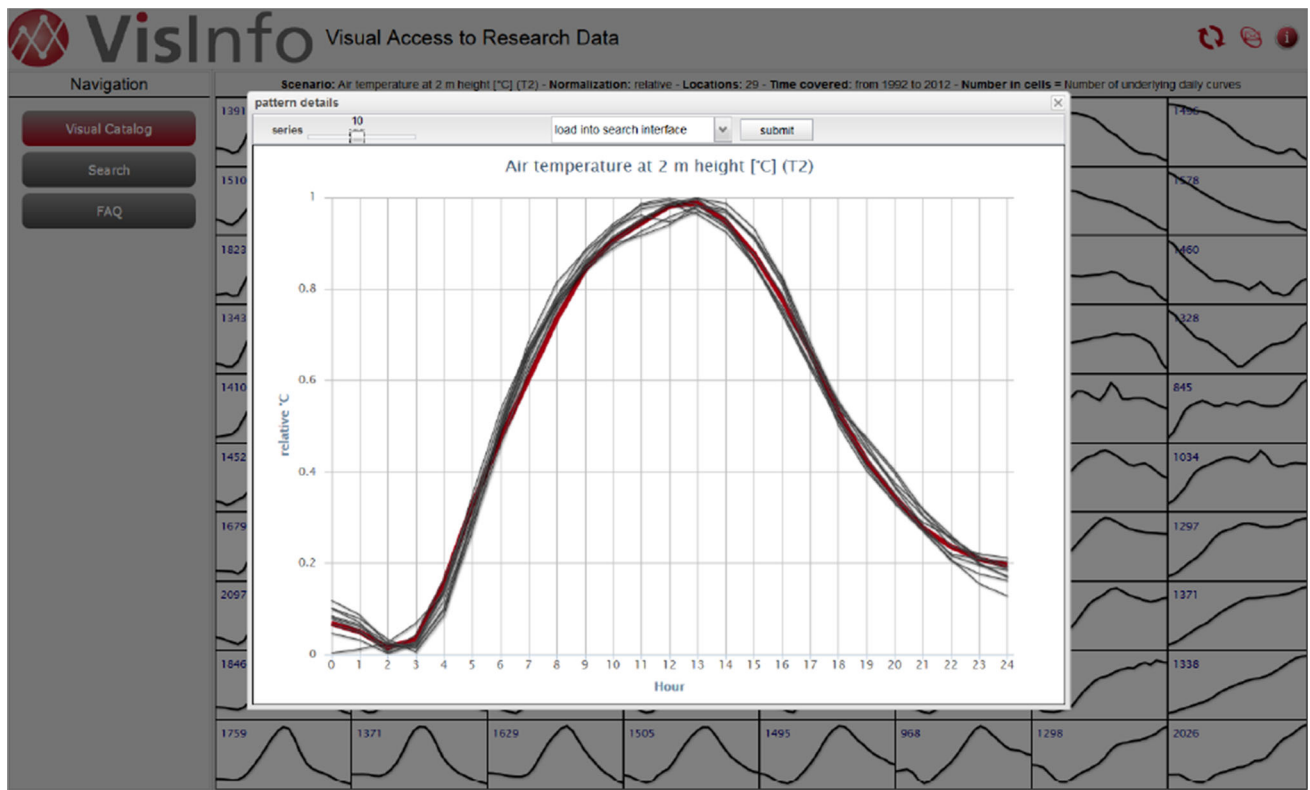


Fig. 6 Details on demand of a single cell in the Visual Catalog (cf. Fig. 5). The most similar patterns are shown. The red temperature pattern can also be used for query-by-example

temperature measurement patterns as black lines in different cells. Each cell represents a number of similar curve patterns (a cluster) of the data set. This number is indicated at the upper left of each cell.

We start the use case by recalling Fig. 5. The overview visualization represents 125,000 daily temperature curves in a single visualization, distributed in the 8x10 cells. For example, the upper left cell represents 1,391 daily temperature curves of the data set. A relative use case scenario is chosen, so that the relative temperature curve within one day can be observed. At the left of the Visual Catalog, all curves show a daily temperature progression with low values in the morning and in the evening, as well as a peak at noon. As expected, this daily pattern corresponds to the most prominent sun-dependent temperature progression on earth as approved by the domain experts. However, other (rather unexpected) daily temperature measurements are available in the data set like, e.g., linear upward trends at the lower right or linear downward trends at the upper right of the Visual Catalog. The scientists confirm that such trends most often occur at polar areas where the influence of the sun is markedly weaker than in other regions on earth. This is a first example of how VisInfo supports learning of intrinsic properties of the search space and enables digital library users to focus on interesting subsets of a research data set.

Another step in the ES process relates to the information drill down when an interesting data aspect was identified [16]. In our use case, we decide to drill down in the search space to a cell with 1,719 curve patterns, as highlighted in Fig. 5. By clicking on the cell, detailed information is provided in an enlarged display as shown in Fig. 6. Here, a number of similar temperature patterns can be explored. Black lines indicate temperature measurements of the data set and the red line shows the representative for the particular cell (the cluster centroid). In this example, most curve patterns have their minimum at hour 2 and their maximum at hour 13. Driving questions may be where on earth these measurements have been taken, or which scientists were responsible for the respective measurement stations. In the following, we describe how the explored data content can be used for the query definition.

5.3 Visual query definition

VisInfo provides visual-interactive means for the query definition to enable efficient and effective retrieval. The basic Search Interface is shown in Fig. 8. In the upper part, a description for the visual-interactive query definition is given. The user can define a query in the Sketch Editor and additionally apply filters on the metadata. An overview of all

Fig. 7 Data-centered workflow. Time series raw data are processed and transformed to feature vectors. The two feature vectors are utilized in four different transformations

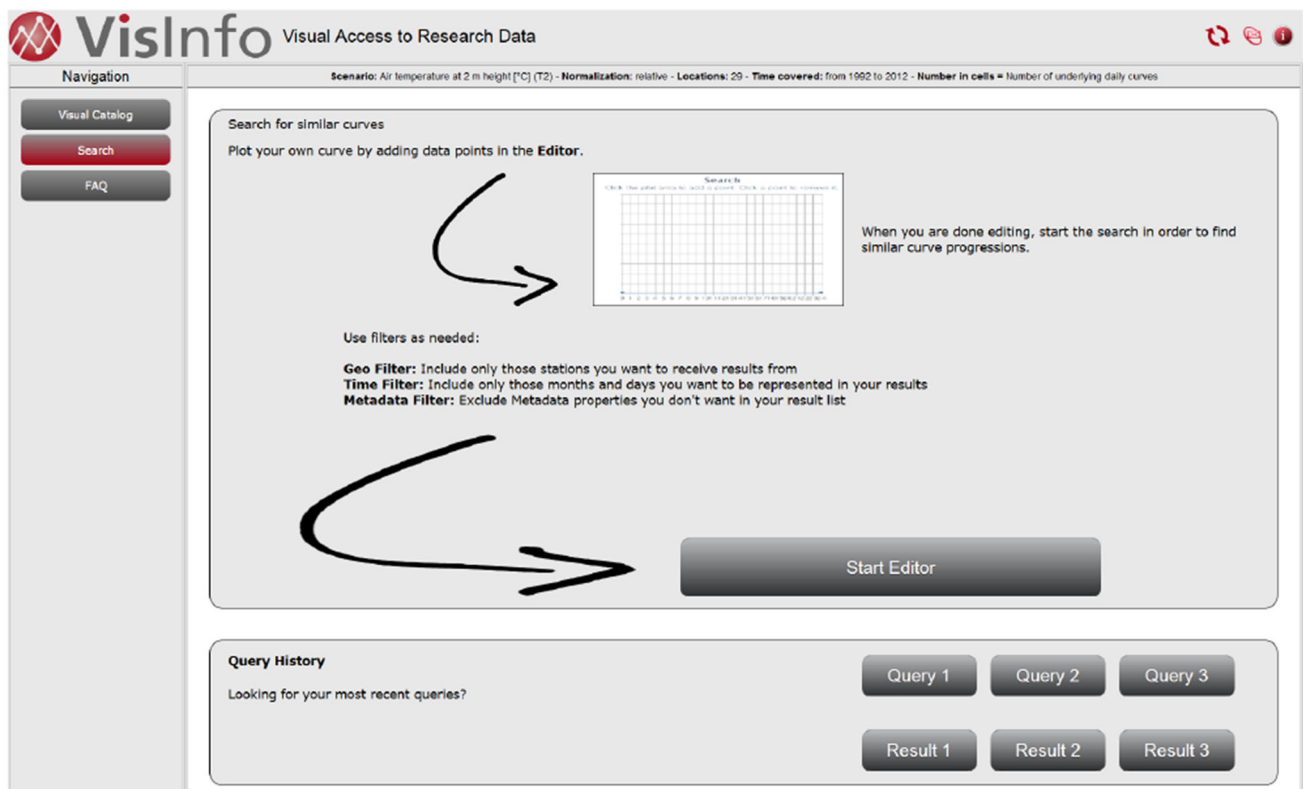
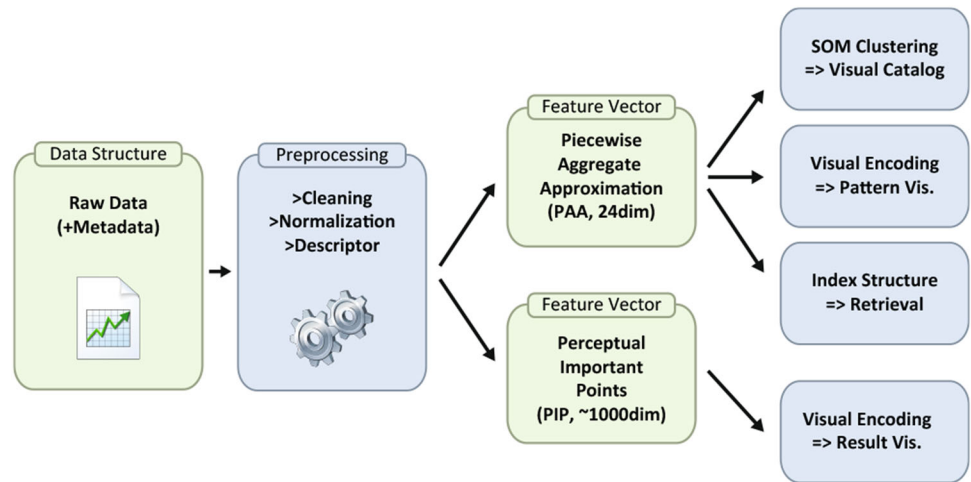


Fig. 8 Search Interface. Users can either start the sketch editor or recall an earlier search result. Both reloading recently defined queries and search results is possible

types of visual-interactive query definition is shown in Fig. 3. In the lower part of the Search Interface, the query history is provided. The users can review the queries executed in a previous search and view the corresponding results. By clicking the 'Start Editor' button, the user is forwarded to the Sketch Editor (see Fig. 9). Users can load one of the cell patterns of the Visual Catalog into the Sketch Editor as the basis for content-based search (query-by-example), or define

a new query by hand in the Sketch Editor (query-by-sketch). The modification of an example pattern in the Sketch Editor is also possible.

In the use case example, the cell pattern discussed in the previous section is loaded into the Sketch Editor, as can be seen in Fig. 9. In this particular case, we did not refine the visual query with the sketching capability but used the loaded pattern directly as a query by example. Before executing the

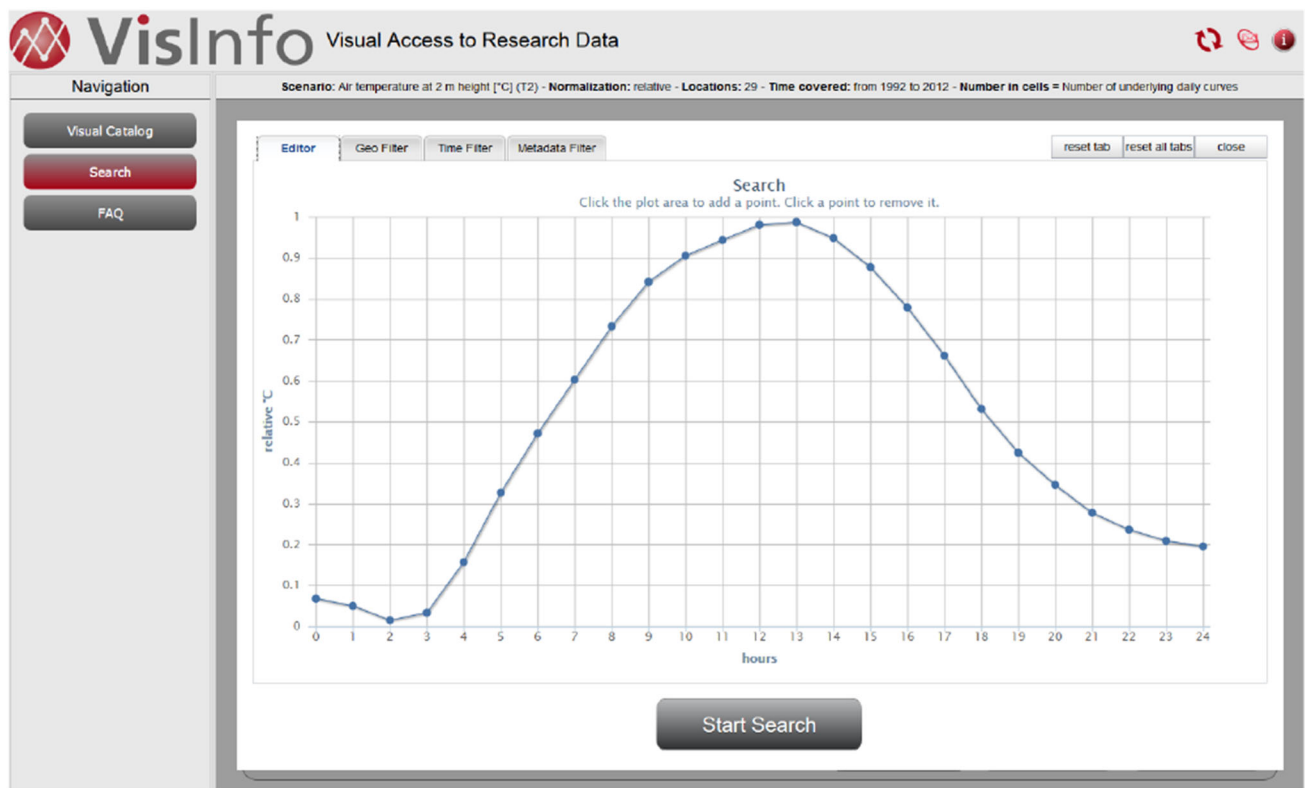


Fig. 9 Sketch Editor for the creation and modification of content-based queries. In this example, the user loaded an example curve pattern from the Visual Catalog (cf. Fig. 6)

query, the search space can be reduced by additional filters applied on the metadata. For example, if a user only wants to search for data sets between the years 2000 and 2002, the Time Filter in Fig. 10 can be applied. Additionally, a Geo Filter is provided (Fig. 11 illustrates). Finally, the remaining metadata properties can be filtered in the Metadata Filter (see Fig. 12). When the user is satisfied with the specified curve to be searched and the filters on the metadata, the search can be executed.

5.4 Visual search result analysis

The visual presentation of the search results defines a second exploration space. The retrieved results are shown in the list-based Result View as shown in Fig. 13. At this point, the PIP feature vector (cf. Fig. 7) comes into play. The feature vector serves as the basis for the detailed visual representation of the (raw) time series content at the center of the display. The content of the BSRN data set [6] consists of monthly measurements, visually encoded as blue line charts. The daily patterns retrieved with the query of Fig. 9 are highlighted with red background colors. At the right of the Result View, additional metadata are shown and external references such as the DOI URL. At the bottom, the user can press the “find more” button to extend the search result, and thus,

expand the result exploration space and increase the recall of the retrieved instances. Besides the list-based visualization of the search result, the Geo View tab arranges the retrieved documents by their geo-reference in a map-based view (see Fig. 14). Coming back to the use case, it can be seen that all geo-locations of the retrieved measurements are located (a) at the northern hemisphere and (b) at latitudes between 35° and 50°. Findings such as these may serve as an interesting starting point for the digital library users’ in-depth analysis based on a new hypothesis. Finally, the Time View (see Fig. 15) visually represents the temporal information of the retrieved documents in a matrix-based calendar view. In the use case, the search result space was extended to the 50 most similar patterns. It is interesting that all daily patterns of the search result were measured in the summer period. Thus, the result exploration reveals a relation of the retrieved daily pattern to the seasonal occurrence within a year. The three complementary perspectives of the result set (Result View, Geo View, and Map View) help in gaining a deeper understanding of the data, and facilitate the identification of interesting relations. For example, the shape query defined in Fig. 9 seems to have a significant geographical and temporal relation, which could be a starting point for new scientific investigation.

In addition to the three different views of the search result presentation, metadata facets can be applied to reduce the

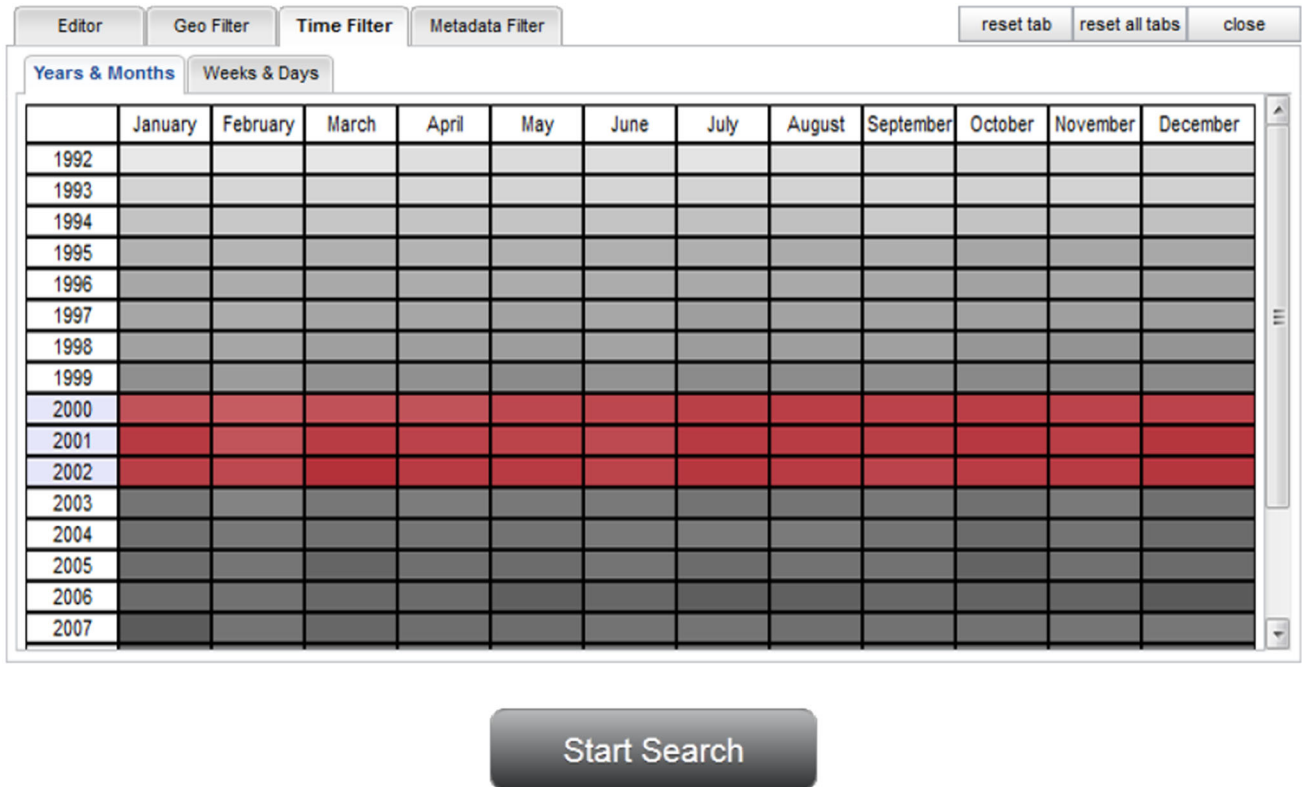


Fig. 10 Time Filter (cutout from the Search Interface). Line-based search space filter to the years of 2000, 2001 and 2002. Column-based and mutual filter definitions are also possible



Fig. 11 Geo Filter (cutout from the Search Interface). The user has selected only measurements from North America to be considered for search

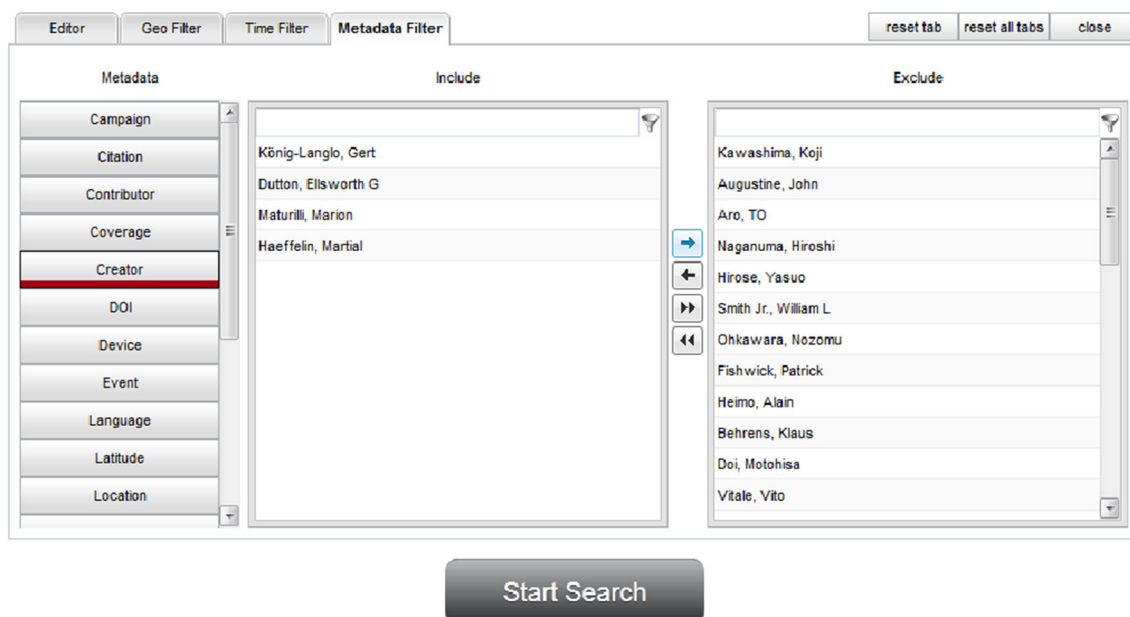


Fig. 12 Metadata Filter (cutout from the Search Interface). Reduction of search space to documents from four creators

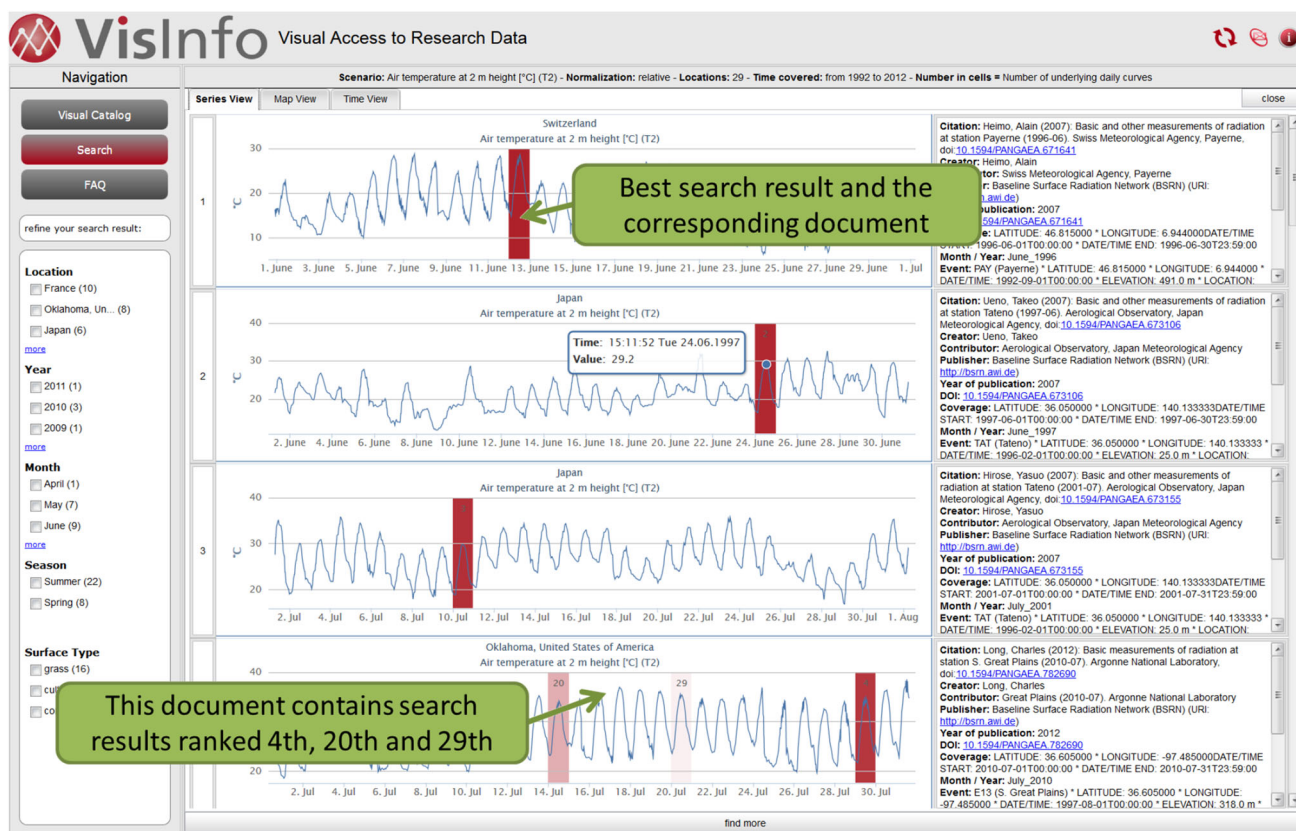


Fig. 13 Result View. Search result visualization including metadata facets (left), list-based visualization of the data content (center), and respective metadata (right). The result set corresponds to the the query of Fig. 9

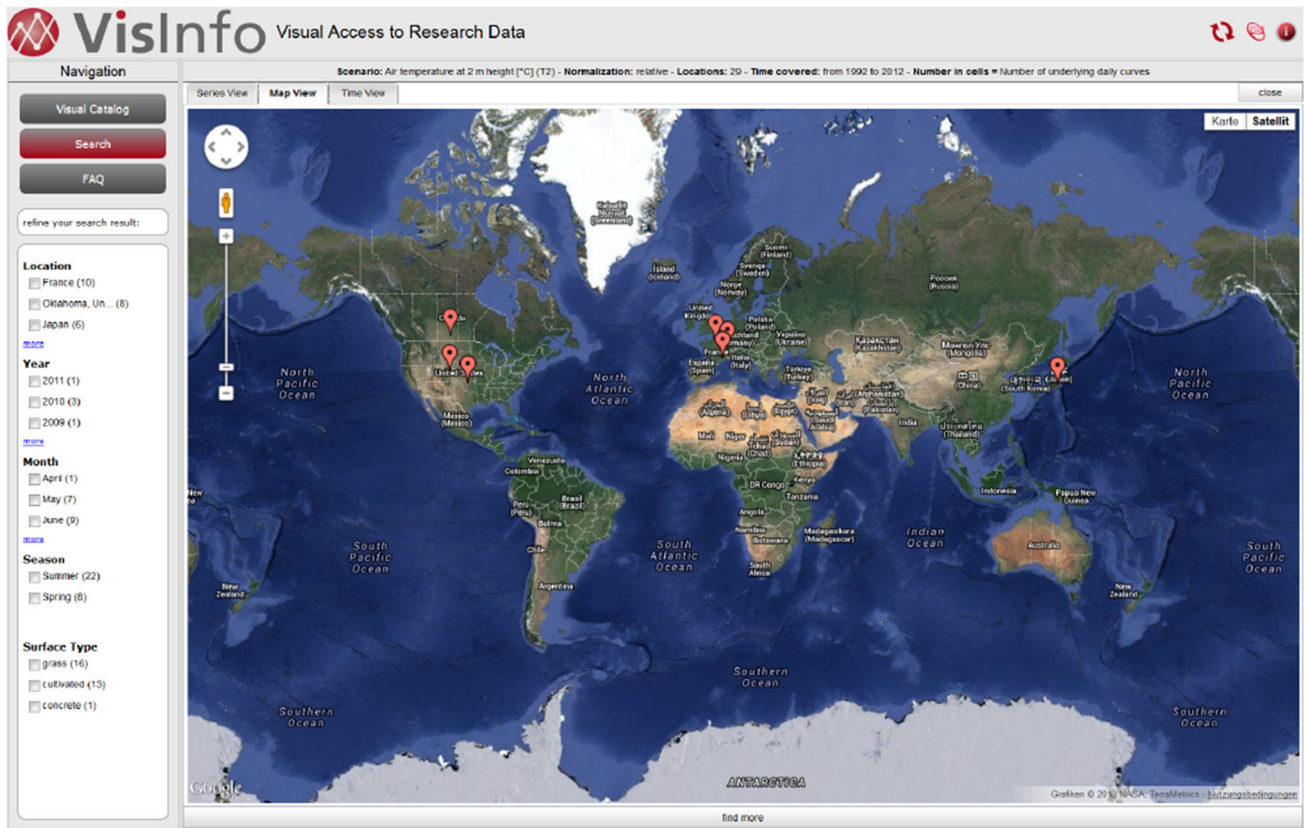


Fig. 14 Geo View. The majority of the retrieved documents are measured in the Northern hemisphere

result space. Since a content-based ES in thousands of search patterns typically retrieves large amounts of relevant items [68], the user is able to choose relevant metadata facets to filter relevant items. A decision was made in favor of 'Location', 'Year', 'Month', 'Season' and 'Surface Type'. These metadata properties are shown in the left of the Result View in a faceted search view (see Fig. 13).

After the search result exploration and a potential information gain, the user can redefine the query and start another search. Concluding the search process, the user can follow external links to the data warehouses to download the retrieved data sets.

5.5 Performance aspects

Since the VisInfo demonstrator is based on web technology, limitations in the communication bandwidth between server and client needed to be considered. The technical challenge to provide efficient search in hundred thousands of daily patterns was solved by a descriptor (a compact and yet precise representation) of the raw time series data. The PAA descriptor compresses the data by more than 98% [7]. An additional search index makes the search engine capable of real-time retrieval for up to 1,000,000 feature vectors. The index is computed via a clustering-based data aggregation

that enables search result calculations with nearest neighbor queries in under one second. The increase in speed compared to a sequential scan search algorithm is measured at 1,250% on average. We calibrated the index in such a way that the search result matches the search in more than 99.9% of all searches. The SOM-based visual data aggregation (the Visual Catalog) also helps to keep data communication at a manageable scale. Based on the ability of pre-calculation, the scalability of the data aggregation concept is independent of the size of the data set. Instead of loading large amounts of daily patterns, only the information for the visual representation of Visual Catalog is transferred initially. Relevant data content is transferred only with a detail-on-demand interaction or a query execution. The visual representation of the raw data content of the search result was another challenge to overcome. Here, the PIP algorithm reduces the number of timestamps in the raw data from about 44,000 per monthly measurement to a PIP feature vector of less than 2,000 timestamps [7]. This results in a compression rate of more than 95%.

6 Evaluation

Applying a user-centered design approach, the VisInfo system was analyzed and subsequently improved based on the

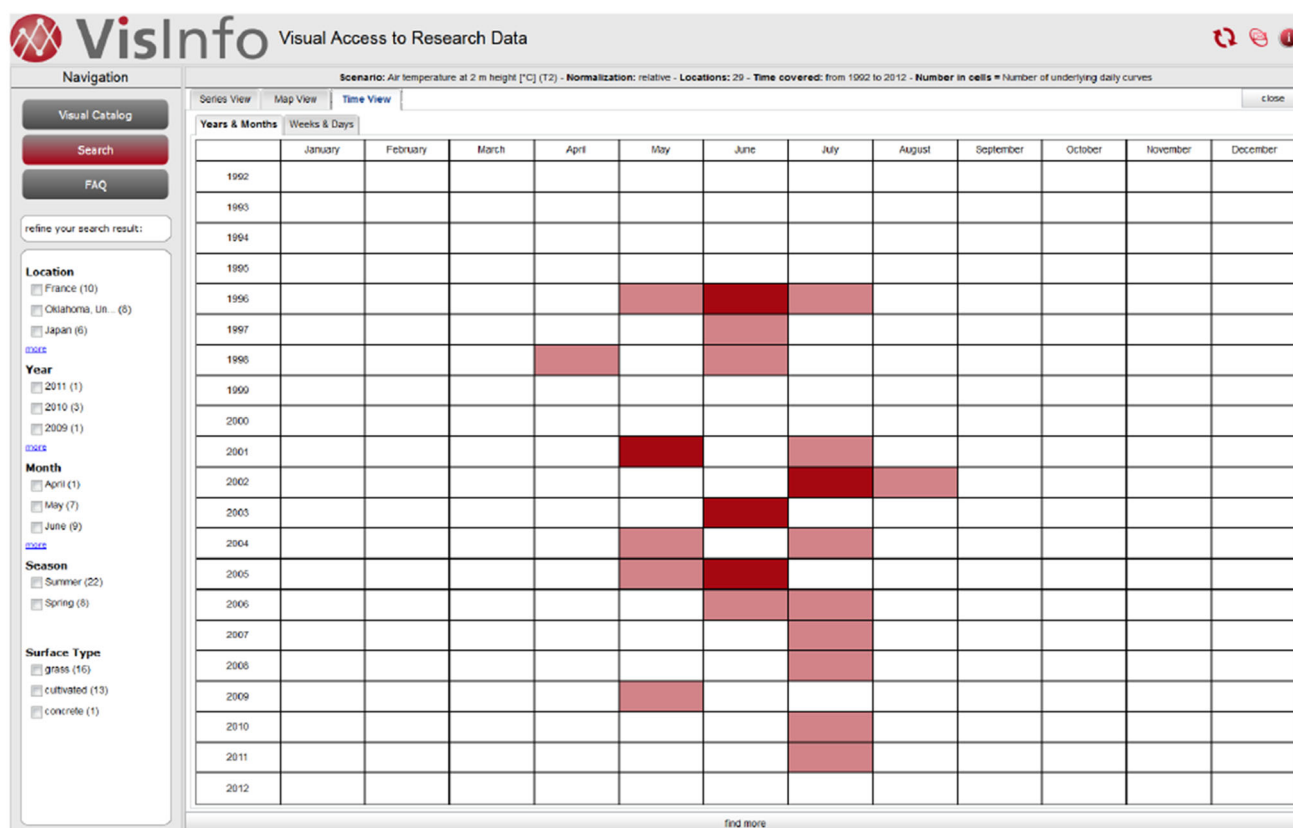


Fig. 15 Time View. The majority of the retrieved documents contain data measured in the summer period (middle set)

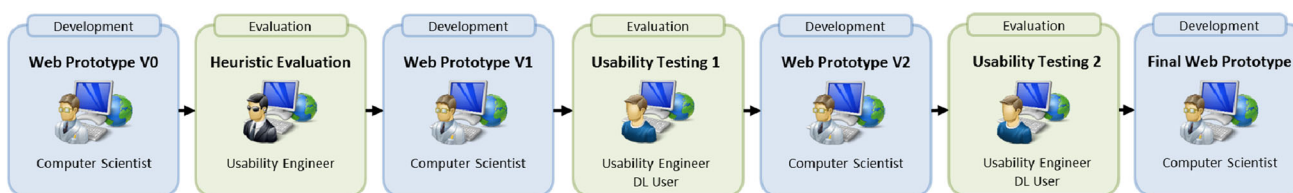


Fig. 16 Development phase of the VisInfo web prototype. Different stakeholders were involved in the iterative process

results of several evaluation rounds. An overview of this iterative design process is shown in Fig. 16.

As a first step, the initial web prototype (V0) was tested by a usability expert using heuristic evaluation methods. Based on the results, the prototype was improved by the computer scientists, resulting in the second version web prototype (V1). As further evaluation steps, two usability testing rounds with domain experts were conducted by the digital librarians and the usability expert, leading to the further evolved web prototype (V2) and the final web prototype, respectively. The tests were conducted with scientists of the earth and environmental sciences as the main target group. Therefore, representative user tasks were identified, and tasks for both usability tests were designed. Both performance and preference data were collected during the usability tests. Performance data were analyzed quantitatively (time, success rate) and qualitatively

(observation of user behavior, reactions). The same applies to preference data (quantitatively: perceived difficulty of the task; qualitatively: user comments). The identified usability problems were ranked by severity.

6.1 Heuristic evaluation

In a heuristic evaluation, a small number of evaluators, ideally usability experts, examine the interface and judge its compliance with recognized usability principles (the “heuristics”). Nielsen et al. stated, that 3–5 usability experts can usually identify up to 75% of all usability problems in a design [44]. For the first heuristic evaluation of the web prototype (V0), a discussion of the involved stakeholders was conducted, led by a usability expert.

Table 1 Tasks and their instruction text

Task	Task instruction
Task 1	Please browse through the given patterns and choose one that you are interested in. Pull up the “distance color map” for this pattern.
Task 2	Use the chosen pattern from Task 1 as query input and load it into the Sketch Editor of the Search Interface. Start the search.
Task 3	Please start a search with your own query input: Draw your own curve. Use the filters to make sure you only receive results from locations in Antarctica, North- and South-America. Look at the results using the Map View.
Task 4	Look at the results of the last task. Please refine the results further so that only results from one country (of your own choice) are shown.

As a result, the usability expert conjectured that the user would have no orientation within the system, because of unclear differentiation between interfaces for search, browsing, and search result visualization. As a consequence, the developers improved the web prototype with the objective of clear interface separations.

Since the first evaluation resulted in significantly altered interfaces and interactions, a second heuristic evaluation was conducted as a cognitive walkthrough. The user behavior was anticipated by a usability expert. After the first web prototype (V0) was implemented with a strong focus on functionality design, the next prototype (V1), after the heuristic evaluation, resulted in an improved, more intuitive user interface and interaction design anticipating common web behavior.

6.2 Usability testing

The usability tests were conducted with scientists from the earth and environmental sciences in two rounds, the first at the ‘BSRN Scientific Review and Workshop’ at Potsdam, Germany, and the second at the ‘Alfred-Wegener Institute’ in Bremerhaven, Germany. The scientists were asked to perform four tasks. The detailed task instructions can be seen in Table 1.

Task 1 addressed all major functionalities available for the cells of the Visual Catalog (exploration). Task 2 addressed the cell patterns that can be used as search input in the Search Interface (query-by-example). Task 3 covered the Sketch Editor, and therefore the concept of the visual search by drawing a curve for defining a query (query-by-sketch). Task 4 assessed the usability of the Result View and the possibilities to refine the search by using facets (result exploration). The tasks covered the most important visual encodings and the interaction designs of the VisInfo application. During the task execution, the think-aloud method was applied, to gather qualitative feedback from the participants while they were interacting with the web interface.

Table 2 Usability expert: task time and success definition

Success	Task Time	Value
Success without any difficulties	under 4 minutes	3
Success with difficulties	under 4 minutes	2
Partial failure	4 – 8 minutes	1
Failure	> 8 minutes	0

Three metrics were gathered for each of the tasks. Firstly, the *task completion time* described the time needed by the test participant to complete a task. Secondly, the *task success rate* was observed by the test facilitator, the definition is shown in Table 2. Thirdly, the *difficulty level* of each task was collected by the facilitator, based on the subjective preference of the test candidates. The scale of difficulty ranged from -3 (very difficult) to +3 (very easy). To capture a notion of the system’s usefulness, the test participants were asked to comment on whether the general concept of VisInfo is adaptable to their own research questions and whether they would recommend VisInfo to colleagues.

During usability testing round 1 in Potsdam, 21 scientists were recruited to test the VisInfo web prototype (V1, see Fig. 16), 19 scientists completed the test. The quantitative results based on the metrics mentioned above can be seen in Fig. 17. Regarding the task completion time in the first usability testing round, task 2 (covering the Search Interface) was solved quickest, while task 3 (concerning the Sketch Editor) was the longest-lasting task on average. The same result was identified with the task success rates and the difficulty level. Task 1 uncovered problems regarding the interface of the Visual Catalog. The applied colormap concept caused confusion since it was an unknown term for 10 participants (53%). Therefore, the colormap was removed in the following implementation phase. In task 2, most severities dealt with the problem that users were unsure how to use a cell pattern of the Visual Catalog as an example query and how it could be loaded into the Sketch Editor. Fig. 18 shows the preliminary version of the Sketch Editor, red exclamation marks indicate the severities of task 3. With this version of the Sketch Editor, the users were unsure how to draw a curve pattern. Task 4 revealed that the concept of faceting was new to many users. It became clear that the facet design was not intuitive.

The reflection of usability testing round 1 yielded a total of 19 usability problems. Together with frequency, their severity rating and their suggested solutions, the problems were reported to the developers of the web prototype. Moreover, the qualitative feedback of the scientists regarding the usefulness was considered.

The usability tests of the second round were conducted with the web prototype (V2) at the Alfred-Wegener Institute in Bremerhaven, Germany, with a total of six scientists. The

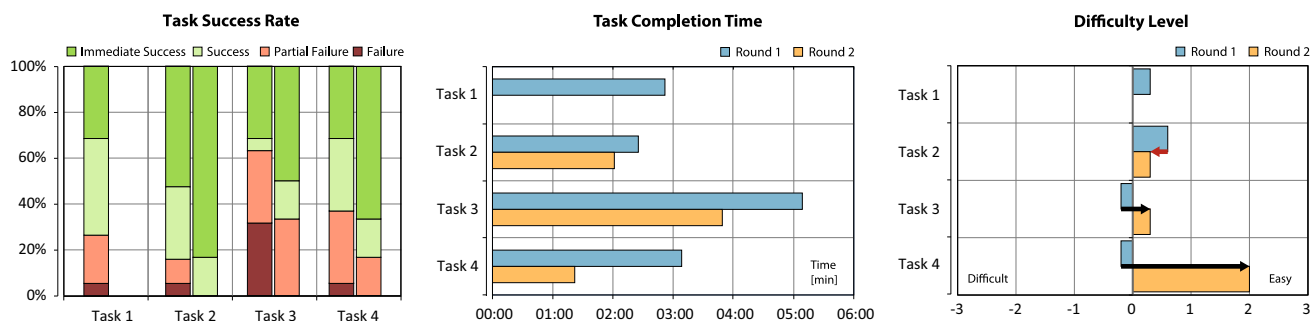


Fig. 17 Metrics gathered for the task quality assessment in the usability testing rounds

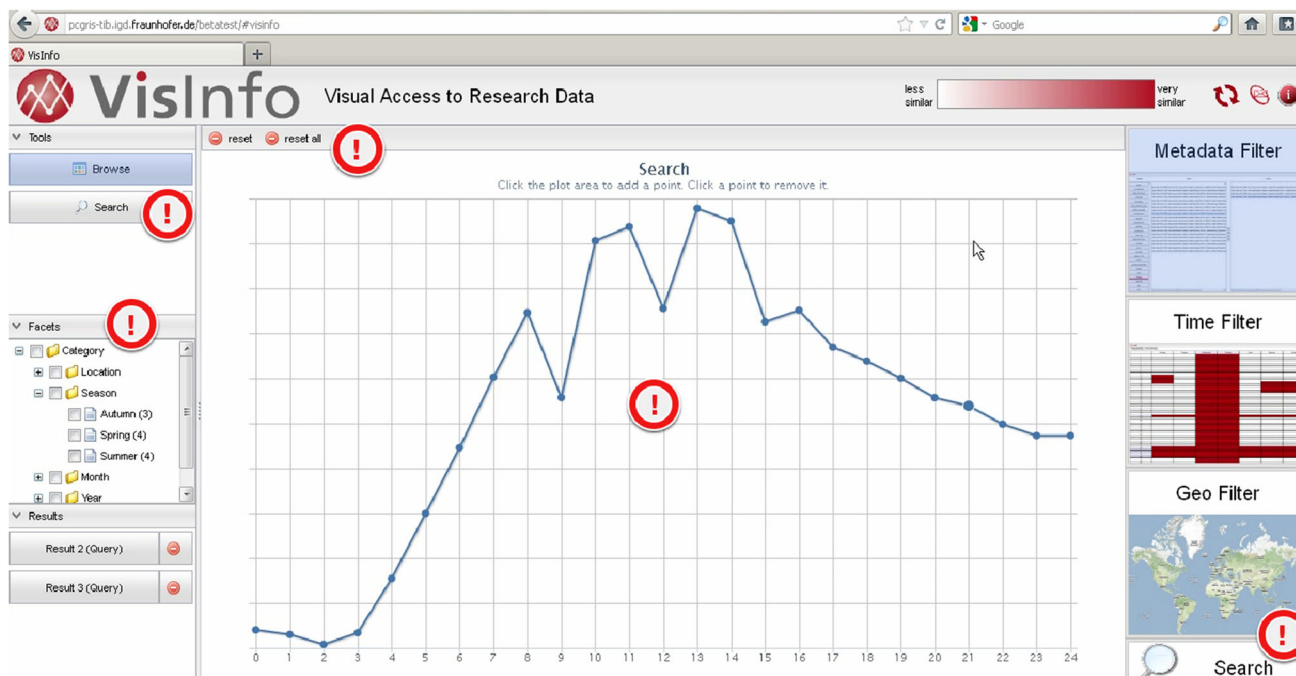


Fig. 18 The VisInfo Sketch Editor with marked usability problems

same test setup as in round 1 was utilized, except for the obsolete task 1. The quantitative results of the second testing round can also be seen in Fig. 17. In this round, nearly all tasks were completed within the 4 minute time range. Ten remaining problems have been identified in the second usability testing round. These problems were again rated by severity and frequency of occurrence. On the basis of the result of usability testing round 2, the computer scientists developed the final VisInfo web prototype.

6.3 Summary of usability testing

The most striking difference between the two usability testing rounds was that in round 2, no participant had a 'failure' (completion time > 8 minutes per task) with respect to the task success rate. This clearly underlines the improvement of the interface design since round 1. The left chart of Fig. 17

illustrates the improvement of the task success rates of each individual task between the two evaluation rounds.

In the severity ratings the occurrence of 'usability catastrophes' identified by more than 50% of the users dropped from 4 to 1. As shown at the task completion chart at the center of the figure, the average time per task decreased substantially as well. As an example, the task completion time for task 4 was reduced to less than 50%. The right chart of Fig. 17 illustrates that the users' subjective difficulty estimation of the system was lower in round 2 than it had been in round 1. Since task 1 was excluded from the second round, task 2 became the entry task of the experimental setup. This might explain the slightly increased subjective difficulty experienced for task 2. Task 3 and especially task 4, however, show a significant improvement of the difficulty level, indicating an overall increase of the system's ease of use. With an overall success rate of 83%, we verified that "70 percent of participants can meet

established successful completion criteria” [51] and therefore verified the usability, robustness, and intuitiveness of VisInfo. The system’s usefulness is reflected by the recommendation rate. While in round 1 68% of the participants would recommend VisInfo to their colleagues, the recommendation rate increased to 83% in round 2. Final comments of the users indicate the generalizability of the system’s usefulness. One user stated: “From what I did before, where I dealt with discharge curves, that [VisInfo] would have been really interesting to compare discharge curves and then see if there is similarity and maybe what kind of reason lies behind these similarities.”

6.4 Field study

After addressing the remaining usability problems derived from usability testing round 2, we made the VisInfo prototype available for the public use. Moreover, we contacted BSRN scientists, the targeted digital library users of VisInfo, and invited them to follow an introductory screencast. Then, we asked the scientists to “play” with the VisInfo prototype, and provided a questionnaire to gather feedback. The goal of this attempt was to measure the acceptance of the provided system, and the potential usefulness for the users’ daily work. The qualitative feedback showed that scientists confirmed that the proposed system supported their research interest. The usability, the robustness, and the trust in the system was rated highly positive by the participants. However, one major flaw of the system seemed to be the innovative type of the implemented concepts. As with many innovative systems, it will take some time for the users to reflect about the new possibilities such a system offers. This is also one outcome of the questionnaire, with one user stating: “Can greatly help in retrieving data based on relatively fuzzy criteria. Explaining this search would then be a major challenge. Challenges also lie in making the research reproducible.” We want to quote the gratifying feedback of a user appreciating the new system and rephrasing the advantage of VisInfo in her own words:

“If I can see time series of radiation before downloading them, I can easier find out if the station can be used for validation. This is better instead of downloading all the data before, process it, and then throw some stations out.”

7 Discussion of the approach

We next discuss our approach by a set of lessons learned, methodological aspects, and future research challenges.

7.1 Lessons learned

Trust. The applicability of our innovative search method within the domain of earth observation scientist was contro-

versely discussed. Informal interviews with our users, the earth observation scientists, in the domain characterization phase have shown that bringing visual analytics capabilities into their domain is of great value for their work. It is widely recognized that for data-driven scientific discovery, new analytical solutions are expected that realize content-based search in large collaborative DLs in future. However, many scientists prefer to analyze primarily their own data, or data they receive from colleague scientists they trust. As a result, the merits of an effective re-usage of ‘third-party’ research data have to be further emphasized. The introduction of a data curation process certainly helps to establish trust in the data collection, e.g., regarding quality, completeness, precision, etc. Moreover, using advanced visualization capability was seen in a bipolar way by some scientists. On the one hand, it was recognized that visual overviews can greatly enhance screening and comparison tasks. On the other hand, scientists were questioning the precision of the visual representations of the data. Overview visualizations work by providing abstractions and aggregations, which by their nature incur a loss of detail information. A lesson learned from that is to strongly focus on tightly integrated detail-on-demand views which provide accurate, high-resolution visual representations to further increase the trust in visually presented data.

Similarity Notion. In discussions with the users of our system, we identified a variety of analytical tasks that may be performed with the underlying research data. This had an impact on the question *what* the scientists wanted to search for. Still, we learned that scientists are often not entirely sure about *how* to formally define the respective similarity for their search. The similarity definitions introduced in our approach present a starting point for this. As scientists pointed out, they ideally want to have *many* different similarity definitions at hand, and compare the search results they obtain, to decide upon the most appropriate similarity function. Yet our assumption is that overloading the search system with too many parameters (e.g., similarity search parameters) would detriment the usability and make the learning process more complicated.

Usability. Since the system is the first of its kind in the considered user domain, a user with no prior exposure to visual search may face a barrier to entry before she can use the system effectively. Thus, one of the lessons learned is that the usability of the system needs to be well grounded. Looking back, an even more integrative approach, especially in the early design stages, would have been desirable. A lesson learned for future projects is to make early assessment of usability aspects mandatory. An assessment of usability considerations from the very beginning of the project, to reduce the users’ familiarization efforts with the final product, is

especially important in the development of innovative search concepts.

Evaluation Methods. However, it appears that it is not always possible to anticipate every usability barrier in advance for users working with unknown systems. Choosing appropriate evaluation methodologies in the course of the implementation phase was important. Based on the iterative evaluation strategy, most usability barriers could be identified and subsequently overcome. In comparison to earlier deployed versions, the final prototype was significantly improved in terms of guiding the users through the analytical workflow. The conduction of a heuristic evaluation and two usability testing rounds has proven to be useful. The collected qualitative and quantitative feedback helped to improve the system.

In summary, we feel that the project provided our entire team with a range of valuable lessons learned that will influence and enhance future work in many aspects.

7.2 Generalized project workflow description

In the following, we describe our best practice for projects with the goal to provide content-based visual access to research data. Although we try to generalize our method, we want to emphasize that we do not claim to provide a general methodology for visualization projects in the DL domain. However, we think that sharing our experiences from the VisInfo project will contribute to the DL domain, and a generalized workflow description might help later visualization projects build successful visual-interactive DLSs. Still, a general methodology for visual interfaces to DLs remains future work.

Each research project starts with a high-level goal. In contrast to standard software engineering projects starting with a requirements document by the customer and a functional specification by the developer, the goal of a research project is often not described very precisely, since the objectives are not necessarily fixed at the beginning of the project. For example, the VisInfo project started with the goal to build a DLS to provide visual access to research data. In this statement two terms had to be described in more detail: visual access and research data. These terms are connected to questions that, in general, fit to every visual DLS: (1) what is being searched for, and (2) how will this search be realized.

Considering (1), we started with characterizing the application domain and specifying the actors within this domain, respectively. In Sect. 4.1, we already defined general stakeholders in the DL context: data collector, data curator, digital librarian, computer scientist, and DL user. For each role, at least one representative should be added to the project consortium, or the advisory board. From our experience, it is of great value to have a direct connection to each of these stake-

holders when designing the DLS. As a next step, if not already given, a data collection should be defined. It is helpful if the data structure will not change during the project life time. However, the experience in VisInfo showed that the development of a general data model provides beneficial means when heterogeneous data sources or varying data standards are to be incorporated. Finally, the data structure has to be analyzed with the help of the interdisciplinary team: divide data into metadata and content data, identify most important metadata features, analyze possible features to be searched for in content data, etc. After this step, the question what is being searched for should be answered. This determination also might change over the project life time.

To answer question (2), we recommend to start with an initial design prototype. This can be realized with different techniques, from a paper prototype to a software prototype acting on a subset of the data collection. This is a crucial part of the design process, since users with no experience in visualization somehow need to get a feeling what visual search and exploration can contribute to the respective DLS. Be aware of this classical “chicken-and-egg” dilemma originating from knowledge gaps between different stakeholders [52]: on one hand, it is difficult for the computer scientist to provide a prototype without well-defined user requirements derived from domain knowledge. On the other hand, users will have difficulties in defining requirements without any visual inspiration derived from expertise in visualization designs. An early design prototype may be the basis for an enhanced requirement analysis process, e.g., by conducting observational user studies, interviews, questionnaires, or other forms of visual prototype communication. The results of this requirement analysis are the basis of the first implementation cycle, which is concluded with first usability, usefulness, and performance tests. The first version of the DLS should then be refined in further cycles until the test results satisfy digital librarians as the host of the developed system.

7.3 Future research challenges

In this project, we have established a first operational, web-based prototype for ES in time series research data, proposed for working with large data repositories. The four considered similarity notions are merely a first step which need to be extended to support further use case scenarios. The system still resembles only one part of the larger ideas of Laboratory Information Management Systems (LIMS) [26]. Among others, we identify the following research challenges:

- Advanced search interfaces should go beyond query-by-sketch or -example, but allow scientists to formulate hypotheses in other useful and appropriate ways. Examples are the specification of correlations, or lags in time and space, helping to discover interesting data sets.

- A deeper integration of the content-based search with published research papers. Techniques from Natural Language Processing could be used to automatically link textual descriptions extracted from research papers to phenomena observed in the raw data, possibly enhanced by cross-referencing the DOIs.
- We currently see a lack of collaborative aspects. The visual search system could be extended to an annotation, rating and referencing scheme and improve the exchange of data among researchers. This may include recommender functionality which correlates with, e.g., user profiles, or search sessions.

8 Conclusion

We presented the web-based VisInfo prototype, a digital library system for time series research data. Based on an overview visualization, large collections of research data are presented to the user in an exploratory manner. To support efficient retrieval, VisInfo supports the visual-interactive definition of content-based queries by sketch, or by example. Additional metadata-based filters may be applied. The visual search result presentation sheds light on the retrieved data via different views, including a geographical visualization, a calendar-based visualization, and a visual representation of the time series data content itself.

The development process of VisInfo was performed in terms of the user-centered design principle. We report on a collaborative approach between data collectors, data curators, digital librarians, computer scientists, usability experts and earth observation scientists, from our targeted user domain. The domain and problem characterization phase was conducted including paper prototyping, user interviews, screen casts and user questionnaires. The system implementation phase was accompanied by three evaluation phases which helped to iteratively improve the system. In particular, an heuristic evaluation and two usability testing rounds with expert users were conducted that helped to improve VisInfo, to result in an intuitive and usable digital library system.

Finally, we discussed lessons learned and reported on best practices for projects with the goal to create advanced visual interfaces for digital libraries incorporating information visualization and visual analytics capability. We think that this comparatively new type of digital library system approach with interdisciplinary participation can benefit from reflections on potential pitfalls and best practices.

Acknowledgments The authors would like to thank Gerd König-Langlo, Rainer Sieger, Hannes Grobe, and their group at the Alfred Wegener Institute for generously supporting the project and kindly providing data and expert feedback. Special thanks also to the Baseline Surface Radiation Network for providing the research data. The usability tests were also conducted with scientists from the BSRN community.

The participants provided helpful insights and feedback. This work was supported by a grant from the Leibniz Association as part of the “Joint Initiative for Research and Innovation” program.

References

1. Ahmed, Z., Yost, P., McGovern, A., Weaver, C.: Steerable clustering for visual analysis of ecosystems. *EuroVA*, 49–52 (2011)
2. Aigner, W., Miksch, S., Schumann, H., Tominski, C.: *Visualization of Time-Oriented Data*. Springer, London (2011). doi:[10.1007/978-0-85729-079-3](https://doi.org/10.1007/978-0-85729-079-3)
3. Ailamaki, A., Kantere, V., Dash, D.: Managing scientific data. *Commun. ACM* **53**(6), 68–78 (2010)
4. Belkin, N., Dumais, S., Scholtz, J., Wilkinson, R.: Evaluating interactive information retrieval systems. In: *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04*, p. 1594. ACM Press, New York (2004). doi:[10.1145/985921.986162](https://doi.org/10.1145/985921.986162)
5. Bernard, J., Brase, J., Fellner, D., Koepler, O., Kohlhammer, J., Ruppert, T., Schreck, T., Sens, I.: A visual digital library approach for time-oriented scientific primary data. *Springer Int. J. Digit. Libr., ECDL 2010 Special Issue* **11**(2), 111–123 (2011)
6. Bernard, J., König-Langlo, G., Sieger, R.: Time-oriented earth observation measurements from the baseline surface radiation network (bsrn) in the years 1992 to 2012, reference list of 6813 datasets. (2012). DOI doi:[10.1594/PANGAEA.787726](https://doi.org/10.1594/PANGAEA.787726)
7. Bernard, J., Ruppert, T., Goroll, O., May, T., Kohlhammer, J.: Visual-interactive preprocessing of time series data. In: Kerren, A., Seipel, S. (eds.) *SIGRAD*, vol. 81, pp. 39–48. Linköping University Electronic Press, Sweden (2012)
8. Bernard, J., Ruppert, T., Scherer, M., Kohlhammer, J., Schreck, T.: Content-based layouts for exploratory metadata search in scientific research data. In: *Proceedings of JCDL*, pp. 139–148. ACM (2012)
9. Bernard, J., Ruppert, T., Scherer, M., Schreck, T., Kohlhammer, J.: Guided discovery of interesting relationships between time series clusters and metadata properties. In: *Proceedings of i-KNOW*, pp. 22:1–22:8. ACM (2012). DOI doi:[10.1145/2362456.2362485](https://doi.org/10.1145/2362456.2362485)
10. Bernard, J., Steiger, M., Widmer, S., Lücke-Tieke, H., May, T., Kohlhammer, J.: Visual-interactive exploration of interesting multivariate relations in mixed research data sets. *Comput. Graph. Forum* **33**(3), 291–300 (2014). doi:[10.1111/cgf.12385](https://doi.org/10.1111/cgf.12385)
11. Bernard, J., Wilhelm, N., Krüger, B., May, T., Schreck, T., Kohlhammer, J.: MotionExplorer: exploratory search in human motion capture data based on hierarchical aggregation. *IEEE TVCG* **19**(12), 2257–2266 (2013). doi:[10.1109/TVCG.2013.178](https://doi.org/10.1109/TVCG.2013.178)
12. Bernard, J., Wilhelm, N., Scherer, M., May, T., Schreck, T.: Time-SeriesPaths: projection-based explorative analysis of multivariate time series data. *J. WSCG* **20**, 97–106 (2012)
13. Borgman, C., Wallis, J.C., Enyedy, N.: Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology. In: *ECDL 2006*, pp. 170–183. Springer, LINC (2006)
14. Brase, J.: Datacite - A Global Registration Agency for Research Data. In: *Cooperation and Promotion of Information Resources in Science and Technology*, 2009. COINFO '09, pp. 257–261 (2009)
15. Candela, L., Castelli, D., Ferro, N., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobrev, M. (eds.): *The DELOS Digital Library Reference model*. Foundations for digital Libraries (Version 0.98). ISTI-CNR at Gruppo ALI, Pisa (2008). <http://eprints.port.ac.uk/4104/>
16. Card, S.K., Mackinlay, J.D., Shneiderman, B. (eds.): *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco (1999)
17. Costas, R., Meijer, I., Zahedi, Z., Wouters, P.: The value of research data-metrics for datasets from a cultural and technical

- point of view. A Knowledge Exchange Report. Available from: www.knowledge-exchange.info/datametrics (2013)
18. Davenport, T.H., Patil, D., et al.: Data scientist: the sexiest job of the 21st century. *Harv. Bus. Rev.* **90**(10), 70–77 (2012)
 19. Deelman, E., Gannon, D., Shields, M., Taylor, I.: Workflows and e-science: an overview of workflow system features and capabilities. *Future Gener. Comput. Syst.* **25**(5), 528–540 (2009). doi:[10.1016/j.future.2008.06.012](https://doi.org/10.1016/j.future.2008.06.012)
 20. Duke, M., Day, M., Heery, R., Carr, L.A., Coles, S.J.: Enhancing access to research data: the challenge of crystallography. In: *Proceedings of JCDL*, pp. 46–55. ACM, New York, NY, USA (2005)
 21. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) *Advances in Knowledge Discovery and Data Mining*, pp. 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA (1996)
 22. Fu, Tc: A review on time series data mining. *Eng. Appl. Artif. Intell.* **24**(1), 164–181 (2011). doi:[10.1016/j.engappai.2010.09.007](https://doi.org/10.1016/j.engappai.2010.09.007)
 23. Fuchs, R., Hauser, H.: Visualization of multi-variate scientific data. *Comput. Graph. Forum* **28**, 1670–1690 (2009)
 24. Hearst, M.A.: *Search User Interfaces*, 1st edn. Cambridge University Press, Cambridge (2009)
 25. Herrmannova, D., Knoth, P.: Visual search for supporting content exploration in large document collections. *D-Lib Magazine* **18**, 7–8 (2012)
 26. Hey, A.J.G., Tansley, S., Tolle, K.M.: *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research (2009)
 27. Hill, L.L., Carver, L., Larsgaard, M., Dolin, R., Smith, T.R., Frew, J., Rae, M.A.: Alexandria digital library: user evaluation studies and system design. *J. ASIS* **51**(3), 246–259 (2000). doi:[10.1002/\(SICI\)1097-4571\(2000\)51:3<246::AID-ASIA>3.0.CO;2-6](https://doi.org/10.1002/(SICI)1097-4571(2000)51:3<246::AID-ASIA>3.0.CO;2-6)
 28. Hoeber, O.: User Evaluation Methods for Visual Web Search Interfaces. 2009 13th International Conference Information Visualisation pp. 139–145 (2009). doi:[10.1109/IV.2009.21](https://doi.org/10.1109/IV.2009.21)
 29. Hull, D., Pettifer, S.R., Kell, D.B.: Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput. Biol.* **4**(10):e1000204 (2008)
 30. Jeng, J.: Usability assessment of academic digital libraries: effectiveness, efficiency, satisfaction, and learnability. *Libri* **55**(2–3), 96–121 (2005). doi:[10.1515/LIBR.2005.96](https://doi.org/10.1515/LIBR.2005.96)
 31. Jeng, J.H.: Usability of digital libraries. In: *Proceedings of JCDL*, p. 407. ACM Press, New York, NY, USA (2004). doi:[10.1145/996350.996473](https://doi.org/10.1145/996350.996473)
 32. John, B.E.: Evaluating usability evaluation techniques. *ACM Comput. Surv.* **28**(4es), 139-es (1996)
 33. Kehrer, J., Hauser, H.: Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE TVCG* **19**(3), 495–513 (2013). doi:[10.1109/TVCG.2012.110](https://doi.org/10.1109/TVCG.2012.110)
 34. Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. (eds.): *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, Aire-la-Ville (2011)
 35. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. *J. Knowl. Inf. Syst.* **3**, 263–286 (2000)
 36. Kohonen, T., Schroeder, M.R., Huang, T.S. (eds.): *Self-Organizing Maps*, 3rd edn. Springer-Verlag New York Inc, Secaucus, NJ, USA (2008)
 37. Lewis, C., Polson, P.G., Wharton, C., Rieman, J.: Testing a walk-through methodology for theory-based design of walk-up-and-use interfaces. In: *Proceedings of the SIGCHI*, pp. 235–242. ACM Press, New York, NY, USA (1990). doi:[10.1145/97243.97279](https://doi.org/10.1145/97243.97279)
 38. Liao, T.W.: Clustering of time series data—a survey. *Pattern Recognit.* **38**, 1857–1874 (2005)
 39. Ludäscher, B., Lin, K., Bowers, S., Jaeger-Frank, E., Brodaric, B., Baru, C.: Managing scientific data: from data integration to scientific workflows. *Geoinformatics: Data to Knowledge*, Geolog. Society of America Special Paper 397 pp. 109–129 (2006)
 40. Marchionini, G.: Exploratory search: from finding to understanding. *Commun. ACM* **49**(4), 41–46 (2006)
 41. Marcial, L.H., Hemminger, B.M.: Scientific data repositories on the web: an initial survey. *JASIST* **61**(10), 2029–2048 (2010)
 42. Munzner, T.: A nested model for visualization design and validation. *IEEE Trans. Vis. Comput. Graph.* **15**(6), 921–928 (2009). doi:[10.1109/tvcg.2009.111](https://doi.org/10.1109/tvcg.2009.111)
 43. Nielsen, J.: The usability engineering life cycle. *Computer* **25**(3), 12–22 (1992). doi:[10.1109/2.121503](https://doi.org/10.1109/2.121503)
 44. Nielsen, J., Landauer, T.K.: A mathematical model of the finding of usability problems. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI '93*, pp. 206–213. ACM Press, New York (1993). doi:[10.1145/169059.169166](https://doi.org/10.1145/169059.169166)
 45. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People—CHI '90*, April, pp. 249–256. ACM Press, New York (1990). DOI doi:[10.1145/97243.97281](https://doi.org/10.1145/97243.97281)
 46. Nocke, T., Sterzel, T., Böttinger, M., Wrobel, M.: Visualization of climate and climate change data: an overview. *Digital Earth Summit on Geoinformatics 2008, Tools for Global Change Research*, pp. 226–232. Herbert Wichmann, Heidelberg (2008)
 47. Ohmura, A., Dutton, E.G., Forgan, B., Fröhlich, C., Gilgen, H., Hegner, H., Heimo, A., König-Langlo, G., McArthur, B., Müller, G., Philipona, R., Pinker, R., Whitlock, C.D., Dehne, K., Wild, M.: Baseline surface radiation network (BSRN/WCRP): new precision radiometry for climate research. *Bull. Am. Met. Soc.* **79**, 2115–2136 (1998)
 48. PANGAEA—Data Publisher for Earth and Environmental Science: <http://www.pangaea.de/>. Accessed 14 Oct 2014. doi:[10.1594/pangaea](https://doi.org/10.1594/pangaea)
 49. Plaisant, C.: The challenge of information visualization evaluation. In: *Proceedings of the Working Conference on Advanced Visual Interfaces*, pp. 109–116. ACM Press, New York (2004). doi:[10.1145/989863.989880](https://doi.org/10.1145/989863.989880)
 50. Reeves, T.C., Buhr, S., Barker, L.: Evaluating digital libraries: toward evolution of concepts. In: *Proceedings of JCDL*, vol. 49, p. 420. ACM Press, New York (2005). doi:[10.1145/1065385.1065525](https://doi.org/10.1145/1065385.1065525)
 51. Rubin, J., Chisnell, D.: *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, 2nd edn. Wiley, Indianapolis (2008)
 52. Ruppert, T., Bernard, J., Kohlhammer, J.: Bridging knowledge gaps in policy analysis with information visualization. In: *EGOV/ePart Ongoing Research*, LNI, vol. 221, pp. 92–103. GI (2013)
 53. Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* **11**(5), 561–580 (2007)
 54. Saracevic, T.: Digital library evaluation: toward evolution of concepts. *Libr. Trend* **49**(2), 350–369 (2000)
 55. Scherer, M., Bernard, J., Schreck, T.: Retrieval and exploratory search in multivariate research data repositories using regression features. In: *Proceedings of JCDL*, pp. 363–372. ACM, New York (2011)
 56. Schreck, T., Bernard, J., Tekušová, T., Kohlhammer, J.: Visual cluster analysis of trajectory data with interactive Kohonen maps. *Palgrave Macmillan Inf. Vis.* **8**, 14–29 (2009). <http://www.palgrave-journals.com/doi/10.1057/ivs.2008.29>
 57. SDSS—The Sloan Digital Sky Survey. <http://www.sdss.org/>. Accessed 15 Oct 2014
 58. Sedlmair, M., Meyer, M., Munzner, T.: Design study methodology: reflections from the trenches and the stacks. *IEEE Trans. Vis. Comput. Graph. (Proc. InfoVis)* **18**(12), 2431–2440 (2012)
 59. Shackel, B.: Usability context, framework, definition, design and evaluation. In: Shackel, B., Richardson, S.J. (eds.) *Human Factors for Informatics Usability*, pp. 21–37. Cambridge University Press, New York (1991)

60. Sheble, L.: Greenstone User and Developer Survey 2009 (2009). <http://greenstonesurvey.wordpress.com/infosheet/>
61. Stein, L.D.: Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat. Rev. Genet.* **9**(9), 678–688 (2008)
62. Strötgen, J., Gertz, M.: Event-centric search and exploration in document collections. In: *Proceedings of JCDL*, pp. 223–232. ACM, New York (2012)
63. Suber, P.: *Open Access*. MIT Press, Cambridge (2012)
64. Tominski, C., Donges, J.F., Nocke, T.: Information visualization in climate research. In: *Information Visualisation*, pp. 298–305. IEEE Computer Society, Washington (2011)
65. Tsatsaronis, G., Varlamis, I., Torge, S., Reimann, M., Nørvåg, K., Schroeder, M., Zschunke, M.: How to become a group leader? Or modeling author types based on graph mining. In: *Proceedings of TPD*, pp. 15–26. Springer, Berlin (2011)
66. Van Wijk, J., Van Selow, E.: Cluster and calendar based visualization of time series data. In: *Symposium on Information Visualization*, pp. 4–9. IEEE Comp. Soc. (1999)
67. VisInfo—Visual Access to Time-oriented Research Data. <http://demo.vis-info.info/>. Accessed 14 Oct 2014
68. White, R., Roth, R.: Exploratory search: beyond the query-response paradigm. *Synth. Lect. Inf. Concepts Retr. Serv.* **1**(1), 1–98 (2009)
69. Yuan, X., Zhang, X., Trofimovsky, A.: Testing visualization on the use of information systems. In: *Proceedings of IliX*, p. 365 (2010). doi:[10.1145/1840784.1840840](https://doi.org/10.1145/1840784.1840840)
70. Ziegler, H., Jenny, M., Gruse, T., Keim, D.: Visual market sector analysis for financial time series data. In: *Proceedings of VAST*, pp. 83–90 (2010). doi:[10.1109/VAST.2010.5652530](https://doi.org/10.1109/VAST.2010.5652530)