Semantic Retrieval Interface for Statistical Research Data

Daniel Bahls, Klaus Tochtermann

Leibniz Information Centre for Economics (ZBW), Kiel, Germany

Abstract. Statistical research data is the foundation for empirical studies. Researchers in economics or social sciences often obtain such data from external sources through specially designed retrieval interfaces from statistical offices, commercial data providers as well as from data agencies and other archives. With the advancements in data cataloguing and acquisition of long tail research data sets from individual scientists and institutes, the opportunity is there to install central services for a more holistic data search. In view of a rapid increase in amount of data available and by association an emerging retrieval problem, retrieval interfaces must make effective use of provided metadata in order to help find relevant data sets efficiently.

This paper presents a multi-step retrieval interface that aims to support the researchers' natural approach to data search and composition. Starting with an idea of the concepts that are to be compared, users kick off their search with thesauri terms and successively specify requirements according to their priorities until suitable data can be selected easily from a manageable number of matching data sets. The prototype presented in this paper also provides means for convenient data harmonization, which is an essential aspect especially when combining statistical data from different sources.

Keywords: Research Data Management, Semantic Digital Data Library, Linked Data, Statistics, Data Retrieval

1 Introduction

A significant number of scientific results are based on research data, since research has become increasingly data-driven over the years [1]. Therefore, to understand such scientific publications in depth, documentation on underlying data is a necessary means. To further provide transparency and enable replicability in the end, respective data sets must be available as such, for which a reliable infrastructure is required. Scientific data needs to be maintained and organized in archives.

With the advancement of computer technology, scientific analyses are more and more carried out with the aid of machines, as it allows for large amounts of data being processed in short amount of time which has never been possible before. While this certainly is one reason why science has become significantly data-driven, it also leads to the fact that most scientific data is maintained in digital form already. This circumstance and the rise of the Web opens up possibilities for a powerful information infrastructure for supporting these aforementioned goals. Information resources nowadays can be delivered to any place in the world within seconds, laying the ground for delivering the right information to the right place at the right time, the precept of knowledge management.

The Web together with its well-established Web 2.0 technologies has already been recognized as a powerful media for promoting efficient exchange and advancement in the scientific domain. In this regard, the Leibniz Association has recently started the research alliance Science 2.0[1] with a growing number of 30 associated institutes to jointly venture into a well-organized and integrated environment of Web-based tools and services for the scientific community to support rapid exchange and good scientific practice.

The vision of a thought-out research data infrastructure fits well into this theme, and many initiatives have formed in the last years, a whole movement to effectively enable exchange, citation and preservation of research data. However, this task has proven non-trivial, as it opened up exhaustive discussions on metadata schemes[2], organized preservation and curation [2], responsibilities [3], data publication policies [4] as well as solutions to overcome issues of data protection and usage rights, only to mention a few. Yet, these efforts have already lead to significant advancements (TheDataHub[3], DataCite[4], and other).

At present, efforts are being made to pick up research data as bibliographic artifacts for re-use, transparency and citation[5]. In view of a rapid increase in amount of data available and by association an emerging retrieval problem, retrieval interfaces must make effective use of provided metadata in order to help find relevant data sets efficiently.

In this paper, we investigate how to make use of Semantic Web technologies for providing an efficient and novel approach for the retrieval of statistical data sets that follows a natural approach for data retrieval in the domain of statistics, particularly in the context of economics or the social sciences. Section 2 elaborates on the practice of data acquisition in empirical research to gain a clear picture on the purpose of our system. Related work is discussed in the subsequent section, and Section 4 explains fundamental design decisions and outlines a system architecture. Section 5 describes the user interface itself and how the declared goals have been implemented into features. The paper eventually closes with conclusions and outlook.

[1] http://www.leibniz-science20.de

[2] particularly important, as in contrast to textual publications, data cannot be understood without documentation

[3] http://datahub.io/

[4] http://www.datacite.org/

2 Retrieving Statistical Data

In many cases, empirical researchers in economics and the social sciences are to put together statistical indicators in large data tables. Typically, each column represents one indicator while the rows represent respective data per year, country or other so-called dimension. The data itself may be self-produced in terms of studies and surveys or acquired from external sources such as statistical offices, affiliated institutes or purchased from commercial data providers. However, common practice is to combine several sources, since some indicators may be obtained from one source while the data for other indicators may be obtained from another one. In this regard, researchers have to be extra careful to make sure respective data represents the same or sufficiently similar statistical population.

To gain a clear picture of the goals of this research, we need to clearly understand the purpose of the system. We have conducted interviews with economic scientists which helped us gain insights in their work with research data. Empirical researchers typically start out with an idea of concepts relevant in their research (e.g. living standards, work conditions, economic growth, etc.). In addition, they have further details in mind, for instance on reference periods, regions to be included and distinguished or frequency of data acquisition in case of time series data. As a result, the data set should be as consistent as possible with respect to acquisition method, statistical universe and adjustments. To achieve user acceptance, the system has to be practical in research settings [6], and therefore we aim to support this data harmonization procedure in a light-weight manner.

As a result, user communication should follow the below steps:

1. Prompt for a list of concepts that are to be compared
2. Let user specify additional requirements on the data
3. Let explore and select matching data sets, allow for revisiting Step 2
4. Offer selected data for download

After finishing Step 1, data sets associated with the concepts named should be presented to the user. Specification of additional requirements should be based on the metadata available for the data sets found. As soon as all relevant requirements are given, the user may inspect and decide on these satisfying data sets and proceed to download at last.

3 Related Work

There are many repositories on the Web that provide statistical data. Some of them are provided by statistical offices and data agencies (e.g. Federal Statistical

Office of Germany[5], EuroStat[6], World Bank[7]), some are associated with commercial providers (e.g. Thomson Reuters Datastream[8], Statista[9]) and yet others are maintained by journals, archives, libraries or independent organizations (e.g. GESIS[10], The Data Hub[11], Dataverse repository of Economists Online[12]). All of these portals are as heterogeneous as the kind and spectrum of data they provide. Some of them provide interfaces for composition of customized data tables where users pick and choose indicators and data records according to their needs. Such features are also provided by the Nesstar system[13], one of the most prominent systems for data publishing and online analysis that is being used by a large number of institutes. The Social Science Variables Database at ICPSR[14] allows for direct comparison of indicators with respect to a variety of metadata, giving intuitive means to understand differences in universe, acquisition method and other between data sets. However, users of these systems are to run keyword-based queries and browse through category trees in order to find relevant data sets individually, and therefore our approach follows a different paradigm as presented in Section 2.

Technical challenges in dealing with distributed sources and applying the OLAP paradigm for retrieval of statistical data from the Linked Data cloud have been addressed in [7]. We view this work as a major contribution for building a scalable backend, whereas our work aims to provide a user interface and communication design for data search and retrieval within the specific setting research data sharing.

Other approaches are based on semantic links between data sets and research articles [8] which give textual context for otherwise sparsely described data content and therefore improve data search by established Information Retrieval techniques. These data links, typically given by persistent identifiers, however, point to entire data bundles as a whole, whereas our approach aims to make single indicators and values available for retrieval.

4 System Architecture

Following the steps presented in Section 2, we elaborate on the system architecture of our data retrieval system. To support Step 1, a thesaurus should be used, so that data sets associated with a particular concept can be found easily. To enable the specification of requirements, metadata must be given in detail

[5] https://www.destatis.de

[6] http://epp.eurostat.ec.europa.eu

[7] http://data.worldbank.org

[8] http://online.thomsonreuters.com/datastream/

[9] http://de.statista.com

[10] http://www.gesis.org/en/

[11] http://thedatahub.org

[12] http://dvn.iq.harvard.edu/dvn/dv/NEEO

[13] http://www.nesstar.com

[14] http://www.icpsr.umich.edu

and in association with individual indicators and records rather than a separate metadata block for a zipped data bundle. This enables the system to make sense of the data in depth and allow for requirement specification as explained later in Section 5.

The research on a data retrieval interface is part of our overall research activities on an infrastructure for scientific data for the field of economics. For several reasons we regard Semantic Web technologies most suitable for this purpose, among which is strength in dealing with distributed data and extensibility, which is required whenever highly specific long tail data from individual researchers needs additional vocabulary for description [9]. However, the data format should provide for typical data types, such as floats, strings, dates and other. It must provide metadata on fine-grained level as to open up possibilities for retrieval and composition. As a consequence, the retrieval system operates on statistical data in the format of the RDF Data Cube Vocabulary[15] [10].

The prototype was implemented in Java and JavaScript under the use of the Play Framework[16]. The live system was tested on an Apache Tomcat[17] and a Sesame Triple Store[18], as the system operates on statistical data provided as RDF using the RDF Data Cube Vocabulary[19] [10].

5 User Interface Design

The system implements a multi-step retrieval interface as described in Section 2. In the following, we are going to refer to the screenshots given in Figure 1 to 8 in parantheses. Since the expected result is a data table after all, the main screen starts with an empty spreadsheet (1). For Step 1, the user successively enters the names of the concepts that are to be compared in the empty column headers as shown in (2). This task is supported by autocompletion on the basis of concept terms contained in a thesaurus, STW[20] in our case. With the selection of a concept, the system displays the number of associated data sets beneath the concept label entered before. A click on this number lists all of them in alphanumerical order (3), and another click reveals a detailed description and further information on the particular data set (7). Yet, at this point, the number of data sets might be huge, and the user may decide to formulate requirements for the data first as per Step 2. With the selection of a single column header, the panel on the left lists down the *union* over all properties and property values available in the metadata of all the data sets associated with the concept of the column (4). Hovering over a property or property value produces an info box with documentation on the vocabulary. Selecting a particular property value specifies a requirement and tells the system that only those data sets are relevant for

[15] `http://www.w3.org/TR/vocab-data-cube/`

[16] `http://www.playframework.org`

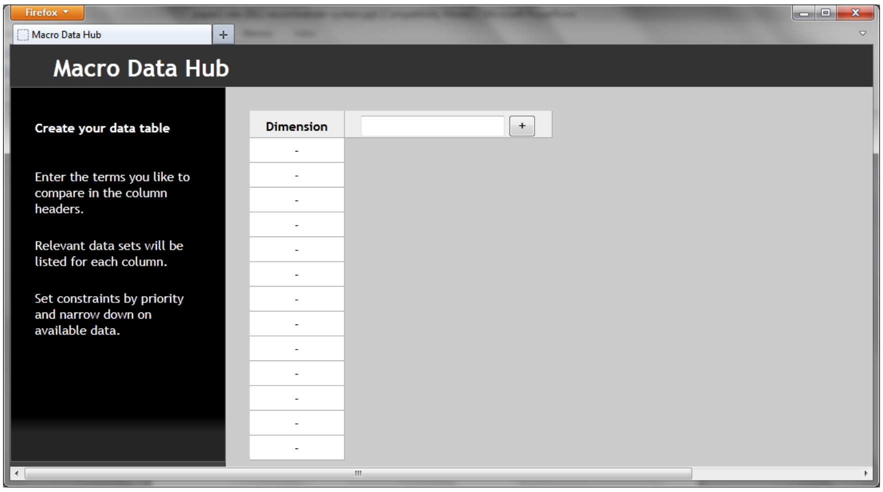
[17] `http://tomcat.apache.org`

[18] `http://www.aduna-software.com/technology/sesame`

[19] `http://www.w3.org/TR/vocab-data-cube/`

[20] STW Thesaurus for Economics, `http://zbw.eu/stw/`

this column that provide this respective property and property value, and the number of relevant data sets drops. With the selection of two or more column headers, the panel on the left shows the *intersection* between the properties and values of the single columns (5). This feature facilitates harmonization of data, as it reveals which data characteristics can be unified among the columns. To specify the contents of the rows, one must specify the *Dimension* property. A click on the respective header highlights all column headers of the entire table as to indicate that the property of choice must be available in the data sets of all columns. The user selects (multiple) values from the properties listed on the left and the Dimension column fills accordingly (6). This again sets requirements for the data sets, as it filters all data sets that do not provide respective records. Eventually, when all requirements are set, the user examines and selects from the remaining list of data sets for each column (7). If all remaining properties with multiple options are bound to a value, the table fills with actual data content (8). As a last step, the table is offered for download.



Create your data table

Enter the terms you like to compare in the column headers.

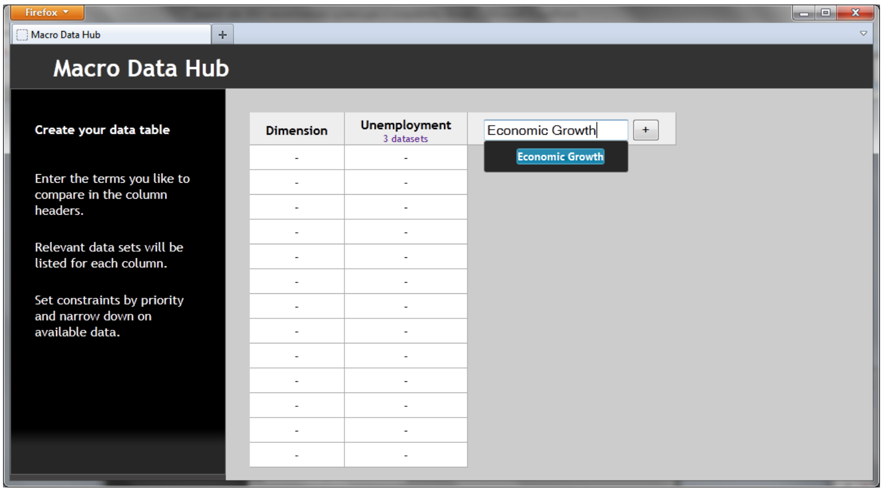
Relevant data sets will be listed for each column.

Set constraints by priority and narrow down on available data.

Fig. 1.

6 Conclusions and Outlook

Following the call for a research data infrastructure, we have addressed the issue of data retrieval for the domain of economics and social sciences where large amounts of scientific results are based on statistical data. With the prospect of a rapidly growing amount of data from individual researchers and institutes filed in the future, overviewing all relevant data sets efficiently becomes a problem. For this purpose, we have designed an innovative retrieval interface that aims



Macro Data Hub

Create your data table

Enter the terms you like to compare in the column headers.

Relevant data sets will be listed for each column.

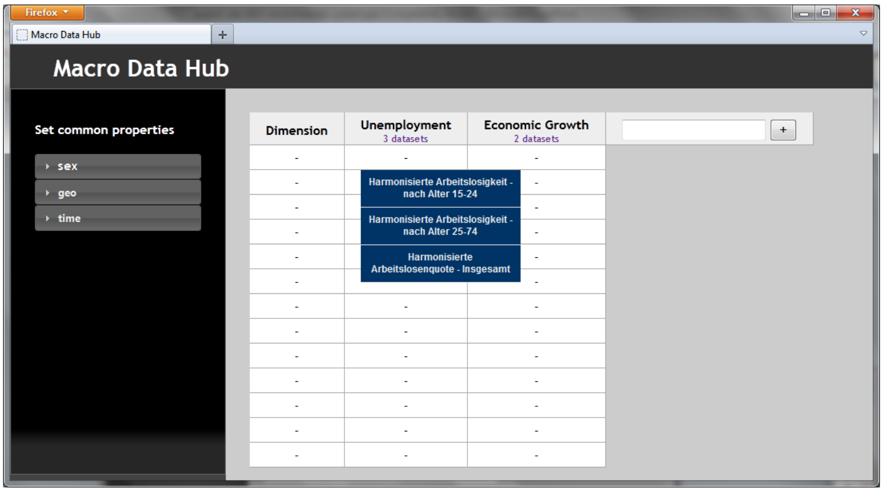
Set constraints by priority and narrow down on available data.

Dimension	Unemployment 3 datasets
-	-
-	-
-	-
-	-
-	-
-	-
-	-
-	-
-	-
-	-
-	-
-	-
-	-

Economic Growth +

Economic Growth

Fig. 2.



Macro Data Hub

Set common properties

- sex
- geo
- time

Dimension	Unemployment 3 datasets	Economic Growth 2 datasets
-	-	-
-		-
-		-
-		-
-		-
-		-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-

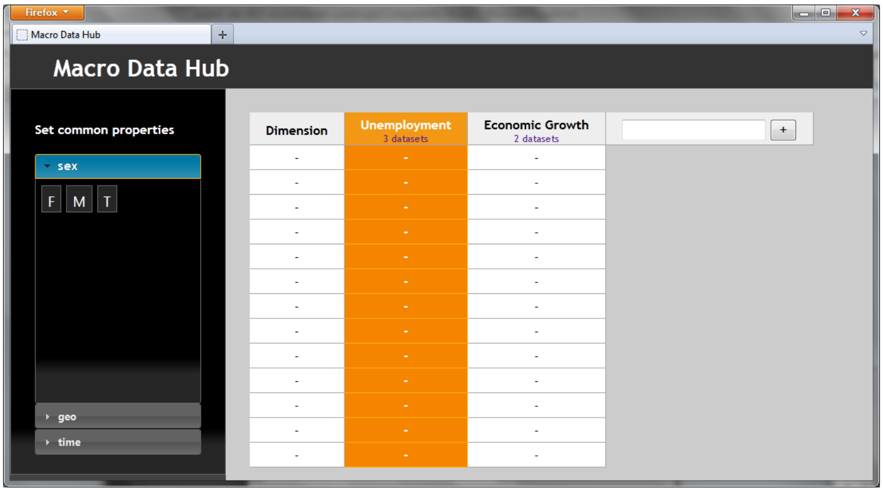
Harmonisierte Arbeitslosigkeit - nach Alter 15-24

Harmonisierte Arbeitslosigkeit - nach Alter 25-74

Harmonisierte Arbeitslosenquote - Insgesamt

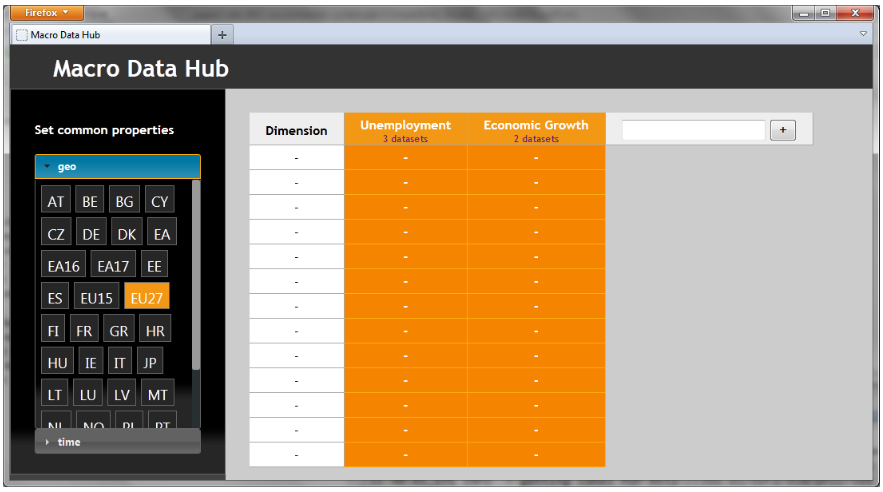
+

Fig. 3.



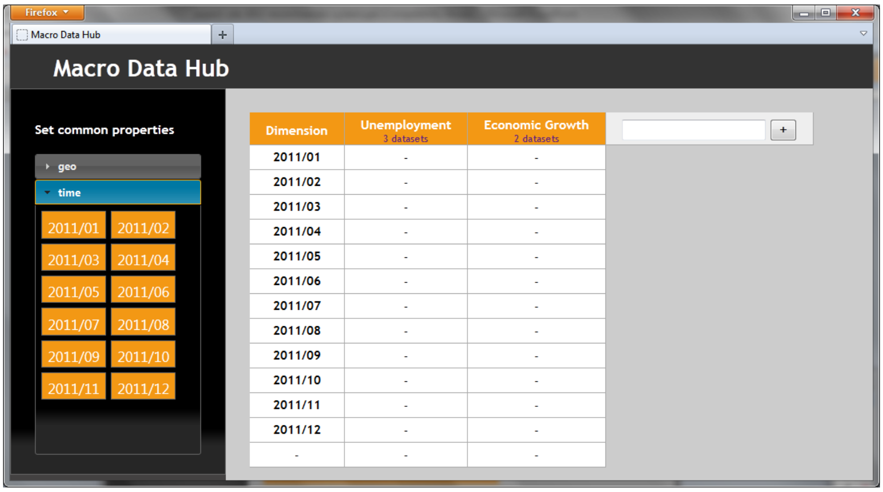
Dimension	Unemployment 3 datasets	Economic Growth 2 datasets
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-

Fig. 4.



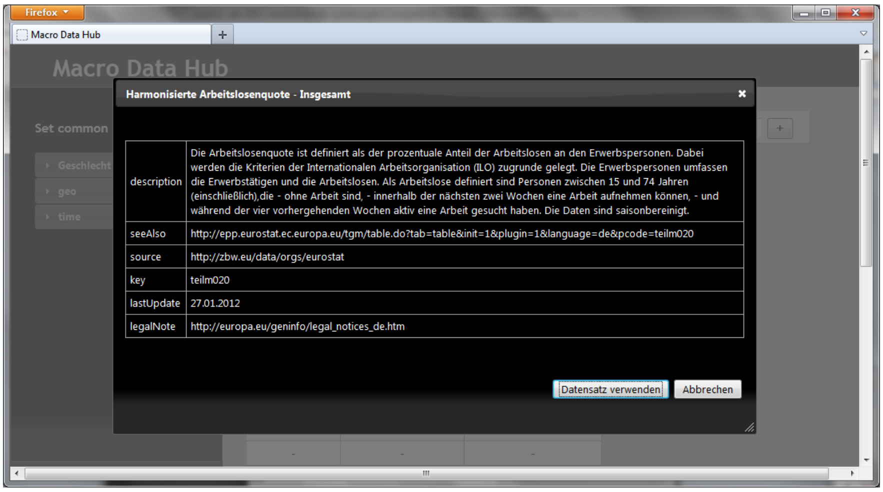
Dimension	Unemployment 3 datasets	Economic Growth 2 datasets
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-

Fig. 5.



Dimension	Unemployment 3 datasets	Economic Growth 2 datasets
2011/01	-	-
2011/02	-	-
2011/03	-	-
2011/04	-	-
2011/05	-	-
2011/06	-	-
2011/07	-	-
2011/08	-	-
2011/09	-	-
2011/10	-	-
2011/11	-	-
2011/12	-	-
-	-	-

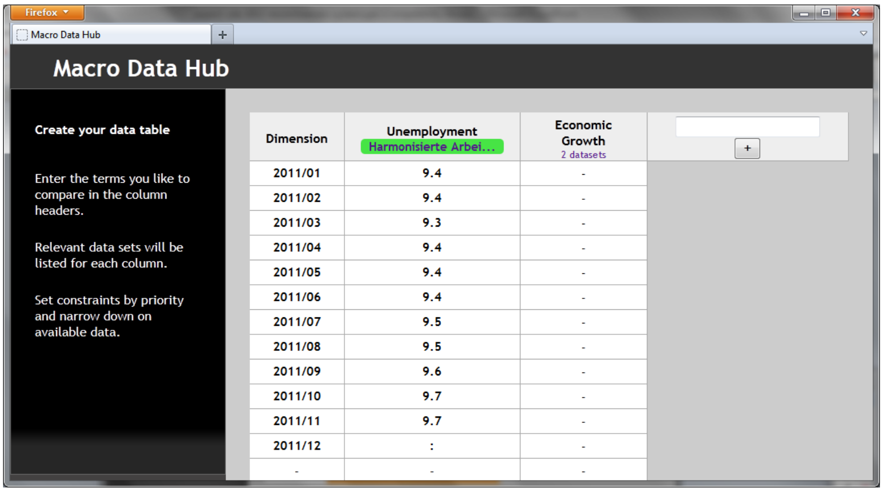
Fig. 6.



Harmonisierte Arbeitslosenquote - Insgesamt

description	Die Arbeitslosenquote ist definiert als der prozentuale Anteil der Arbeitslosen an den Erwerbspersonen. Dabei werden die Kriterien der Internationalen Arbeitsorganisation (ILO) zugrunde gelegt. Die Erwerbspersonen umfassen die Erwerbstätigen und die Arbeitslosen. Als Arbeitslose definiert sind Personen zwischen 15 und 74 Jahren (einschließlich),die - ohne Arbeit sind, - innerhalb der nächsten zwei Wochen eine Arbeit aufnehmen können, - und während der vier vorhergehenden Wochen aktiv eine Arbeit gesucht haben. Die Daten sind saisonbereinigt.
seeAlso	http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&plugin=1&language=de&pcode=teilm020
source	http://zbw.eu/data/orgs/eurostat
key	teilm020
lastUpdate	27.01.2012
legalNote	http://europa.eu/geninfo/legal_notices_de.htm

Fig. 7.



Macro Data Hub

Create your data table

Enter the terms you like to compare in the column headers.

Relevant data sets will be listed for each column.

Set constraints by priority and narrow down on available data.

Dimension	Unemployment Harmonisierte Arbei...	Economic Growth 2 datasets	+
2011/01	9.4	-	
2011/02	9.4	-	
2011/03	9.3	-	
2011/04	9.4	-	
2011/05	9.4	-	
2011/06	9.4	-	
2011/07	9.5	-	
2011/08	9.5	-	
2011/09	9.6	-	
2011/10	9.7	-	
2011/11	9.7	-	
2011/12	:	-	
-	-	-	

Fig. 8.

to support researchers in finding and composing data sets according to their natural way of approaching a research question. The prototype presented in this paper provides simple means for data harmonization to enable consistency within statistical population in intuitive ways. Under the use of these features, we expect a significant decrease of time needed for data search and composition in comparison to the current practice, although this is yet to be evaluated.

Future improvements of the system should include retrieval from distributed sources, as this version operates on a single triple store endpoint only Moreover, the advantages of using subproperty relations should be investigated and made available to the user. Many other valuable ideas for improvements can be found with regard to user assistance, e.g. warning notifications when selected time series data include breaks, errors or changes in acquisition method which can be derived from well-maintained metadata.

Finally, this approach needs to be tested on a large archive of various kinds of statistical data and evaluated with end users from the target group of empirical researchers.

References

1. Gray, J.: Jim Gray on eScience: A Transformed Scientific Method (January 2007)
2. Treloar, A., Harboe-Ree, C.: Data management and the curation continuum: how the Monash experience is informing repository relationships. Proceedings of VALA 2008 (2007)
3. Rümpel, S.: Data Librarianship : Anforderungen an Bibliothekare im Forschungsdatenmanagement (2010)
4. Vlaeminck, S., Siegert, O.: Welche rolle spielen forschungsdaten eigentlich für fachzeitschriften? eine analyse mit fokus auf die wirtschaftswissenschaften. Technical report, German Council for Social and Economic Data (RatSWD) (2012)

5. Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D.R., Los, W., Marinucci, M., Romary, L., Van de Sompel, H., Vigen, J., Wittenburg, P., Giaretta, D.: Riding the wave: How Europe can gain from the rising tide of scientific data. European Union (2010) Final report of the High Level Expert Group on Scientific Data: A submission to the European Commission.
6. Feijen, M.: What researchers want - a literature study of researchers' requirements with respect to storage and access to research data (February 2011)
7. Kämpgen, B., Harth, A.: Transforming statistical linked data for use in olap systems. In: Proceedings of the 7th international conference on Semantic systems, ACM (2011) 33–40
8. Boland, K., Ritze, D., Eckert, K., Mathiak, B.: Identifying references to datasets in publications. In Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F., eds.: Theory and Practice of Digital Libraries. Volume 7489 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 150–161
9. Bahls, D., Tochtermann, K.: Addressing the long tail in empirical research data management. In: Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies. i-KNOW '12, New York, NY, USA, ACM (2012) 19:1–19:8
10. Cyganiak, R., Field, S., Gregory, A., Halb, W., Tennison, J.: Semantic statistics: Bringing together sdmx and scovo. In Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M., eds.: LDOW. Volume 628 of CEUR Workshop Proceedings., CEUR-WS.org (2010)