Erschienen in: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries : June 10 - 14, 2012, Washington, DC, USA / Karim B. Boughida ... (eds.). - New York, NY : ACM, 2012. - S. 139-148. - ISBN 978-1-4503-1154-0

Content-Based Layouts for Exploratory Metadata Search in Scientific Research Data

Jürgen Bernard
Fraunhofer Institute for Computer Graphics Research
Fraunhoferstr. 5
D-64283 Darmstadt, Germany
juergen.bernard@igd.fraunhofer.de

Tobias Ruppert
Fraunhofer Institute for Computer Graphics Research
Fraunhoferstr. 5
D-64283 Darmstadt, Germany
tobias.ruppert@igd.fraunhofer.de

Maximilian Scherer
TU Darmstadt
Fraunhoferstr. 5
D-64283 Darmstadt, Germany
maximilian.scherer@gris.tu-darmstadt.de

Jörn Kohlhammer
Fraunhofer Institute for Computer Graphics Research
Fraunhoferstr. 5
D-64283 Darmstadt, Germany
joern.kohlhammer@igd.fraunhofer.de

Tobias Schreck
University of Konstanz
Universitätsstr. 10
D-78457 Konstanz, Germany
tobias.schreck@uni-konstanz.de

ABSTRACT

Today's digital libraries (DLs) archive vast amounts of information in the form of text, videos, images, data measurements, etc. User access to DL content can rely on similarity between metadata elements, or similarity between the data itself (content-based similarity). We consider the problem of exploratory search in large DLs of time-oriented data. We propose a novel approach for overview-first exploration of data collections based on user-selected metadata properties. In a 2D layout representing entities of the selected property are laid out based on their similarity with respect to the underlying data content. The display is enhanced by compact summarizations of underlying data elements, and forms the basis for exploratory navigation of users in the data space. The approach is proposed as an interface for visual exploration, leading the user to discover interesting relationships between data items relying on content-based similarity between data items and their respective metadata labels. We apply the method on real data sets from the earth observation community, showing its applicability and usefulness.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3 [**Information Storage and Retrieval**]: Digital Libraries; J.2 [**Physical Sciences and Engineering**]: Earth and atmospheric sciences

Keywords

Information Visualization, Exploratory Search, Similarity Measures, Scientific Research Data, Visual Layouts

1. INTRODUCTION

Digital Libraries (DLs) organize, preserve, and make available to users all kinds of information. Often, textual and multimedia documents such as images, video, or audio are among the document types supported by DL systems. Recently, also *scientific research data* has come into focus as an important type of information that should be treated by DL efforts. New technologies for data collection are leading to data production at high rates, giving rise to large amounts of relevant data which users may be interested in. Examples, among many others, include the scientific domains of earth and space observation, where repositories such as PANGAEA [26] and SLOAN [31] host large amounts of relevant data, respectively. Jim Gray's *Fourth Paradigm* [15] suggests that the scientific discovery process as a whole could benefit tremendously if research data could be consistently collected, shared, and made available by means of a research data infrastructure.

To date, the data storage facilities in DLs have increased substantially, and the storage of large data becomes less an issue than appropriate forms of user access. Data in existing scientific data repositories are typically accessed by users according to metadata properties of the data items, e.g., time or location of observation, or the name of the creator of the respective data item. User access based on the actual *content* of the data, however, remains a difficult problem as the concept of data is very diverse, and appropriate search methods depend on the type of data and user intents.

The visual analysis has the goal to make big amounts of data and information processing transparent in a way that combines the strengths of humans *and* computers. Actual approaches unify automated data analysis with visual-interactive data exploration [18]. Exploratory search [38]

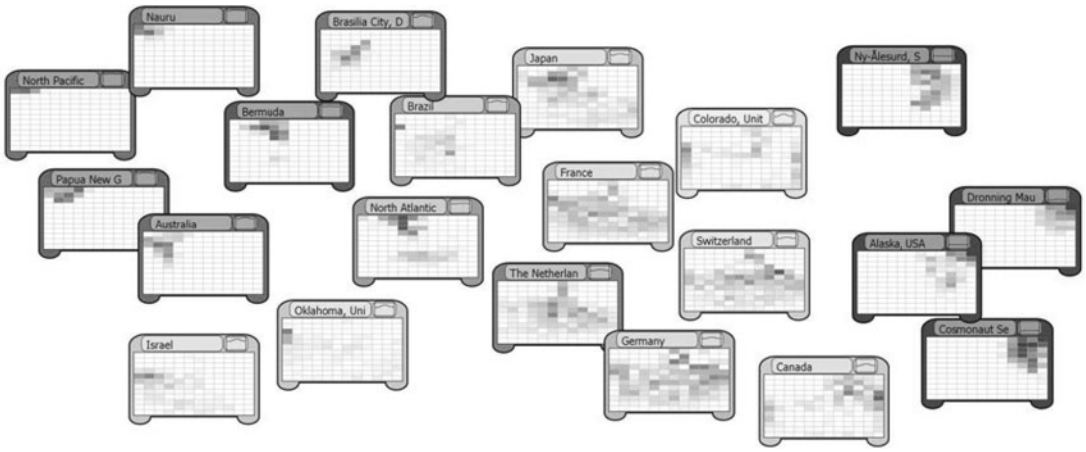

Figure 1: Graphical layout of metadata property 'Location' enriched by data content summaries based on temperature measurements. Hot temperature measurements to the left (red), measurements from polar regions to the right (blue), cluster of moderate temperatures at the center (yellow, green).

addresses the problem of users not having a well-defined information need that could be translated into a formal query. Appropriate overview and navigation facilities allow users to obtain a better understanding of the overall data space, before narrowing down the search to more specific queries. Facetted search approaches have proven to be very useful in the exploratory search process. They group documents according to categories of metadata, such as 'Author', 'Year', or 'Location'. The appropriate representation of relationships and patterns in data can lead to the discovery of expected and unexpected relationships [42] potentially relevant for the users' information need.

We propose a new approach for exploratory search in time-oriented research data. It reflects both metadata and the actual data content. A visual-interactive display is designed which allows users to navigate the metadata search space, discover relationships between data items, and access individual data records. Based on the notion of *metadata properties*, users are enabled to select a specific metadata field of interest. The similarity between instances of the selected metadata property (we call them *metadata entities*) is automatically computed based on the content of the underlying data items. A similarity-preserving 2D layout of the metadata entities, based on a force-directed network layout method, is then generated. This layout can be freely navigated by the user to explore interesting relationships between entities. The visual exploration is supported by showing, for each entity, a content summary of the represented data items based on the clusters within the data.

Our approach is applied to a repository of time-series data from the earth environment domain. We show the principal applicability of the method and recommend it as an approach for visual and exploratory search in this kind of data domain. Examples of the many useful application scenarios made possible by our approach include finding answers to questions such as "which data creator found the earth observation measurements most similar to mine?" or "which measurement stations produce the most similar measurement series?". In a case study, we show the practical applicability of our search and exploration display by incorporating scientific researchers from the earth observation field.

2. RELATED WORK

2.1 Scientific Research Data in the Digital Library Context

In modern information infrastructures, DLs are a key component. DLs provide and manage facilities for accessing information and data resource, and allow distributed access to large user communities. DL systems have developed over the years to a 'digital library universe' with a great variety of user roles, resources, technology and relationships, and yield many advantages in modern information infrastructure [22, 40]. Many DL systems provide mechanisms to search in well-defined metadata. Emerging nominatives for metadata properties are the Dublin Core standard [27]. Many other domain-focused metadata standards exist. For example, in research data, the DataCite metadata kernel [33] is proposed. Open data libraries cover a number of data and application domains like crystallography [13], earth observation [26], or chemistry [39]. A clear trend indicates that the number and size of public data repositories will continue to grow. Respective challenges in the DL area include the question how to integrate and access that data to support scholarly work flows.

DLs are fundamental components of E-science [15, 32], and the necessity of treating scientific research data by library services is generally recognized [17]. Moreover, significant challenges like handling complex data in the sense of (a) 'big data' and (b) heterogeneous data exist [19, 18]. Furthermore, a challenge with scientific research data retains: how to organize and access various kinds of scientific data content? In Section 2.3 we will identify the problem of creating content summary solutions in the scientific research data context. Hull et al. expect future DL tools with metadata and data content to be less isolated and rigid [17]. Based on our experience and a review of DL publications distributed among various scientific data domains, it is of invaluable importance to work closely together with researchers to appropriately meet the two-way requirements on target DL analysis systems. This need for multidisciplinary collaboration is stated in biological sciences [32], neuroscience [10],

E-Science [16] and other application areas [6, 13]. Making effective use of data sets based on a combined metadata-based and content-based approach is promising [3]. Like stated by Foster et al., understanding the scientific workflow and providing effective information tools has the potential to eventually realize massive benefit to science [17].

2.2 Similarity relations generated by metadata and content-based approaches

When are two entities related or similar? This is the driving question in related work about metadata-based and content-based approaches.

In *metadata*-based approaches, annotations primarily serve as a measure for similarity. If two annotations are similar, the annotated entities are considered similar [27]. Such metadata approaches rely heavily on the quality of the metadata – and accordingly on the curation process. Another metadata-based approach relies on usage statistics. Such approaches are most popular in collaborative filtering systems, where the assumed relatedness (or similarity) of two entities is increased, whenever they are being bought together, looked at the same time, have the same co-citation patterns etc. [5]. Hybrid algorithms employing both – content-based and collaborative filtering techniques – have also been successfully used [34]. A third approach comprises ontology-driven techniques and systems. Here, a strict ontology is imposed for a given domain. Under such constraints, retrieval with a focus on semantics becomes viable. Such approaches already started to enable new semantic applications in a wide span of areas such as bioinformatics, financial services, web services, business intelligence, DLs and national security [19]. Two of the most prominent examples include DBPedia [2], which enables semantic search of wikipedia knowledge, and Wolfram Alpha [41] which even allows for semantic searches with natural language querying.

Measuring *content*-based similarity is subject to research in many areas. These include multimedia information retrieval tasks [23], like 2D shape analysis, 3D object retrieval and content-based image retrieval [11], as well as information retrieval in time-series data [21]. This process usually involves (a) some kind of descriptor that represents data under concern, and (b) a distance measure between two descriptors that represent the dissimilarity of the data object. One prominent descriptor-scheme is the computation of feature vectors. For computing the distance between two feature vectors, a vast amount of distance functions is available [9]. Given a descriptor and a distance measure, users are allowed to search for data objects not only by similarity of the annotation, but also by similarity of content. Such queries often consist of query-by-example or query-by-sketch [14].

A key difference between content-based and metadata-based similarity notions is that the former can be computed fully automatically, but often suffer from limited discrimination capability and in general, the semantic gap problem [29]. Metadata-based approaches in the best case can provide better discrimination and semantic description, yet are often depending on manual annotation and quality control. Most retrieval and exploration systems consider either the content-based or the metadata-based similarity notions for supporting user queries or computing visual layouts for explorative search and browsing. Our work is novel in that we integrate both similarity notions in a joint approach. Specifically, we compute content-based similarity between data items, using it to visually map metadata properties of the data, for explorative analysis and correlation.

2.3 Visualization of Search and Exploration Spaces

There exist two classes of techniques for visualizing data entities on a 2D display reflecting their interdependencies and thereby generating a visual entity map. These are projection-based and graph-based layouts. The projection-based approaches map high-dimensional data spaces to spaces of lower dimensionality. In our case, the high-dimensional time-series data is projected to a space of lower dimensionality in order to visualize it in 2D. Examples for projection-based approaches include principal component analysis (PCA), multi-dimensional scaling (MDS) or self-organizing map (SOM) [12] layouts. All of these techniques are topology-preserving. Their goal is to preserve the pairwise distances between its entities, which is the main requirement for our 2D map. As a drawback of these projection-based layouts, their output is neglecting the overfitting problem on the display space. Entities of similar content will overlap in the visualization, which is a drawback for the user's exploration task.

As a second possibility, data entities and their pairwise similarities can be interpreted as complete graphs. Hence, the visual map can be generated via a graph-based layout. For the preservation of the edge lengths especially force-based graph layouts are suitable. An overview of graph layouts is given in [37]. In our approach, we apply the Weighted Edge-Repulsion LinLog model, an extension of the Edge-Repulsion LinLog model [24], to layout our metadata entity graph. Besides their topology-preserving characteristic, it addresses the overfitting problem by applying repulsive forces on overlapping entities.

There also exist metadata visualization approaches in the field of DL. In [35], powergraphs are used to visualize clustered co-author relationships from a bibliographical database. The experimental library software INVISQUE [42] uses an index card metaphor to realize a visual interactive exploration of library content. While approaches in providing 'content summary solutions' for generic data types like audio or image, or other multimedia data exist for years [28, 1], concepts for organizing scientific data by its content in the DL workflow are scarce. In [30], content summaries are provided via a k-means clustering approach. As another example, Bernard et al. [3] generate content summaries based on scientific research data via a self-organzing maps approach. A solution for visualizing icon-based cluster content summaries combined with graph layouts can be found in [8] from the information visualization research field.

In the information visualization field, mapping of data variables on the display space is often performed by means of visual attributes like color, transparency, object size, or object position. The use of color needs predefined color palettes, so called color maps. In our approach we apply color on the basis of a so called visual catalog to visually discriminate between distinct variations of data content. Therefore, a variety of two dimensional color maps can be applied [7]. We follow the idea of Vesanto et al. [36] where the output of the self-organizing map algorithm is color-coded.

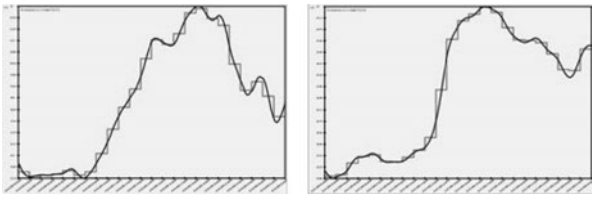

Figure 2: Visualization of time-oriented research data content. Daily temperature curves (black) with a similar visual appearance are shown, while absolute values differ (left: around 20°C, right: below -20°C). PAA descriptor of the curves (red) denotes the basis of our content-based similarity concept.

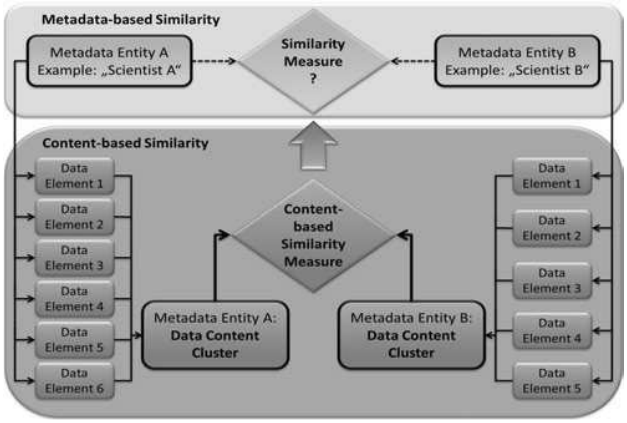

Figure 3: Content-based similarity for two given metadata entities: Measurement data of 'Scientist A' is compared with measurement data of 'Scientist B' to compute a content-based distance between these two (semantic) entities.

3. APPROACH

In this section we present our method for visual search and exploration of metadata properties. In current DL systems metadata often refers to nominal entities (authors, geo-annotation, sources, etc.) As a consequence, potential relations between different entities must be neglected in a metadata search: Only the library objects matching the name defined in the query can be considered as a valid result. For example, searching for a specific author will result in a list of all library objects composed by the author. Relations between different metadata entities (e.g., different authors) are neglected.

In the following Section 3.1 we propose a similarity measure in order to describe the relations between different metadata entities. This measure is computed based on the data content associated to the metadata property. In our use case, the content consists of time-oriented measurements taken by the author. The result of these content-based similarity computation is a distance matrix describing all pairwise similarities between the metadata entities. In Section 3.2, this distance matrix is visualized for a first comparison of different metadata entities regarding their content. In Section 3.3, we provide visual access to the metadata entity relations via a 2D layout. Here, authors that created similar measurements appear close to each other. In Section 3.4 we detail on our visual catalog which allows for a global overview of the data content. Moreover, we employ this visual catalog metaphor as a visual representation to enrich each of our metadata entity glyphs with associated data content information, see Section 3.5.

3.1 Content-based Metadata Similarity Measures

In our approach, we cluster entities based on similar metadata annotations like 'Author' or geo-spatial 'Location'. Given such a clustering, we wish to compute the content-based similarity between these semantic entities. In the application we consider here, the content of each entity consists of time-oriented climate measurements series.

A simple, yet powerful technique to describe sequential data is the so-called Piecewise Aggregate Approximation (PAA) [20] (see Figure 2). It is suitable for our purposes of describing time-series. The basic idea is to split a sequence of length n into m segments and compute the mean value of all data-points in each segment. Such a block-wise average can be computed extremely fast, only n (sequence length) additions and m (number of segments) divisions are required.

To measure content-based similarity we chose the Euclidean distance, because the PAA descriptor tries to approximate the original time-series as closely as possible. This avoids deviating any further from the true Euclidean distance of two given time-series.

However given two sets of descriptors and a distance measure, there are still several possibilties to assess the similarity. We can (a) compute the average of each pair-wise distance; (b) average the descriptors for each set and then compute the distance; (c) determine the median descriptor for each set and compute their distances.

We chose option (b), because we believe this to be the best trade-off between the discriminativeness of (a) and the outlier robustness of (c) (see Figure 3).

Please note that the choice of data descriptor and distance function is a user-parameter in general. As long as the descriptor computation result is a vector and the distance measure is a true metric, the particular concept of content-based similarity is interchangeable.

3.2 DistanceMatrixView: Visualizing Pairwise Distances of Metadata Entities

The similarity measures presented in Section 3.1 are used to identify relations between different metadata entities. A matrix consisting of all pairwise similarities between the metadata entities of a certain metadata property (e.g. 'Author') can be interpreted as a distance matrix. In Figure 4 we present a triangular distance matrix visualization, that supports the visual exploration of these similarities. Each diamond-shaped box represents the relation between two metadata entities. The color describes the similarity value from similar (white) to dissimilar (blue). We provide two possible scenarios for the user to interact with the visualization. First, similar or dissimilar metadata entities can be detected by clicking on bright or dark blue diamonds. The corresponding metadata entities are highlighted in the list (see Figure 4). With this interaction mode, extreme similarity values can be explored. Secondly, a user may want to

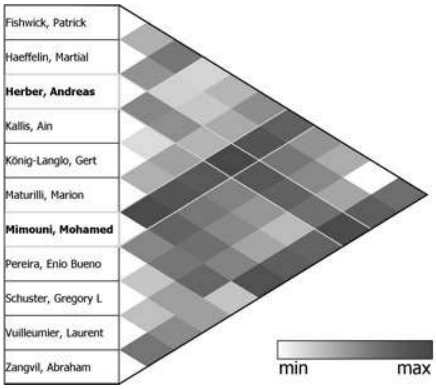

Figure 4: DistanceMatrixView: Visualization of a similarity matrix regarding metadata property 'Author'. Similarities are computed based on shortwave downward radiation measurements, which are prominent measures in climate research, especially for giving statements about cloud occurences. Highlighting of lowest similarity between the measurements of the two authors 'Andreas Herber' and 'Mohamed Mimouni'.

identify the similarities to one specific metadata entity. By selecting this entity in the list, the corresponding diamonds are highlighted. This helps the user to analyze the similarities of one specific metadata entity to the other entities (e.g, see Figure 7). The distance matrix visualization allows the user to visually explore the metadata space implying its inner relations regarding a specific metadata property (e.g. 'Author'). Please note, that the user can freely choose which metadata property to explore. Our concept can be applied to any metadata property provided in the data set. The properties only define the grouping of the data content.

3.3 MetaMap: Visual Mapping of Metadata Entities

Another method to support exploratory search in the metadata space is to display the metadata entities to be explored in a 2D map. One requirement to the resulting visualization is (a) that similar entities are depicted close to each other while dissimilar entities should have a greater visual distance. That way, the user can explore similar metadata entities in the vicinity of an entity of attention. Moreover, (b) the visualization has to arrange the icons representing the metadata entities with a minimum of overplotting, since the nearest neighbors of a metadata entity are most relevant for the exploration task.

From the visual mapping presented in the related work (see Section 2.3) the force-based graph layouts prove to fulfill both requirements for the visual mapping. Therefore, the metadata entities and their pairwise similarities are interpreted as nodes and edges respectively.

In our approach, we apply the Weighted Edge-repulsion LinLog model to layout our metadata entity graph. It is topology-preserving, and therefore, meeting requirement (a). Moreover, it addresses the overfitting problem, requirement (b).

In the following the metadata entity map is enriched by visual content summaries.

Figure 5: Visual layout of metadata entities represented by their content summary icons. Topology preservation (e.g., lowest similarity between authors 'Andreas Herber' and 'Mohamed Mimouni') and minimization of overlapping is shown.

3.4 Visual Catalog to Summarize the Data Content

So far, we introduced the computation of similarity measures for metadata entities based on their underlying data content, and two possible ways for visualizing the metadata entities and their similarity values. In both visualizations (presented in Section 3.2 and 3.3) the associated data content that is used to compute the metadata similarities is not visualized. Now, we present a visualization method for giving the user a 'global overview' of the underlying data content. In our use case, the data content consists of time-oriented measurement curves, each with the duration of one day. We make these daily patterns visually accessible via a visual catalog metaphor. This visual catalog has to (a) represent the underlying curve patterns, and (b) arrange them in an intuitive order showing similar curve patterns close to each other.

In our approach, we use a SOM algorithm to generate a visual catalog. The resulting visual catalog (see Figure 6 (left)) consists of a $n \times m$ grid of cells, each representing a cluster of curve patterns. Within each cell, a representative curve pattern and the number of curves contained in the cluster is depicted. The cells are automatically arranged in a way, that similar curve clusters appear close to each other. With these characteristics our visual catalog meets both requirements, visual overview and topology preservation. Please note, that the grid should not be interpreted as a coordinate system with a semantic meaning for both coordinate axes. The visual catalog is just a topology-preserving arrangement of clusters.

As a last step for generating our visual catalog, a color map is applied to the SOM grid. The purpose of this color map is to increase the recognition value of special areas, which is relevant for the content summaries in Section 3.5. Please note, that the color map does not depend on the data content. It is just applied to the grid structure of the visual catalog. This is important to visually discreminate different areas of the visual catalog. The color map can freely be chosen and adapted to specific use case scenarios. For example, in the application part of this paper (see Section 4), we choose a color map that supports the notion of warm (red) and cold (blue) temperatures.

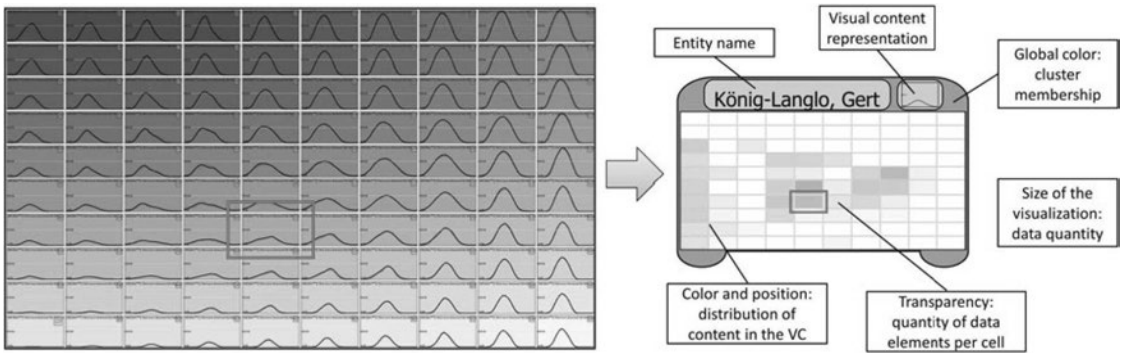

Figure 6: Left: Visual catalog of daily curve progressions generated by a SOM algorithm. Underlying data consists of shortwave downward global radiation measurements. Each cluster cell shows a representing curve pattern, the number of associated curves is depicted at the upper right, respectively. Independent of the actual data, a color map is applied to the 10×10 grid. Right: Content summary visualization representing the measurements of author 'Gert König-Langlo'. Transparency is applied to highlight associated measurements. Global color and visual content representation show the mean curve progression and cluster cell affiliation.

3.5 Enriching our Metadata Entity Glyphs with Visual Content Mappings

In the last section we introduced a visual catalog to give an overview of the data content, consisting of daily curve progressions. Now, we will combine the metadata map (see Section 3.3) with the visual catalog to obtain an integrated visualization depicting both metadata and content.

Therefore, each metadata entity in the metadata map is represented via visual content representations, we call them content summaries. A content summary consists of the visual catalog, presented in Section 3.4, with a color map adjusted to the corresponding metadata entity (see Figure 6 (right)). For example, each 'Author XY' has a certain number of associated measurement curves. These measurement curves are highlighted in the visual catalog to obtain a content summary for the respective author. That means that areas in the visual catalog with a low number of occurences regarding the associated measurement curves are visualized with a higher transparency on the color map.

With this method for each metadata entity a content summary represented by a visual catalog reflecting the occurences of measurement curves is calculated (see Figure 6 right). Now, these content summaries are visualized in the metadata map introduced in Section 3.3 (see Figure 5).

With this visualization the user can directly identify which metadata entities are similar and how their similarity can be interpreted regarding the data content. As an additional feature, the content summary border is depicted with the color most representative with respect to the visual catalog. With this method we provide a content summary solution that supports the user in the visual exploration task.

4. APPLICATION

In our case study we apply our visual search and exploration designs to a real-world example. On the basis of a scientific research data set, we aspire two aspects: (1) prove the functionality of our approach, and (2) explore interesting characteristics in the search space. The challenge in dealing with scientific research data in the DL context is on the one hand (a) to satisfy the expert user expectations and on the other hand (b) to come up with illustrative use cases to attract interested but non-expert user groups. In compliance with this need, we designed our case study in close collaboration with experts from the Alfred Wegener Institute (AWI) for Polar and Marine Research in Bremerhaven, Germany.

4.1 Data Set and Application Domain

Our incorporated data set [4] is acquired from the open data repository PANGAEA, operated by the AWI. PANGAEA archives and publishes georeferenced scientific earth observation data in the research areas of water, ice, sediment and atmosphere. Our data set focuses on atmospheric weather measurements, gathered in the scope of the Baseline Surface Radiation Network (BSRN)[25], a PANGAEA compartment. In general, data of BSRN concerns the development of radiation and meteorological measurements over time, expressed by up to 100 physical parameters, recorded up to a temporal resolution of one measurement per minute. Common physical units include atmospheric pressure, relative air humidity, temperature and a variety of radiation-based measurements like shortwave downward radiation and longwave upward radiation.

4.1.1 Definition of Data Content

We decided to use temperature neasurements as our data content for two reasons. Temperature measurements are especially qualified to satisfy the public interest of non expert users (see Section 4.2, case study A) and the critical examination view of expert users (see Section 4.3, case study B). In consultation with researchers from AWI, we decided to choose a temperature measurement set of 22 BSRN stations recorded within the year 2006. Considering temperature measurements of a whole year is especially important to get a complete impression of all thermal behaviors to be expected within the most typical climatic time period - one year. Another important time period in considering climatic examinations concerns the duration of single days. Different regions on earth evoke entirely different measurement curves within the duration of one day. Scientists call this the phenomena of diverse *diurnal variations*. To comply with this requirement, we define temperature measurements of a single day, taken at one distinct measurement station on earth

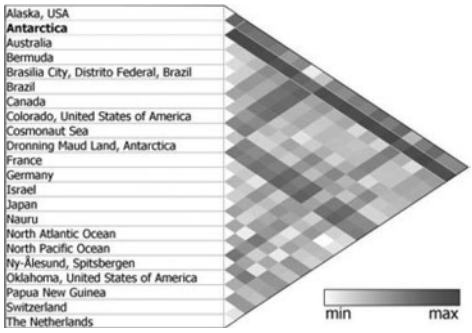

Figure 7: The DistanceMatrixView shows pairwise content-based distances of metadata property 'Location'. 'Antarctica' (selected) is dissimilar to all other metadata entities.

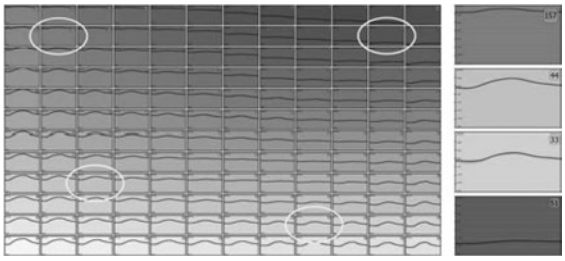

Figure 8: Left: Visual catalog of all daily temperature curves. Highest temperature values are discovered at the left (red, orange, brown), moderate values at the lower right (green, cyan), lowest temperatures at the upper right (blue). Right: Enlargements of four discriminative cells for closer examination.

as one data element (see Figure 2 for two exemplary daily curve progressions). In other words, the outcome of a daily temperature measurement is one 'pattern' that is applied as our content-based similarity modality and thus, the basis of our visual catalog that gives an overview of all (daily) temperature patterns measured in the data set (see Figures 8 and 9). Altogether we process 7677 daily patterns, since the 22 available BSRN stations produce 365 daily curve progressions each. About 400 patterns have been sorted out based on failed data consistency checks (e.g., missing values detection).

We want to point out that it only takes one mouse click to switch the entire data content to other available physical units. The same holds for the aggregation of measurement patterns, the similarity definition and the selection of metadata entities to be visually explored.

4.1.2 Selection of a Metadata Property

PANGAEA data files are divided in a clearly arranged metadata header and a data table with columns of measurements with distinct physical units. The interpretation of the metadata headers provides sufficiently enough metadata properties to comply the DataCite metadata standard we use in our platform. Besides mandatory DataCite attributes ('Identifier', 'Creator', 'Title', 'Publisher', 'Publication Year') and further optional DataCite metadata properties, we extract PANGAEA-specific metadata properties, like 'Location' or 'Coverage'. We decided to choose 'Location' as our focused metadata property, because the location of measurement stations on earth gives a good characterization of various climates to be expected. BSRN stations are distributed all over the world. Due to the heterogeneity of measured climates, we expect interesting exploration results within the defined search and exploration space.

4.2 Case Study A: Metadata Property 'Location' Explored by Absolute Temperature Measurements

Our first case study describes temperature measurements from 2006, taken from 22 BSRN stations all over the world. We choose the metadata property 'Location' as our metadata exploration space. Comparing *absolute* temperature values it can be expected that measurement stations geographically located close to each other score a high similarity. This scenario is qualified for expert and non-expert user groups in particular, since every user has a notion about temperature behavior of different locations on earth.

First, we will explore the result of our content-based similarity concept. The distance matrix (Figure 7) shows all metadata entities in alphabetical order. Remember that white diamonds denote high similarities, while blue diamonds show low similarities between two distinct metadata entities. We highlight 'Antarctica', which has considerably dissimilar temperature values to effectively all other locations in the data set, the highlighted row of 'Antarctica' comes along with the bluest color values in the entire exploration space. We record this statement and receive the validation of the researchers from AWI: Antarctica produces temperature measurements about 30°C colder than any other BSRN station. In contrast, the most similar metadata entities are 'Switzerland' and 'Germany', which is geographically comprehensible. Other similar entity pairs are the 'Cosmonaut Sea' and 'Alaska', or 'Papua New Guinea' and 'Australia'. Together with the researchers, we conclude that these pairs of locations are geographically grouped close together or rather on the same latitude. Due to the fact that 'Antarctica' is completely analyzed, we exclude this entity from the data set for more detailed exploration purposes.

Next, we calculate the visual catalog (see Figure 8) to get an overview over all daily measurement patterns of the data set. We choose our colormap in a way that red and orange color values denote high temperature measurements (left), moderate temperatures are color-coded in green and cyan (lower right) and coldest daily measurements are represented by blue cells (upper right). Furthermore, temperature progressions with higher peaks tend to be located at the lower half of the visual catalog (yellow, green, cyan). The overall temperature interval ranges from -29.2°C) to a maximum of 33°C.

With this obtained overview over the data content, we proceed our exploratory case study by examining the MetaMap (Figure 1). Together with the researchers, we discovered several groups of measurement stations with similar characteristics.

Result 1: At the upper left, the stations 'Papua New Guinea', 'North Pacific', 'Nauru' and 'Australia' are marked red. Most of these temperature measurements are located in

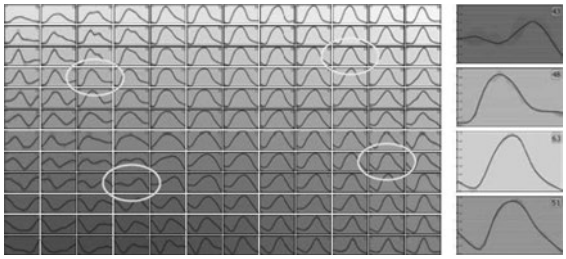

Figure 9: Visual catalog of all relative daily temperature curves. Progressions with an early maximum displayed at upper left (cyan), curves with daily maximum in the afternoon on the right (yellow, orange, red), diverse curve progressions with missing temperature peaks in the mid day at the lower left (blue).

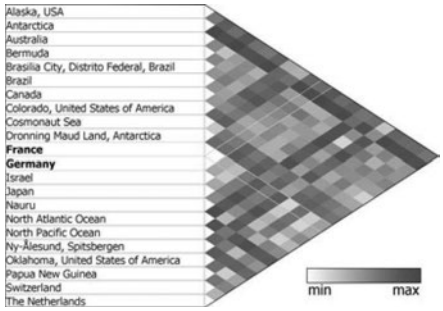

Figure 10: The DistanceMatrixView displays 'France' and 'Germany' as the most similar metadata entities in the data set (white diamond).

the upper left of the visual catalog and also coded with red color values. This group of metadata entities comprises the hottest temperatures in the entire data set. We infer that these stations are located in tropical or desert-like climates.

Result 2: Another group of stations is located at the center of our layout, marked with yellow, green and cyan color. Most of the stations are located in moderate climates on earth like Europe ('Netherlands', 'Germany', 'Switzerland', 'France').

Result 3: Finally we want to explore the group at the upper right of our layout, denoted with blue and purple colors. This group comprises the coldest measurements in the data set, all content summaries of the entity glyphs contain measurements taken from the blue and purple region of the visual catalog (see Figure 8), denoting low temperatures (mostly below 0°C). On inspection, the locations 'NY-Ålesund', 'Alaska', 'Cosmonaut Sea' and 'Dronning Maud Land' provide the lowest temperature values in the data set, which is evidence for the validity of both our similarity concept and our topology-preserving MetaMap.

4.3 Case Study B: Metadata Property 'Location' Explored by Relative Temperature Measurements

With our second, more scientific case study, we follow a request of researchers at the AWI. We explore the metadata property 'Location', based on *relative* temperature curve progressions. Consider that the analysis of patterns can either be performed on absolute (temperature) values (like in case study A) or on relative curve progressions. The progression of each temperature curve within one day disregarding the absolute temperature values is of utmost importance for our present notion of similarity. To illustrate the benefit of relative curve analysis, the reader is referred to stock market analysis tasks, where stocks of similar branches often develop in a similar manner, even if their absolute values are entirely different. One reason might be that all concerning stocks are dependent on similar impacts.

Back to our case study, we firstly consider the visual catalog in Figure 9. We identify daily temperature progressions with a maximum before noon on the upper left (cyan) and curves with their peaks in the afternoon on the right (yellow, orange, red). In the upper left, the visual catalog shares curves with indefinite behavior (blue, purple). Many temperature progressions have maxima in the evening or even at night. Even though researchers from AWI describe this behavior as ordinary in arctic regions, we were surprised about this outcome of the visual catalog. Finally, we discover curves with clear daily maxima in the middle of the day to be located at the right of the visual catalog.

Next we investigate the DistanceMatrixView in Figure 10. We select the whitest diamond with the minimum pairwise distance and consider 'France' and 'Germany' as the most similar metadata entity pair in the data set. Furthermore we point out 'France' and 'Switzerland', 'North Pacific Ocean' and 'Bermuda', 'Switzerland' and 'Germany', as well as 'Switzerland' and 'Japan' to have above average degrees of similarity. A great many of the most dissimilar locations in the data set are addressed by the entities 'Antarctica' and 'North Atlantic Ocean'. We regard these findings as important and to deserve further exploration.

Together with the researchers from AWI, we explore our MetaMap of all locations in the data set based on the similarity measure of relative daily temperature curves. In the following, we detail the results of our explorations:

Result 1: The biggest group of locations is arranged at the lower right regions of the MetaMap (orange, brown). Measurements taken from the corresponding stations have sufficiently distinctive temperature peaks in the middle of the day, typically after noon. This attitude is obtained by moderate and continental climates. Locations from Central Europe and measurements taken from 'Oklahoma' and 'Canada' prove this hypothesis.

Result 2: A second cluster is found at the upper of the MetaMap (cyan). It is particularly noticeable that measurements within the distinct content summaries arise from the green and cyan colored regions of our visual catalog, where curves with a daily maximum before noon are discovered. The measurement stations are predominantly located at tropical, maritime places on earth - known to be influenced by gathering clouds at noon and rainstorms in the afternoon, which explains aggregated daily temperature maxima even before noon. The location 'Colorado' is depicted as an outlier. We inquired the researchers and found out that the measurement station is located on more than 1500 meters above sea level in a rainy climate region, which explains the cluster affiliation.

Result 3: A third group of metadata entities is recognized on the lower left (blue), all measurement stations are located at arctic regions. The crossover to the orange clus-

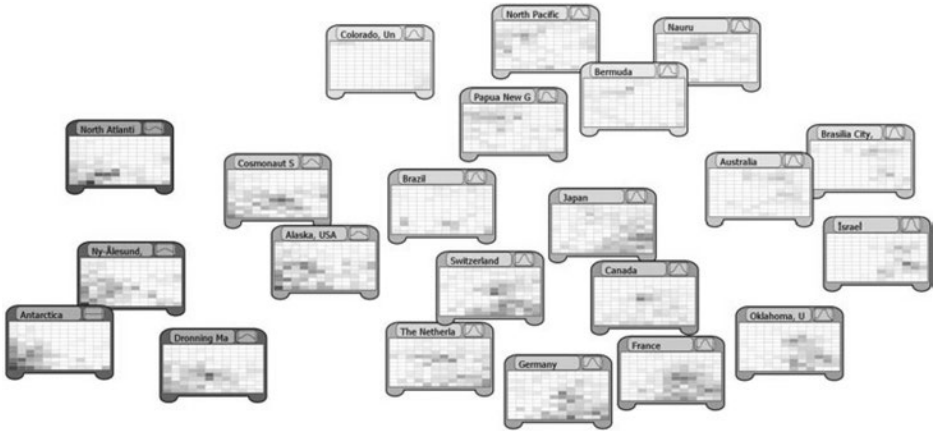

Figure 11: MetaMap of all 'Locations' in the data set. Predominantly maritime and tropical regions are arranged at the top (cyan), locations of moderate and continental climates are located at the bottom (orange, brown). At the lower left, a group of arctic locations is shown (blue).

ter is built by 'Cosmonaut Sea' and 'Alaska', which is also geographically comprehensible. Arctic climates are characterized by temperature behaviors apart from solar zenith angles, typically without a major temperature peak at noon. This behavior can be stated by the blue measurements in the content summaries denoting rather heterogeneous temperature progressions in the visual catalog. Note that the station 'Antarctica' even has no average temperature deviation within the whole day, which makes the station significantly dissimilar to all other locations in the data set.

5. CONCLUSIONS

We presented an approach for exploratory search in time-oriented research data. The approach is based on user-selectable metadata properties for which visual layouts are presented that reflect similarity of the underlying metadata entities. Our approach is novel in that it derives a notion of similarity from the content of the data, and allows exploration of the data based on metadata properties. It allows users to examine relevant relationships between entities and is a basis for subsequent drill-down queries. The approach is particularly useful for navigating in large data spaces for which users have no a-priori knowledge.

This work is only a first approach in our idea to provide users with a more encompassing access to research data both from the metadata and content perspectives. Future work includes extending the approach to further domain-specific similarity notions in time-series data, considering, e.g., similarity based on time-series motif analysis or correlation between measurements. Also, we currently neglect similarity between metadata properties themselves. For example, similarity between geospatial distance of measurement locations or similarity between authors according to their scientific institution could be considered. To this end, comparative exploration of groups of data according to content and metadata should be considered. The 2D layout could be enhanced by additional visual representations of data properties, such as the scientific methodology applied for obtaining measurements. Finally, validation of the approach by expanded domain user studies should be performed in the future.

Acknowledgments

We thank the Alfred Wegener Insitute (AWI) in Bremerhaven, particularly Rainer Sieger, Hannes Grobe and Gert König-Langlo, and everyone involved with PANGAEA for supporting this research effort. We are especially grateful to the many scientists that contributed the data available through BSRN and other research projects.

6. REFERENCES

[1] M. Agosti, S. Berretti, G. Brettlecker, A. D. Bimbo, N. Ferro, N. Fuhr, D. A. Keim, C.-P. Klas, T. Lidy, D. Milano, M. C. Norrie, P. Ranaldi, A. Rauber, H.-J. Schek, T. Schreck, H. Schuldt, B. Signer, and M. Springmann. Delosdlms - the integrated delos digital library management system. In *DELOS Conference*, pages 36–45, 2007.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Semantic Web Conf.*, pages 11–15. Springer, 2007.

[3] J. Bernard, J. Brase, D. Fellner, O. Koepler, J. Kohlhammer, T. Ruppert, T. Schreck, and I. Sens. A visual digital library approach for time-oriented scientific primary data. *Springer International Journal of Digital Libraries, ECDL 2010 Special Issue*, 2011.

[4] J. Bernard, T. Ruppert, M. Scherer, J. Kohlhammer, and T. Schreck. Reference list of 269 sources used for exploratory search. doi:10.1594/pangaea.778638, 2012.

[5] J. Bollen and H. Van de Sompel. An architecture for the aggregation and analysis of scholarly usage data. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '06, pages 298–307, New York, NY, USA, 2006. ACM.

[6] C. L. Borgman, J. C. Wallis, and N. Enyedy. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, 7(1-2):17–30, 2007.

[7] S. Bremm, T. von Landesberger, J. Bernard, and T. Schreck. Assisted descriptor selection based on

visual comparative data analysis. *Comput. Graph. Forum*, 30(3):891–900, 2011.

[8] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2581–2590, 2011.

[9] S. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307, 2007.

[10] K.-H. Cheung, E. Lim, M. Samwald, H. Chen, L. N. Marenco, M. Holford, T. M. Morse, P. Mutalik, G. M. Shepherd, and P. L. Miller. Approaches to neuroscience data integration. *Briefings in Bioinformatics*, 10(4):345–353, 2009.

[11] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107, 2008.

[12] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2001.

[13] M. Duke, M. Day, R. Heery, L. A. Carr, and S. J. Coles. Enhancing access to research data: the challenge of crystallography. In *JCDL*, pages 46–55. ACM, 2005.

[14] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. A descriptor for large scale image retrieval based on sketched feature lines. In *Eurographics Symposium on Sketch-Based Interfaces and Modeling*, pages 29–36. ACM, 2009.

[15] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[16] T. Hey and A. Trefethen. Cyberinfrastructure for e-Science. *Science*, 308(5723):817–821, 2005.

[17] D. Hull, S. R. Pettifer, and D. B. Kell. Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web, 2008.

[18] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering the information age: solving problems with visual analytics*. Eurographics, 2011.

[19] D. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Information Visualization*, pages 9 –16, 2006.

[20] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Inf. Systems*, 3(3):263–286, 2001.

[21] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[22] C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: an architecture for complex objects and their relationships. *Int. J. Digit. Libr.*, 6(2):124–138, 2006.

[23] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *Transactions on Multimedia Computing, Communications, and Applications*, 2(1):1–19, 2006.

[24] A. Noack. An energy model for visual graph clustering. In *International Symposium on Graph Drawing*, pages 425–436. Springer-Verlag, 2003.

[25] A. Ohmura, E. G. Dutton, B. Forgan, C. Fröhlich, H. Gilgen, H. Hegner, A. Heimo, G. König-Langlo, B. mcarthur, G. Müller, R. Philipona, R. Pinker, C. H. Whitlock, K. Dehne, and M. Wild. Baseline surface radiation network (BSRN/WCRP): New precision radiometry for climate research. *Bull. Amer. Met. Soc.*, 79:2115–2136, 1998.

[26] PANGAEA Data Publisher for Earth & Environmental Science. `http://www.pangaea.de/`.

[27] A. Powell, M. Nilsson, A. Naeve, and P. Johnston. Dublin core metadata initiative - abstract model, 2005. White Paper.

[28] A. Rauber and M. Frühwirth. Automatically analyzing and organizing music archives. In *ECDL*, pages 402–414. Springer, 2001.

[29] S. Rueger. *Multimedia Information Retrieval*. Morgan and Claypool Publishers, 2010.

[30] M. Scherer, J. Bernard, and T. Schreck. Retrieval and exploratory search in multivariate research data repositories using regressional features. In *JCDL*, pages 363–372. ACM, 2011.

[31] Sloan Digital Sky Survey. `http://www.sdss.org/`.

[32] L. D. Stein. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews Genetics*, 9(9):678–688, 2008.

[33] The DataCite consortium. DataCite: Helping you to find, access, and reuse data. `http://datacite.org/`.

[34] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl. Enhancing digital libraries with techlens+. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '04, pages 228–236, New York, NY, USA, 2004. ACM.

[35] G. Tsatsaronis, I. Varlamis, S. Torge, M. Reimann, K. Nørvåg, M. Schroeder, and M. Zschunke. How to become a group leader? or modeling author types based on graph mining. *TPDL*, pages 15–26, 2011.

[36] J. Vesanto. SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126, 1999.

[37] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. vanWijk, J.-D. Fekete, and D. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. 2011.

[38] White and Roth. *Exploratory Search - Beyond the query-response paradigm*. Morgan and Claypool, 2009.

[39] A. J. Williams. A perspective of publicly accessible/open-access chemistry databases. *Drug discovery today*, 13(11-12):495–501, 2008.

[40] I. H. Witten, R. J. Mcnab, S. J. Boddie, and D. Bainbridge. Greenstone: A comprehensive open-source digital library software system. In *International Conference on Digital Libraries*. ACM, 2000.

[41] Wolfram Alpha. Wolfram|Alpha: Computational Knowledge Engine. `http://www.wolframalpha.com/`.

[42] B. Wong, S. Choudhury, C. Rooney, R. Chen, and K. Xu. Invisque: technology and methodologies for interactive information visualization and analytics in large library collections. *TPDL*, pages 227–235, 2011.