Retrieval and Exploratory Search in Multivariate Research Data Repositories using Regressional Features

Maximilian Scherer Jürgen Bernard Tobias Schreck

Interactive Graphics Systems Group
Technische Universität Darmstadt
Fraunhoferstr. 5, 64283 Darmstadt, Germany
{maximilian.scherer, juergen.bernard, tobias.schreck}@gris.tu-darmstadt.de

ABSTRACT

Increasing amounts of data are collected in many areas of research and application. The degree to which this data can be accessed, retrieved, and analyzed is decisive to obtain progress in fields such as scientific research or industrial production. We present a novel method supporting content-based retrieval and exploratory search in repositories of multivariate research data. In particular, functional dependencies are a key characteristic of data that researchers are often interested in. Our methods are able to describe the functional form of such dependencies, e.g., the relationship between inflation and unemployment in economics. Our basic idea is to use feature vectors based on the goodness-of-fit of a set of regression models, to describe the data mathematically. We denote this approach *Regressional Features* and use it for content-based search and, since our approach motivates an intuitive definition of interestingness, for exploring the most interesting data. We apply our method on considerable real-world research datasets, showing the usefulness of our approach for user-centered access to research data in a Digital Library system.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Algorithms

Keywords

Research Data Repositories, Content-Based Access, Parametric Fitting, Interestingness Analysis, Clustering

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.

1. INTRODUCTION

In many scientific disciplines relying on empirical data, e.g., earth observation, experimental physics, medical and biological science, economics and the social sciences, vast amounts of research data are produced or gathered on a daily basis. Often being funded by the public, demand for *open access* to the produced data is arising. Making research data publicly available has several benefits. First, *reproducibility* and transparency of obtained results is a principal requirement for good scientific practice and publishing. Second, finding data *related* to one's own work is crucial for many researchers.

As a motivating example, consider that several top-tier scientific conferences and journals increase acceptance as an incentive for researches to actually publish their data. *Econometrica* [8] is such an example, where authors are actually required to submit their data. Often though, research data is provided on an individual basis, with researchers putting up undocumented data, in an arbitrary format on personal web-space. Such data is usually available only for a limited time. Therefore, such practice hardly supports the demand for reproducibility, let alone the possibility to find related data. Hence, a need for *library-oriented* handling of research data exists. This aims at the centralized, long-term availability of data, adhering to specific formatting and documentation requirements. As such, this treatment of research data allows for reproducibility by supplying data associated with scientific publications as well as finding related data by searching for related textual publications. Well-established database techniques and thorough data curation, to guarantee format-adherence and meaningful metadata annotations, allow digital libraries to provide research data in such a way.

Research data typically contains large quantities of non-textual, digital data content for which no native system support beyond *annotation-based* access is provided. In the multimedia digital library context, to date several systems exists that support content-based search relying on automatically extracted descriptors. However, devising meaningful retrieval methods for research data is a difficult problem. Current systems rely on textual (metadata) based search, and content-based search *in* the data is typically not supported. Content-based access in research data is therefore mostly unexplored in the digital library context.

The contribution of this paper is to support content-based retrieval and explorative search in research data, by proposing a novel data similarity notion that is particularly suited

in a user-centered Digital Library context. This similarity notion is based on *functional dependencies* between observation variables in the data and thereby captures a most important and generic data aspect. Classical examples of functional dependencies include the *Phillip's Curve* (relationship between inflation and unemployment), accounted for in empirical econometrics data, or *Ohm's Law* (relationship between current, voltage and resistance) discovered by measurements in an electrical circuit. The basic idea to capture such dependencies, is to describe the data mathematically by forming a descriptor (feature vector) capturing goodness-of-fit parameters of the data to several regression models. Hence we call this approach *Regressional Features* (see Section 3 for details).

We show the utility of our approach on a considerable set of real-world research data. Furthermore, we discuss a unifying framework to extend support for research data in digital libraries by content-based means (using regressional features) in conjunction with established techniques relying on metadata annotations.

2. RELATED WORK

We will begin the section on related work, where we left off in the introduction, and elaborate on related techniques and algorithms for data mining and knowledge discovery in multivariate data. The second subsection details on the related Digital Library context.

2.1 Content-Based Access to Research Data

Content-based access to research data requires data mining techniques for data import, preprocessing and comparison. Research data derived from possibly different data repositories usually is highly heterogeneous. In general, a generic data standard must be derived by which the research data can be imported in the retrieval system.

To allow content-based access, the primary task is the definition of a concept for data comparison. Similarity calculation approaches are highly data and application dependent. For example, Liao [18] surveys a set of similarity approaches for time series data. We consider the description of two dimensional-data by its goodness-of-fit to several functional models, encoded as a feature vector. The feature vector approach is prominently used in multimedia retrieval, for example to capture visual properties of images and shapes for retrieval [16].

Related approaches to describe the functional form of data include methods and references in the book by Ramsay et al. [22] and recent work by Hebrail et al. [13]. These methods though, are usually *non-parametric* (using basis functions, i.e. splines, Fourier series, wavelets, etc.), in contrast to our approach based on parametric models. The major advantage of non-parametric methods is, that the number of parameters (or *coefficients* in that case) can be adaptive to the complexity of the data. As a drawback, they lack the explanatory power of parametric models and are harder to be interpreted by users [12].

Another related topic, that does not use parametric models, but rather strives to derive them directly from the data, is *scientific discovery* [23]. A particular connection to our work is found in [26] by Todorovski and Dzeroski, where they describe the inclusion of *domain knowledge* into the discovery process. We also support this with our approach, by allowing users to add their functional models depending on application needs. For more information on the importance of including *a priori knowledge* and further topics in data-mining please refer to the work of Fayyad et al. [12].

Due to the vast amounts of available research data, there is a need to create a visual overview using visualization and clustering techniques [2]. Assigning instances of data to clusters based on their similarity yields such an overview. Prominent algorithms include `kmeans` [19] and *self-organizing maps* (`SOM`) [14], which are recently employed in [17] and [25], respectively.

2.2 Digital Library Context

Digital Library systems have evolved from mere research prototypes into production stable pieces of software, allowing us to cope with the rapidly increasing numbers of digital documents. Prominent DL systems include [6, 15, 27].

So far, these DL systems focus on *annotation-based* access to documents, as well as rendering *textual*-content accessible (e.g., by full-text search). This is well-suited for textual documents, however support for *non-textual* documents usually relies on some metadata standard (e.g., MPEG7 for multimedia) and is often lacking appropriate *content-based* access (e.g., comparing similarity of images based on color distribution). Multimedia documents (e.g., audio, image, video, 3D models) and recently, research data gathered in natural and empirical sciences, have been recognized as important non-textual documents with a need for library-oriented treatment.

Related, prototypical and commercial systems to support content-based access of non-textual documents include PROBADO [4] (classical music and 3D architectural models), VICTORY [7] (3D models) and Google's *Similar Images* and *3D Warehouse* approaches. Recently, an approach for content-based access to time-series research data in DL systems was proposed [3].

Repositories and data libraries collecting research data from different domains include generic data underlying natural sciences publications [9], geoscientific and environmental data [20], psychological data [21], or biological information [11] and highly motivate research to increase data-accessibility.

3. REGRESSIONAL FEATURES

What kind of function does the scatterplot on the upper part in Figure 1 visualize? Just by looking at the form of this plot, we as humans are certainly not able to formulate a precise mathematical function underlying such a plot, but still assessing the crude functional form (like x^2 or $\frac{1}{x}$) is possible.

One intuitive way to describe the presence of a functional relationship within data mathematically is *correlation*. Computing *Pearson's correlation coefficient* for two-dimensional data returns a value in $[-1, 1]$, where the absolute value of that coefficient relates to how well a line fits to the data and the sign indicates the direction.

Inspired by this, we propose *regressional features* to describe the functional form of two-dimensional data projections in a similar, but extended way. The algorithm to achieve that is outlined in Figure 1. Our basic idea is to generalize the notion of correlation, by fitting the data to a number of representative functional models using regression. Computing the relative goodness-of-fit (GOF) to each of these models and storing these parameters in a vector,

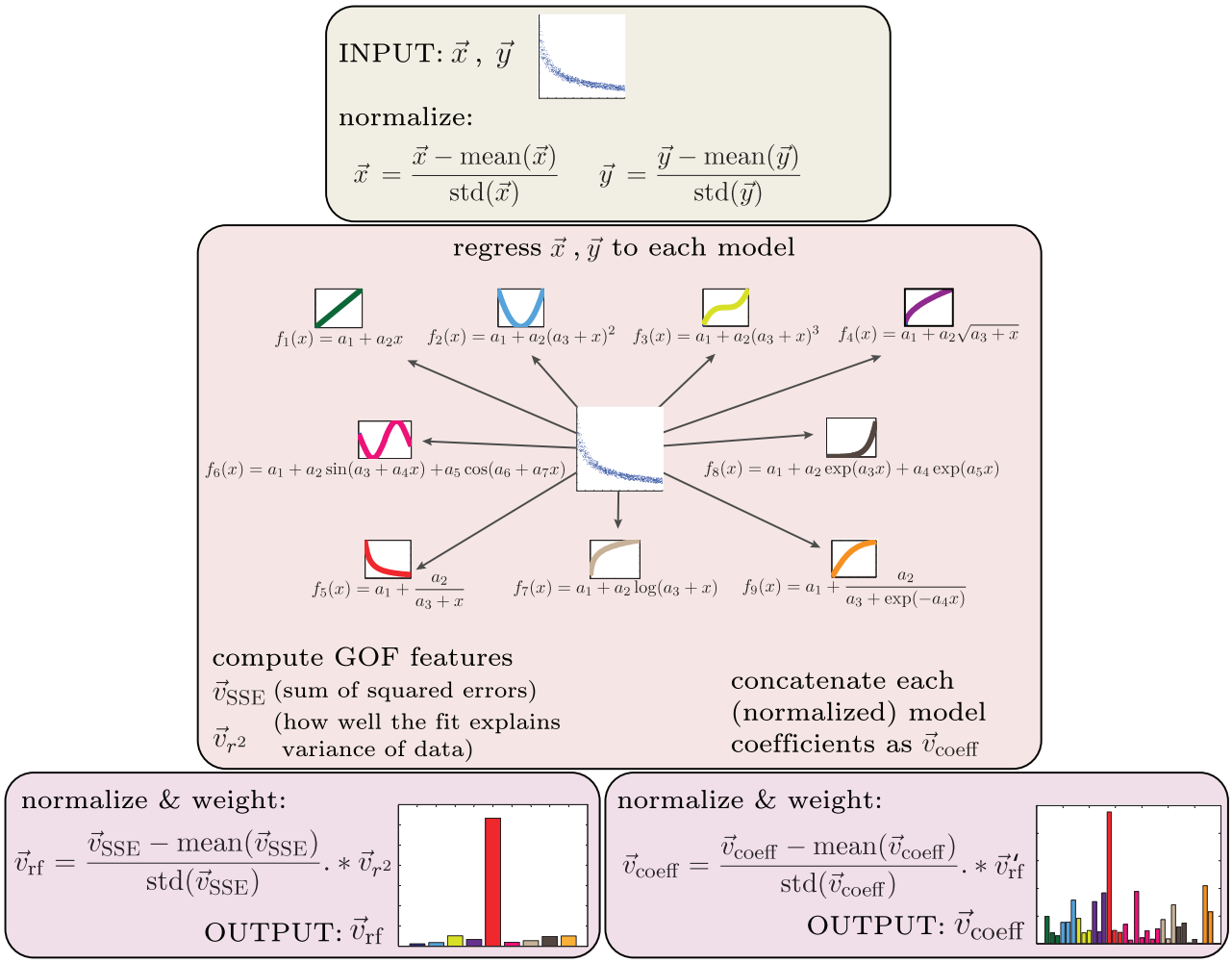

Figure 1: Schematic algorithm outline to compute regressional features. Top: Data input and normalization; Middle: Goodness-of-fit to specified functional models; Bottom: Normalization and weighting of descriptors

yields a powerful descriptor of the functional form of the data. We therefore denote this descriptor as *regressional feature vector* $\vec{v}_{\text{rf}}$. The functional models we use are included in the figure and their respective functional form is visualized as a colored plot. We chose these models with complementarity and completeness in mind, such that at least one of the models should be suitable to describe any kind of functional relationship in the data, while not capturing any functional properties the other models would be able to. Experimental results (see the next section) show that the chosen models work well in general. Our apporach may also easily be extended by further functional models as possibly required by specific application domains. We provide a simple text-based interface for users, to alter, remove or add new functional models to the regressional features extraction algorithm to adjust it to their needs.

In the lower left of Figure 1 we see $\vec{v}_{\text{rf}}$ computed for some exemplary data. It is visualized as a colored histogram, since each entry of the vector relates to the probability of the correspondingly colored functional model being applicable.

The *regressional feature vector* can be used to assess interestingness of a scatterplot (useful for ranking, filtering and highlighting), as well as retrieval and clustering of scatterplots by functional similarity of the two-dimensional data they visualize. For example, we can retrieve all scatterplots visualizing data similar to $f(x) = x^2$ from a database, or give an exploratory overview of several different kinds of scatterplots in a database by clustering according to their functional similarity.

Compared to nonparametric data analysis techniques (see Functional Data Mining [22]), regressional features offers advantageous properties for user-centered applications like Digital Library systems. Since every coefficient relates to a specific functional model, each of them is interpretable by the user. Furthermore interestingness and similarity functions can be intuitively defined (as detailed in the following subsections) and suited to specific user needs. This leads to the transparency of results obtained, which is expected to increases acceptance thereof by domain experts, as they see *how* and *why* a particular result was computed.

To describe multivariate data with this algorithm, we compute regressional features for each pairwise combination of all the variables in the data. Although we are well aware that certain patterns might be hidden in higher dimensions, we opt not to apply any dimension reduction techniques (like PCA), which would result in a combination of the orig-

inal measurement variables. By only projecting one variable orthogonally versus another, we allow domain experts to discover functional relationships between two specific variables as well as retrieving similar relationships among those two variables in other datasets, as they are uniquely identified by an annotated label.

A secondary descriptor called *regressional coefficient feature vector* $\vec{v}_{\text{coeff}}$ is also computed as part of our regressional features approach. It is composed of the actual (albeit normalized and weighted) coefficients obtained by each regression. It is shown on the lower right in Figure 1 and is also visualized as a colored histogram. Note that sets of entries have the same color to show that these entries belong to coefficients from the correspondingly colored functional models (i.e. the three entries colored red correspond to model f_5). This color also indicates which sets of coefficients were weighted with the corresponding entry of $\vec{v}_{\text{rf}}$, to avoid having coefficients of entirely inapplicable models in $\vec{v}_{\text{coeff}}$. This descriptor allows assessment of similarity on a finer level of detail, since $\vec{v}_{\text{rf}}$ captures the *functional family* of the data, while $\vec{v}_{\text{coeff}}$ captures an *concrete instantiation* (the coefficients) of each functional model. As such, $\vec{v}_{\text{rf}}$ is *robust* against changes in functional properties like *offset, slope, amplitude* and *frequency*, and is recommended for a first search on the level of models. If a specific model has been found, $\vec{v}_{\text{coeff}}$ can be used for a refinement step. It is also possible to specify a combination relying on both search notions (cf. Section 3.2). The specifics of combining both descriptors will rely on the data domain and/or the interactive retrieval by the user, who is able to select the descriptor combination to use at any step. Actual examples of regressional features, computed for generated data, are shown in Figures 2 and Figure 3.

3.1 Interestingness

The notion of *interestingness* certainly depends on the domain and each user's individual preferences. We consider two-dimensional data projections interesting, if they can be (unambiguously) well described by a functional relationship. Due to the construction of regressional features, we are able to compute precisely this notion of interestingness using the following function:

$$I_\alpha(\vec{v}_{\text{rf}}) = \alpha \cdot \texttt{sum}(\vec{v}_{\text{rf}}) + (1-\alpha) \cdot \sqrt{\texttt{var}(\vec{v}_{\text{rf}}) \cdot \texttt{length}(\vec{v}_{\text{rf}})} \quad (1)$$

Recalling Figure 1, we know that each entry of $\vec{v}_{\text{rf}}$ relates to the probability of one of the functional models being applicable to describe the underlying data. Each entry was weighted with the R-squared measure *after* normalization, thus we compute the *overall applicability* of $\vec{v}_{\text{rf}}$ by summing up all entries. This is the first part in Equation 1.

The second part computes the *unambiguity* of the functional models by the *standard deviation* (normalized to $[0,1]$) of $\vec{v}_{\text{rf}}$. The user adjustable weight α assigns more or less importance to *overall applicability* or *unambiguity*, depending on user preference.

An example for interestingness of regressional features, in particular with regard to overall applicability and unambiguity, is provided in Figure 2.

3.2 Similarity

By means of regressional features we obtained a descriptor for the functional form of two-dimensional data projections.

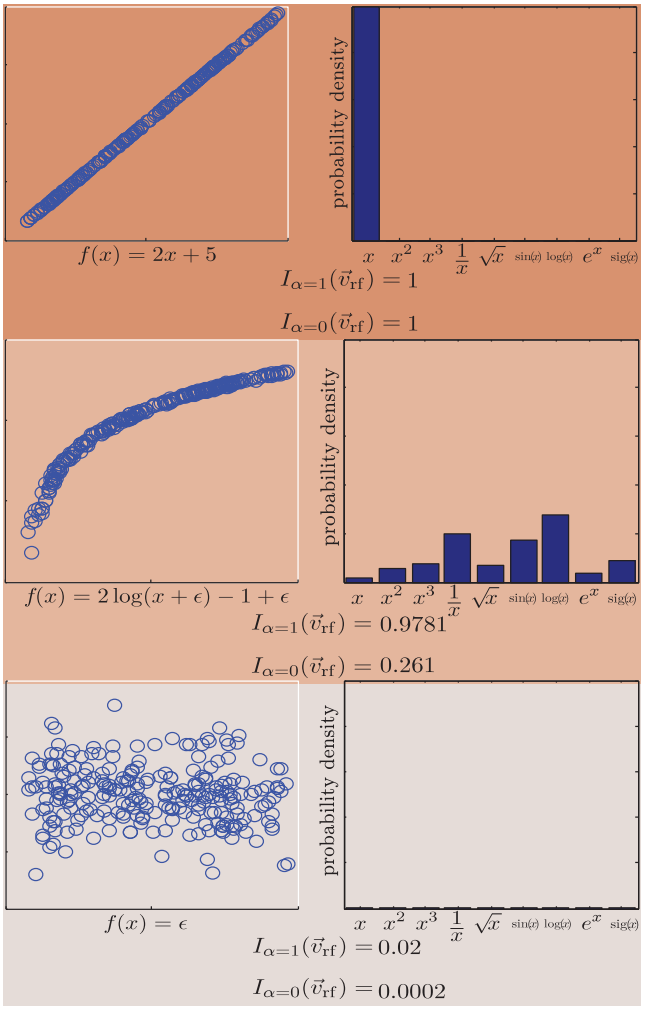

Figure 2: Interestingness of regressional features for $\alpha = 1$ (overall applicability) and $\alpha = 0$ (unambiguity). Top: applicable and unambiguous; Middle: applicable but ambiguous; Bottom: not applicable and ambiguous

Computing the distance between the descriptors of two data projections yields a measure for the (dis-)similarity of their respective functional form. Such a distance function is a requirement for any retrieval or clustering algorithm.

Since our descriptor consists of two vectors, $\vec{v}_{\text{rf}}$ and $\vec{v}_{\text{coeff}}$, we propose to use a weighted L_1 distance of both vectors to measure the overall dissimilarity.

$$d_{\text{R}}(A,B) = \beta \cdot |\vec{v}^A_{\text{rf}} - \vec{v}^B_{\text{rf}}| + (1-\beta) \cdot |\vec{v}^A_{\text{coeff}} - \vec{v}^B_{\text{coeff}}| \quad (2)$$

The superscripts A and B indicate the two sets of data being described. Increasing the user-adjustable weight β gives more importance to *overall* similarity of functional form (as described by $\vec{v}_{\text{rf}}$). If the user decreases β, more importance is given to similarity of the functional *coefficients* (as described by $\vec{v}_{\text{coeff}}$).

Figure 3 provides an example measurement for this distance with $\beta = 1$ and $\beta = 0$ to show the significant influence of the weight parameter.

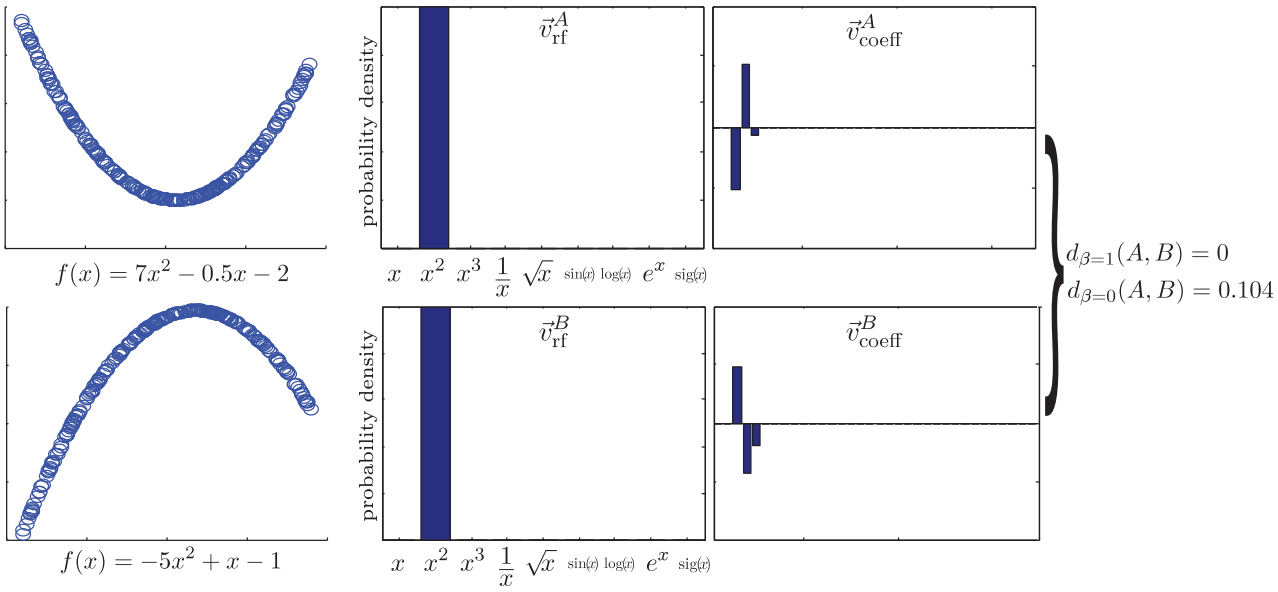

Figure 3: Dissimilarity of two datasets A and B by computing the distance between their regressional features focusing on *functional form* ($\beta = 1$), or on *functional coefficients* ($\beta = 0$). Our similarity measure allows users to search between those two notions, by selecting β.

4. APPLICATION

In Section 3 we presented the idea and algorithmic outline of regressional features, that allows us to describe the functional form of two-dimensional data projections in a very convenient way. However, to demonstrate applicability of this algorithm in Digital Library systems, we will evaluate our approach on two real-world research datasets. These datasets are described in the following subsection. Two use-cases for content-based access to this data (retrieval and exploratory search) are described and evaluated afterwards. The results obtained therein motivate the proposal in subsequent Section 5, to extend Digital Library systems with support for these content-based search modalities in multivariate research data.

4.1 Datasets

We created two datasets consisting of primary research datafiles available through the scientific data library PANGAEA [20] operated by the Alfred-Wegener-Institute for Polar and Marine Research in Bremerhaven, and the Center for Marine Environmental Sciences in Bremen. PANGAEA archives, publishes, and distributes geo-referenced scientific earth observation data, collected by scientists in many different research efforts.

The data comprises observations and measurements of four main areas of study of several research projects and includes water, sediment, ice and atmosphere. PANGAEA supports data export for post-processing and analysis purposes. This covers metadata and tabular raw data, as provided by individual research projects, e.g., BSRN (see below). Metadata includes citations, spatial and temporal conditions, parameter description and – most importantly for our approach – variable names and physicals units.

Most of the research data itself is sequential and multivariate. That is, several dependent variables (i.e. *pressure* or *ozone*) are measured for one or more independent variables (most prominently *time* and i.e. *altitude*). This primary research data is curated and subsequently annotated with metadata and cite-able via a persistent identifier (DOI) according to the *DataCite* [5] standard.

The first dataset, `dataMixed` [24], consists of 1110 datasets spanning all research domains published through PANGAEA. These datasets were obtained manually through PANGAEA's search functionality[1], by selecting several collections of datasets from different research domains. Consequently, `dataMixed` consists of quite heterogeneous research data. Since we are interested in analyzing *multivariate* research data, we are looking at each pairwise combination of variables within each research-dataset. Please note, that we thereby ignore sequentiality, as we are interested in global dependencies between variables. In total, this results in 84,461 two-dimensional variable combinations (scatterplots) and approx. $47 \cdot 10^6$ data-points. Based on this data, we compute a descriptor for each scatterplot using regressional features to support content-based access.

The second dataset, `dataBSRN` [24], is composed of 6770 datasets as provided by the *Baseline Surface Radiation Network* [1] (BSRN) through PANGAEA. The data tables have up to 100 columns (variables / measurements), and up to 50,000 rows (number of observations). Data provided by BSRN is dominated by measurements of radiation (short-wave, long-wave, diffuse, direct), temperature, humidity and wind (speed, direction). Again we computed regressional features for each pairwise combination of variables in each dataset. Here we provide content-based access to a total of 295,475 scatterplots, consisting in total of more than $3 \cdot 10^9$ data-points.

4.2 Retrieval

A very prominent approach in content-based retrieval, especially in multimedia retrieval, is query-by-example. Here,

[1] `http://www.pangaea.de`

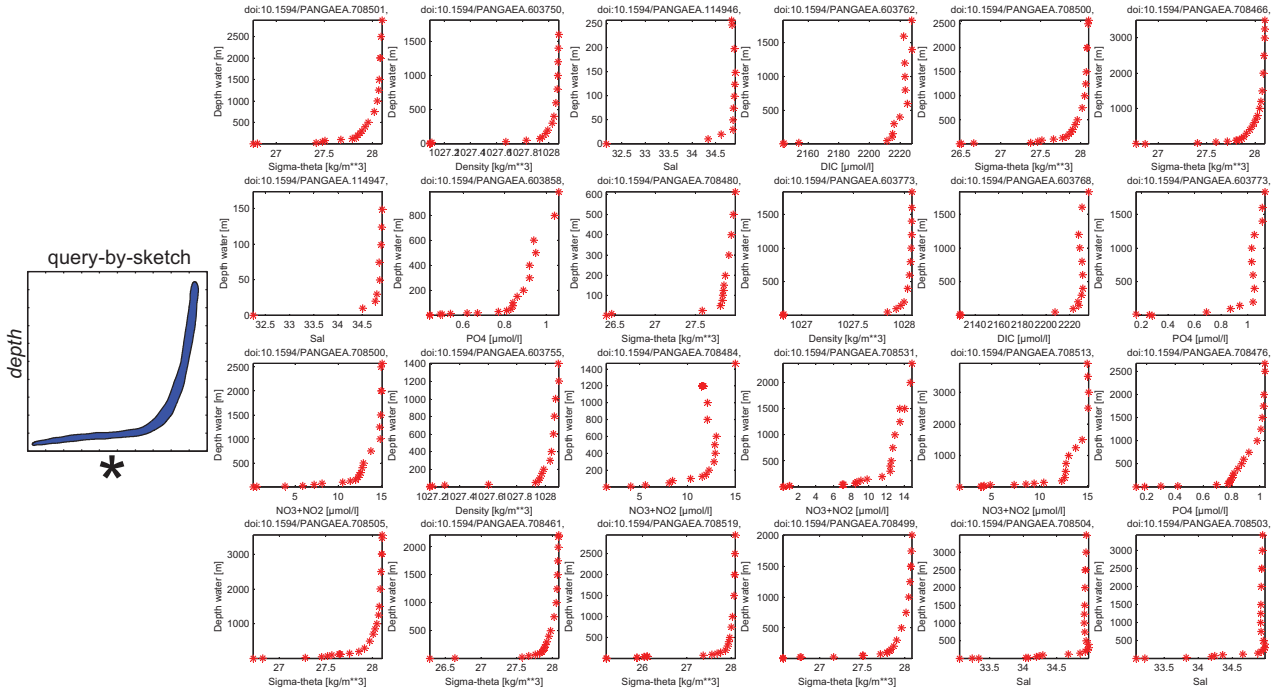

Figure 4: Retrieval Results using query-by-sketch for $*$ `vs.` *depth* in `dataMixed`. Looking at the different x-axis labels, we see that *salinity*, *density* (σ_θ), *phosphate* (PO4), *dissolved, inorganic carbon* (DIC) and *nitrate and nitrite* (NO3+NO2) all exhibit a functional dependency on *water depth* similar to our sketch.

a user supplies an example object (i.e. an example image) and is returned a ranking of all the objects in a database according to their similarity.

Thus, content-based retrieval of research data allows scientists to retrieve data that is similar to exemplary data. A particular use-case here is finding data that either disproves or verifies a certain hypothesis, by querying with one's own data underlying the given hypothesis.

Using regressional features, the datasets are indexed by functional relationships between every two variables. Therefore scientists can formulate a query by selecting two variables they are interested in (i.e. *Altitude [m]* vs. *PPPP [hPa]*) and specifying a functional relationship (i.e. $\frac{1}{x}$) these two variables need to be similar to, according to the query.

So in principal we also follow the query-by-example paradigm here. But a user does not necessarily have to provide explicit example data. Other query modalities (that implicitly generate example data) to query for a functional relationship are available. A user can either enter a functional relationship directly as a mathematical formula, by sketching a scatterplot or by retrieving the *most interesting* scatterplots. These query modalities are used in the query examples depicted in Figure 4 and Figure 5.

Given such a query, we filter all available research-datasets via textual search for those containing the variables in question. After this filtering, we are left with all scatterplots depicting the two query variables, which we then rank according to their functional similarity to the query data. This is accomplished by computing the distance between the respective regressional features (recall Eq. 2).

Although we do not go into detail on algorithmic properties like space-requirement and run-time, we briefly discuss our empirical observations here. For `dataBSRN` we require about six GB of main memory to store the raw data, metadata and the precomputed regressional features for each variable combination (recall, approx. 300,000). We reach average query times below 2 seconds on a modern desktop PC (Core i7 2,6Ghz, 12GB RAM) using our prototypical Matlab implementation. Space-requirement (for regressional feature precomputation) and query run-time depend linearly on the number of variable combinations, so they depend quadratically on the number of dimensions in each dataset.

4.3 Exploratory Search

In contrast to retrieval scenarios, where the user usually has a specific pattern to search for in mind, exploratory search is concerned with guiding the user to *interesting* patterns. By applying clustering algorithms, we can assign data to different groups and give an overview of the data.

In particular, regressional features allows us to cluster all scatterplots in our research database according to their functional similarity. We apply a `kmeans` clustering, which tries to optimize a specified number of *centroids* (in a least-squared-distances sense) and assigns each regressional feature vector to the nearest centroid.

Using this clustering method, we can create a visual overview of all functional relationships in the research database. We visualize the nearest neighbor of each cluster centroid as a scatterplot, to show the functional relationship it represents. Since we can compute interestingness for each centroid, we are also able to sort the visualized scatterplots in a meaningful way.

Of course this overview is only a staring point, and details-on-demand are available to users by selecting one of the

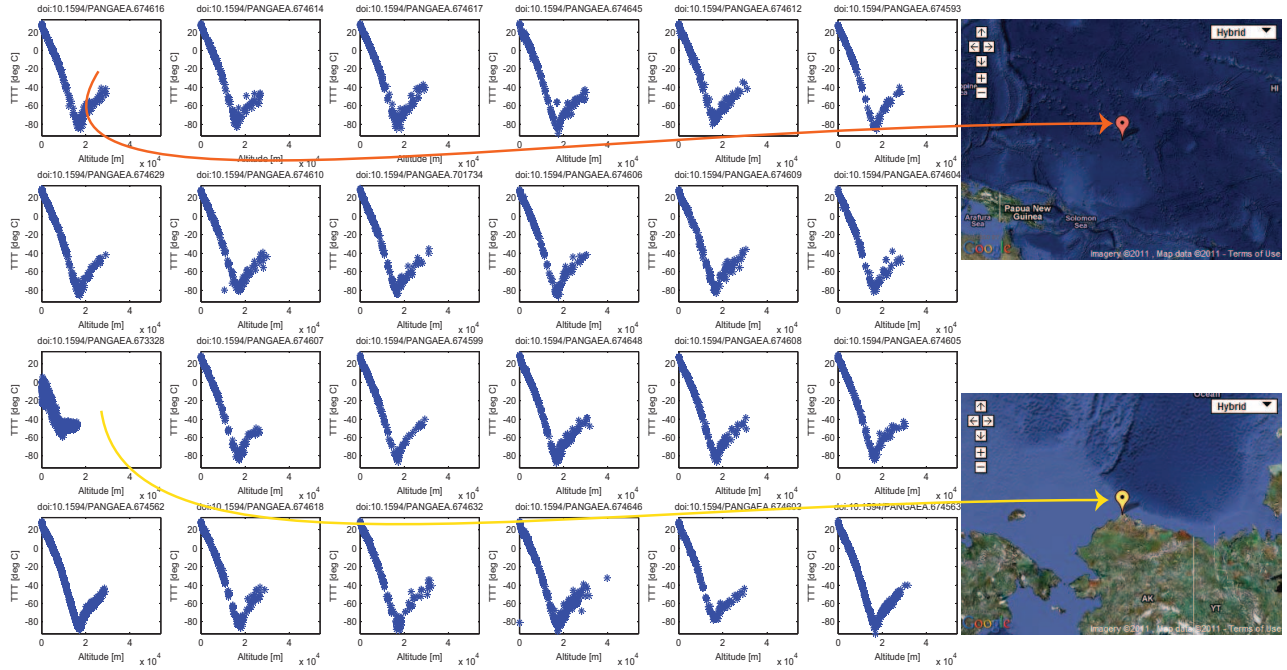

Figure 5: Retrieval Results using query-by-interestingness for *altitude* vs. *deg C* (temperature in degrees Celsius) in dataBSRN. The most interesting scatterplots depicting a relationship between those two variables are presented. The geo-references reveal, why one of the scatterplots deviates from the rest. It shows temperatures measured in northern Alaska [10] (both map images are attributed to Google, Imagery and Map data). Additonally we see that temperature is a linear falling function of altitude within the troposphere (17km). Then it rises (with the root or logarithm) of altitude, after the measurement probe has left the troposphere and is approaching the sun.

centroids. The top-most interesting scatterplots, that were assigned to this centroid are then visualized. During this step, we can optionally normalize all presented scatterplots *globally*, so that the user can easily spot differences in numeric values among these plots (since they usually are of the same functional form).

There are two ways to incorporate metadata (the variable names) into the exploratory search.

First, to restrict the exploratory search to certain variables or a particular functional relationship, we offer the query-by-formula technique (from retrieval) to filter the datasets before clustering. This allows for filtered clustering to create an overview of all functional relationships in the data or the functional relationships between one particular variable and all the others. Entering one or two particular variables filters out all those scatterplots not depicting these variables by textual metadata comparison. Supplying a specific functional relationship ranks all scatterplots according to their similarity, and then filters the datasets for the most similar ones (top 20%) and clusters only those. By entering the wildcard operator $*$ for either variable or function we avoid prefiltering. Figure 6 shows cluster results for exploratory searches, along with the applied prefilter commands.

The second way to incorporate metadata follows a different paradigm, to enable users to explore the *most interesting* variable combinations in the data. By using regressional features' interestingness measure, we rank all variable combinations with the same label (as identified by their metadata annotation, e.g., all combinations with the label *depth* *water* vs. *salinity*) according to their aggregated interestingness. This intermediate result gives important insight into the data simply by textual means (listing the interesting combinations), but also allows to create a cluster overview of the most interesting variable combinations as described before. Figure 7 illustrates an example.

5. A DIGITAL LIBRARY FRAMEWORK FOR CONTENT-BASED ACCESS TO MULTIVARIATE RESEARCH DATA

In the previous section we introduced two research datasets and showed some qualitative results for retrieval and exploratory search. In this subsection we propose a framework how research datasets can be integrated into a Digital Library system, to make content-based access, in conjunction with annotated metadata, available to users.

Figure 8 shows the proposed framework. The upper left part, labeled as *annotation-based*, provides access to research data in a way similar to other kinds of documents, by following an established library processing chain.

Data is gathered and annotated by a researcher, submitted to the library and after quality control by a curator made accessible through indexing the textual metadata. Thus, by means of textual querying, other scientists can retrieve data of interest.

On the upper right in Figure 8, labeled as *content-based*, we see the application of regressional features to index the content of the research database as described before. This

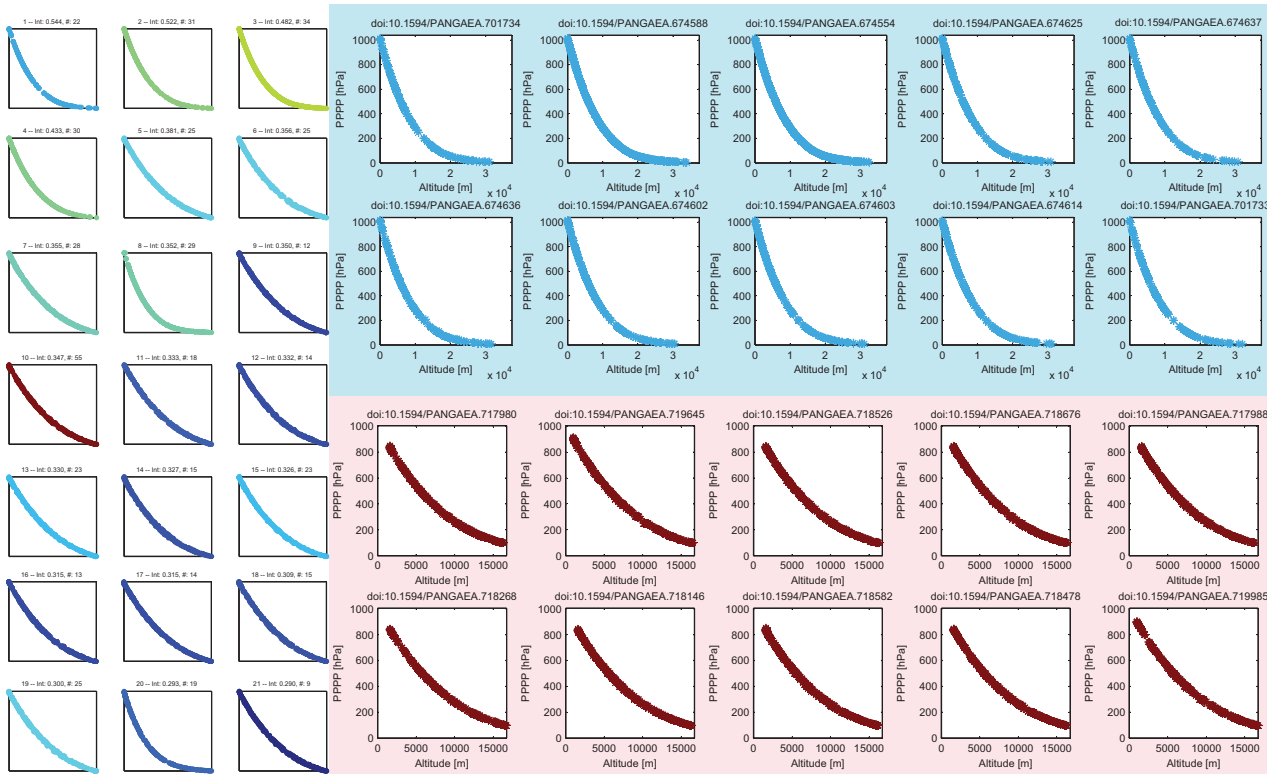

Figure 6: Exploratory search results for `CLUSTER ALL DATASETS WITH` $v_{\text{hpa}} = f_{e^{-x}}(v_{\text{altitude}})$ in `dataBSRN`. We assume an inverse exponential relationship between atmospheric pressure and altitude and see a cluster overview of all datasets that support this assumption on the left. By selecting clusters #1 and #10, we see the plots and DOI references of actual datasets, that can be cited to support this assumption or to compare one's own data against.

allows the aforementioned, content-based query modalities of retrieval and exploratory search as illustrated on the lower part of the figure. By using an intuitive query interface to specify variables and a functional relationship of interest, we allow for access to the data in a content-based way. We note that this content-based approach benefits from available *high-quality metadata*. With correct, meaningful descriptions of the variables, semantically meaningful content-based access is supported, and retrieval results can be easily interpreted by the domain expert.

6. CONCLUSION AND FUTURE WORK

Data is an increasingly decisive factor in scientific research and industrial applications. It represents a valuable asset and if made accessible in a transparent and user friendly way, can improve the scientific process as a whole. Digital Library support for research data is therefore highly desirable. We presented a novel approach for content-based and exploratory search in repositories of research data based on a similarity concept relying on functional dependencies between pairs of variables in a data set. This is a key data aspect in multivariate data. Our application on a large, real-world data set showed the utility of our approach to support content-based access.

Functional dependencies are one important, yet not the only key aspect in data collections. Our approach is a first step in supporting effective retrieval in research data repositories. Future work needs to consider complementary content-based search methods for data repositories, to be applicable to a user community as large as possible. To this end, we want to consider not only bivariate, but also multivariate dependencies in the future. Our approach currently abstracts from temporal aspect of data. Future work will consider approaches to include these aspects in the data description. To this end, a large research and design space exists, and we expect that completely new user interfaces need to be designed as well, for query specification and result visualization.

Acknowledgements.

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.

7. REFERENCES

[1] Baseline Surface Radiation Network (BSRN). http://www.bsrn.awi.de/.

[2] P. Berkhin. A survey of clustering data mining techniques. *Grouping Multidimensional Data*, pages 25–71, 2006.

[3] J. Bernard, J. Brase, D. W. Fellner, O. Koepler,

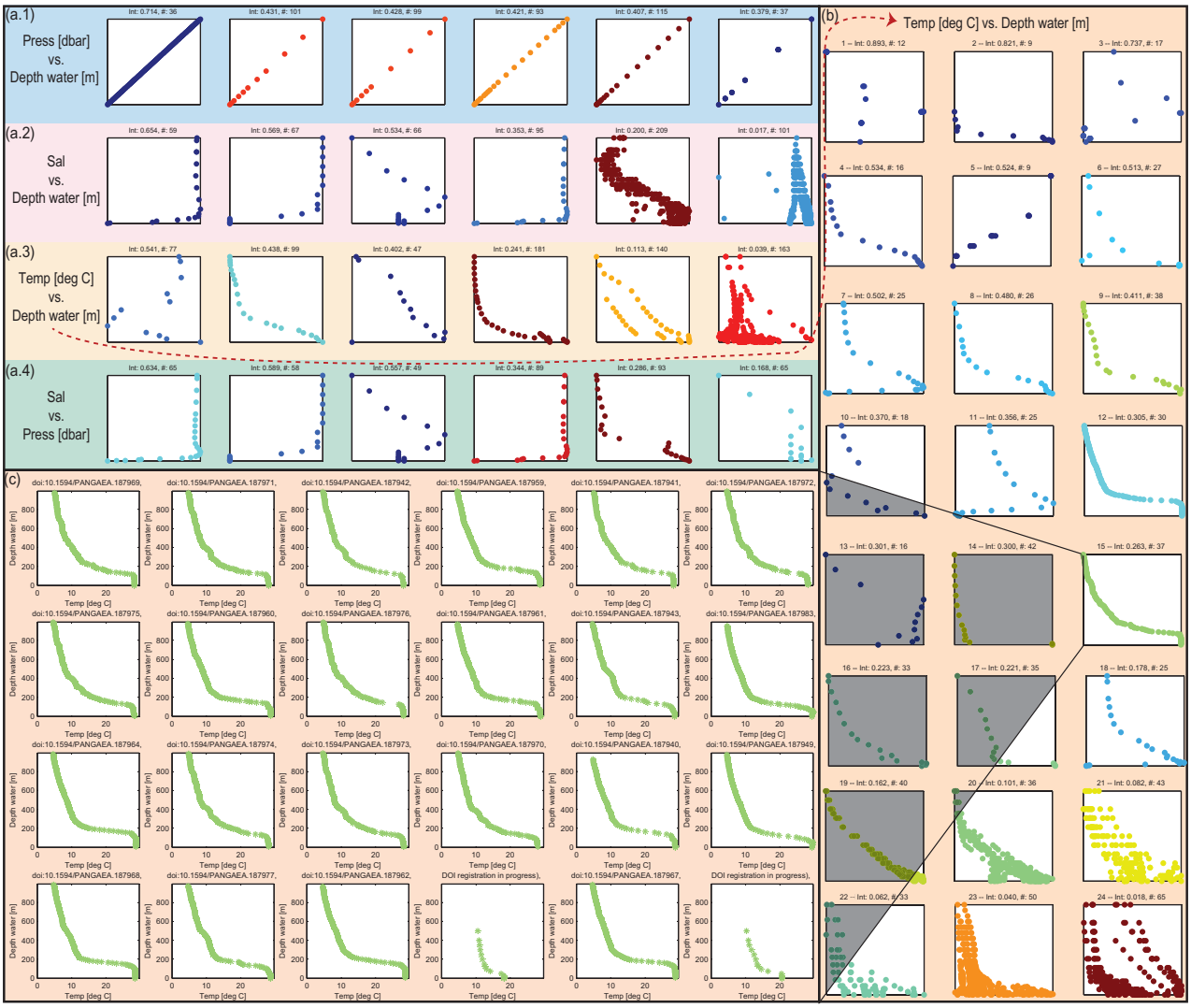

Figure 7: Exploratory search of the most interesting variable combinations in `dataMixed`, see (a.1) to (a.4) (grouped by annotation, sorted by Eq. 1). We select the 3rd variable combination (*temperature* vs. *depth*) for further details and see a detailed clustering in (b). Cluster #15 is inspected further in (c).

J. Kohlhammer, T. Ruppert, T. Schreck, and I. Sens. A visual digital library approach for time-oriented scientific primary data. In *ECDL*, pages 352–363, 2010.

[4] R. Berndt, I. Blümel, M. Clausen, D. Damm, J. Diet, D. W. Fellner, C. Fremerey, R. Klein, F. Krahl, M. Scherer, T. Schreck, I. Sens, V. Thomas, and R. Wessel. The probado project - approach and lessons learned in building a digital library system for heterogeneous non-textual documents. In *ECDL*, pages 376–383, Sept. 2010.

[5] J. Brase. DataCite – A global registration agency for research data. In *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, pages 257–261, 2009.

[6] D. Castelli and P. Pagano. Opendlib: A dl service system. In *ECDL*, 2002.

[7] P. Daras, D. Tzovaras, S. Dobravec, J. Trnkoczy, A. Sanna, G. Paravati, R. Traphoener, J. Franz, T. Kastrinogiannis, C. Malavazos, N. Ploskas, M. Gumz, K. Geramani, and G.-J. Wintterle. Victory: a 3d search engine over p2p and wireless p2p networks. In *4th International Conference on Wireless Internet*, 2008.

[8] E. Dekel, G. Ellison, J. Horowitz, C. Meghir, and A. Postlewaite. The econometric society annual reports report of the editors 2008-2009. *Econometrica*, 78(1):433–436, 2010.

[9] Dryad Digital Repository for Data Underlying Published Works. http://www.datadryad.org/.

[10] E. G. Dutton. Radiosonde measurements from station barrow (1992-05). In *Climate Monitoring & Diagnostics Laboratory, Boulder*, Baseline Surface Radiation Network. http://dx.doi.org/10.1594/PANGAEA.673328, 2007.

[11] ELIXIR European Life Sciences Infrastructure for Biological Information. http://www.elixir-europe.org/.

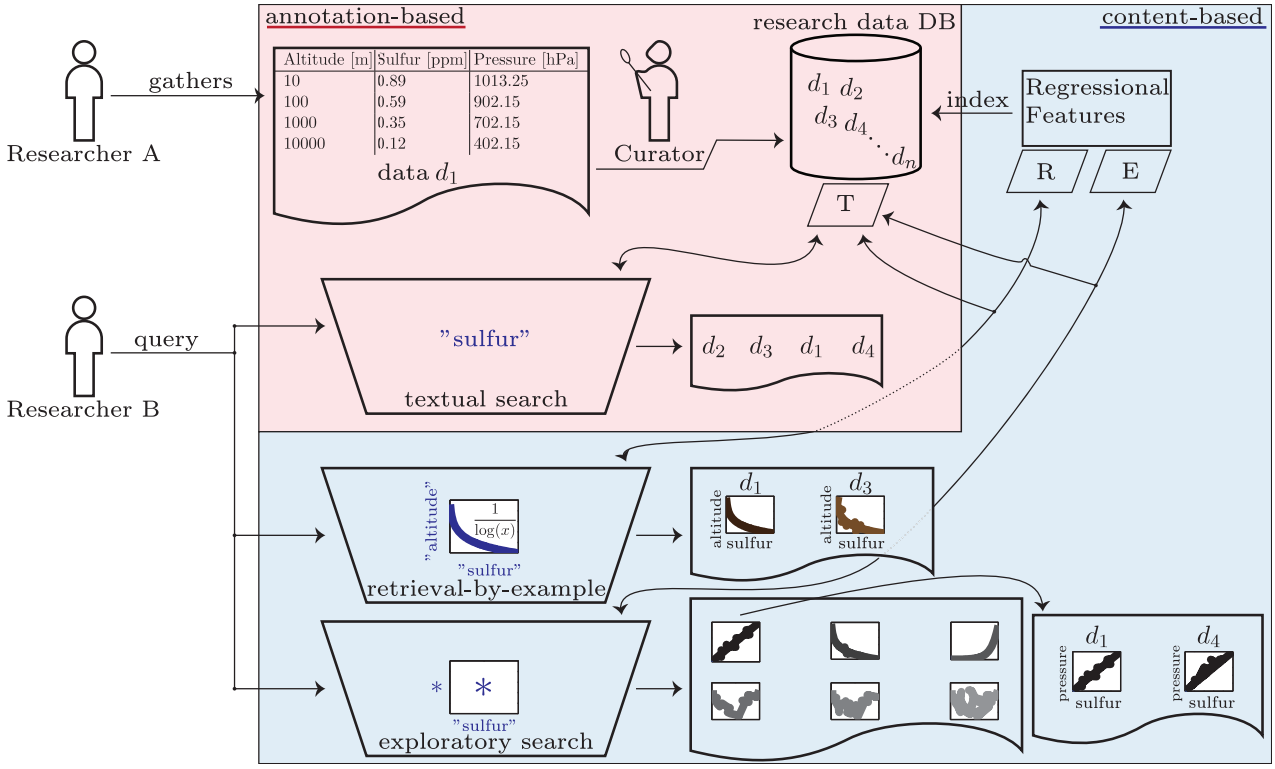

Figure 8: Framework to support content-based access to research data in a Digital Library system. On top of established techniques and practices for annotation-based access (light-red background), content-based access (light-blue background) to research data itself is provided via regressional features.

[12] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 82–88, 1996.

[13] G. Hébrail, B. Hugueney, Y. Lechevallier, and F. Rossi. Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomput.*, 73(7-9):1125–1141, 2010.

[14] T. Kohonen. *Self-Organizing Maps*. Springer, 3rd edition, 2001.

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

[16] L. J. Latecki, R. Lakämper, and U. Eckhardt. Shape descriptors for non-rigid shapes with a single closed contour. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 424–429, 2000.

[17] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg. Comparative analysis of multidimensional, quantitative data. In *IEEE Transactions on Visualization and Computer Graphics*, pages 1027–1035, 2010.

[18] T. W. Liao. Clustering of time series data-a survey. *Pattern Recognition*, 38:1857–1874, 2005.

[19] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[20] PANGAEA Publishing Network for Geoscientific & Environmental Data. http://www.pangaea.de/.

[21] PsychData National Repository for Psychological Research Data. http://psychdata.zpid.de/.

[22] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd edition, June 2005.

[23] C. Schaffer. Bivariate scientific function finding in a sampled, real-data testbed. *Mach. Learn.*, 12(1-3):167–183, 1993.

[24] M. Scherer, J. Bernard, and T. Schreck. Reference list of sources used for two experimental data files databsrn and datamixed. In *Publishing Network for Geoscientific & Environmental Data*. http://dx.doi.org/10.1594/PANGAEA.756307, 2011.

[25] T. Schreck, J. Bernard, T. Von Landesberger, and J. Kohlhammer. Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, 8:14–29, January 2009.

[26] L. Todorovski and S. Dzeroski. Integrating domain knowledge in equation discovery. In *Computational Discovery of Scientific Knowledge*, pages 69–97, 2007.

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