Addressing the Long Tail in Empirical Research Data Management*

Daniel Bahls
Leibniz Information Centre for Economics (ZBW)
Neuer Jungfernstieg 21
20354 Hamburg, Germany
d.bahls@zbw.eu

Klaus Tochtermann
Leibniz Information Centre for Economics (ZBW)
Düsternbrooker Weg 120
24105 Kiel, Germany
k.tochtermann@zbw.eu

ABSTRACT

At present, efforts are being made to pick up research data as bibliographic artifacts for re-use, transparency and citation. When approaching research data management solutions, it is imperative to consider carefully how filed data can be retrieved and accessed again on the user side. In the field of economics, a large amount of research is based on empirical data, which is often combined from several sources such as data centers, affiliated institutes or self-conducted surveys. Respecting this practice, we motivate and elaborate on techniques for fine-grained referencing of data fragments as to avoid multiple copies of same data archived over and over again, which may result in questionable transparency and difficult curation tasks. In addition, machines should have a deeper understanding of the given data, so that high-quality services can be installed.

The paper first discusses the challenges of data management for the management of research data as used in empirical research. We conclude a comparison of referencing and copying strategies and reflect on their implications respectively. As a result from this argumentation, we elaborate on a data representation model, which we further examine in regard to considerable extensions. A Generating Model is subsequently introduced to enable citation, transparency and re-use. Eventually, we close with the demonstration of an explorative prototype for data access and investigate a distance metric for assisting in finding similar data sets and evaluating existing compositions.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Miscellaneous; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic networks

General Terms

Design, Algorithms, Economics, Standardization

Keywords

Research Data Management, Semantic Digital Data Library, Linked Data, Statistics

1. INTRODUCTION

With the advancement of information technologies, researchers of all areas increasingly follow the practice of data-driven research, which is often referred to as the fourth scientific paradigm [11]. Computers help conduct analyses and simulations efficiently in large scale, such that knowledge can be produced in areas that were unapproachable before. This shift to digital technology yields great advantages on the one hand, but on the other jeopardizes the strong foundation of science when unorganized data management practices compromise its fundamental principles, namely transparency and repeatability of results. The argumentation follows a rather practical sense, considering the fact that generating research data is often very costly and time-consuming, and thus hard to replicate. An exhaustive study on research data management practices across Europe gives evidence for the extent of disorganization in preserving digital research data [12].

Funding bodies all over the world have recognized this problem and came up with research agendas for building a sustainable infrastructure, in which data can be cited, shared and reviewed for good scientific practice and a higher degree of exchange and efficiency [5]. Since research data varies in size, format and after all in nature among the disciplines [4], a one size fits all solution is unlikely to bring up satisfying tools and services as requested by the scientific community [6]. In this paper, we elaborate an approach for the field of economics and specialize on empirical data as it accounts for a large share of economic research data. Due to the fact that empirical data is also used in the social sciences, the behavioral sciences and other, we attempt to drive this approach by the particular data nature and the way researchers work with it rather than narrowing down on a specific scientific discipline, whereas most of our understanding was derived from interviews with economic scientists.

2. EMPIRICAL DATA AND RESEARCH PRACTICES

We use the term *empirical data* to refer to any kind of data used in empirical research for making certain statements about a particular population, which is statistically significant with a calculable degree of confidence, be it derived from surveys or process-generated, and comprises micro as well as macro data. The concept of *population* or *universe* will be used in the statistical sense throughout the paper and denotes a group of individuals a data set represents, as for example the households in a certain country or the persons of a particular age group. This understanding is important, since populations play a significant role when combining data from different sources, which is a very common practice. Researchers do so in order to link some indicators with others for a variety of reasons, and they must take particular care as to avoid unintentional mismatches between the respective populations, which would easily bias the results.

Empirical research is based on observations, and thus based on evidence. However, many research questions address concepts that are not directly measurable like wealth, justice or level of education for instance. Access to clean water and electricity, television and radio, child labor participation and motorized mobility are often used in aggregate models for measuring the advancement of living standards in a formerly colonized country. Although a reasonable approach, it can never serve as a ground truth measure, considering that the named factors have become less significant in other societies. In addition, producing representative data is an art of its own, as its contained values are very much influenced by the acquisition method applied, which is a well-known circumstance [15]. Even if the researcher's data is perfectly valid with respect to the above issues, it yet remains a challenge to compare for example the economic situation in one country to the one of another, since cultural and other country-specific factors have to be considered, and observation data therefore often needs adjustments[1], which again is often an arguable procedure and easily exposed to subjective thinking due to the absence of accurate deductive means as they often are available in the natural sciences.

This was to mention only a few of the difficulties empirical researchers have to deal with. We consider it crucial to understand all these aspects and practices as to get into a position that allows us to carefully design an infrastructure that suits best to our community, laying the ground for a sustainable environment of valuable—and hence accepted—tools and services. From interviews with domain experts in economics, we learned that the amount of data produced by a researching facility makes about an estimated thirty percent only. Most of the data is drawn from external sources, ranging from prominent data providing agencies like OECD[2], EuroStat[3] or statistical offices to individual research institutes with data of a more specific kind. Producing such data on one's own—if possible at all—is expensive in time and money, which is in line with the statement that availability of data is a key driver for research activities.

Some empirical data is freely available while other is not, for which several reasons are to be distinguished. When data is protected for reasons of data security, some policies grant on-site access only. Even though provided data is in anonymized form, researchers may inspect sample data only to understand structure and content by which they design an analysis script, which is then carried out locally on the actual data. Script and results are therefore inspected by the agency, making sure that data protection is maintained[4]. Another reason is given by commercial interest, so that data must be paid for in order to be granted access. And yet others are related to the quite comprehensible attitude of researchers who have put a lot of work into producing and revising a particular data set and strive to be the first to publish results on it. By the time the researcher has finished her work, the data may be assumed to be not of interest anymore, and in connection with the large documentation efforts necessary to make it reusable and a missing infrastructure, the data remains unavailable to others.

The other problem with data availability is the fact that there is no central catalog, portal, whatsoever enabling to even find out about the existence of particular data. Best practice is to visit data center websites[5], the websites of familiar institutes, or make use of regular web search engines. Researchers exchange news on available data at conferences or find out about interesting ones when reading papers. However, the interviews laid open that continuous polling in all sorts of directions is their strategy to keep up to date. A clever cataloguing solution together with notification services[6] might be quite welcome.

Further details on research practices and today's situation will be explained as we proceed in the argumentation. Based on the findings, we attempt a few careful thoughts on infrastructure design, implementing solutions and possible applications in the following sections.

3. DATA REPRESENTATION

Before arguing over a particular data model, one should be clear on the kind of services and processes that are to operate on. We begin with a discussion on two different and fundamental paradigms of data representation and their implications. Subsequently, we formulate some requirements by which we then decide on the basic model. We further discuss ideas and possible extensions in the closing.

3.1 Referencing vs. Copying

Repository software usually supports uploading and retrieval of arbitrary data bundles that come in zip or other archive files and provide metadata fields of all kinds for description of content [19]. Generally, one advantage is that any kind of data can be handled that way, irrespective of its particular structure or formats. It can be considered a black box approach, as the system does not make meaning of the contents

[1] with respect to purchasing power parities or seasonal adjustment just to name a few

[2] `http://www.oecd.org/`

[3] `http://ec.europa.eu/eurostat`

[4] Meeting thresholds for cell sizes in aggregation are common criteria to make sure one cannot trace back to (small groups of) individuals

[5] browse and search interfaces vary in design and quality

[6] cf. Google Alerts

and builds retrieval and presentation techniques entirely on the provided metadata instead. As one drawback in principle, this separation may lead to inconsistencies and may also result in frustration in cases where users download a resource bundle and cannot make meaning of its contents either. A brief examination of the data sets submitted within the NEEO context [18] reveals that these experiences are shared in the economics as well. On the other side, there is the structured approach, which captures data by a specific model and thus avoids separation of content and metadata. Therefore, it demands a more or less troublesome mapping and runs the risk of being unsuitable for some data sets. We examine the implications of either approach in the following.

One item in this discussion addresses curation efforts. As we have stated earlier in this paper, the bulk of data used in empirical research is drawn from external sources and may or may not be combined with own data. As the overall target is to pick up the data used in research publications, we argue that following the black box concept leads to difficult and hardly managable curation tasks. Assuming the numbers provided by the interviewed researchers are correct, the amount of duplicate data in a data publication would make around seventy percent in average. Besides the deluge of data, one problem lies in carrying over notes, remarks or changes that were made in curation of the original data set. Assume, a researcher uses unemployment data from EuroStat in her research. After submitting paper and data, the data agency realizes a mistake in this aggregated data set which had its origin in one of the many data producing institutions in one of the EU countries. To some extent, the correction affects the results presented by our researcher. Immediately modifying her data set may not be the right action, as she is not to be blamed and her reputation should not be affected, yet a note linked with her publication would certainly be of valuable information to others. Doing so for hundreds of data sets on the basis of metadata description can be considered a high cost factor, might result in discontent among curators and is likely to lack in completeness - if done manually. A precise reference model lays the ground for exact tracking algorithms and enables powerful machine support. Furthermore, if a reference model can be designed that eliminates all plausible scenarios in which duplicate data is required, all curation tasks have to be carried out once only - right by the hands of those who are responsible, having the highest expertise in the particular field.

In our particular domain, following a black box approach would lead to another more serious problem that is of a legal kind. A good deal of the data used in empirical research is protected due to data security or simply cannot be shared with third parties due to data usage rights, partly because some of the providers are commercial. Consequently, a researcher often is not allowed to upload the entire data set as a whole to any independent data repository. Hence, whoever wants to re-use this very same data set is to collect the respective data from various sources individually. Succeeding in such endeavor again depends on several factors, and in case the expected results cannot be produced, the researcher might be wondering whether or not this is due to his own failure in correct composition, and thus, the aproach fails in achieving the goals of transparency and re-usability.

As a result of the discussion, we conclude that any infrastructure design that is based on the submission of entire data set copies falls short of providing a comprehensive data management solution. Platforms like Google's Fusion Tables [7], the Data Hub[7] and similar give excellent quick win services but are insufficient for the wide spectrum of research data of the long tail. We therefore choose the principle of referencing over the principle of copying and argue that a carefully designed reference model lays the ground for a more managable data curation and archiving infrastructure.

3.2 Summarization of requirements

As we state that every data should be hosted at one place only, the referencing model must be able to deal with highly distributed content. Moreover, since data is often re-used partially, in subsets and slices, it is not sufficient to allow for pointing to entire data sets, but rather a very fine-grained model is required in which single data items can be referenced in very detail.

In order to enable retrieval as well as clear understanding of data, metadata must be provided to describe all properties necessary for researchers to work with it. As to keep curation overhead low and to let metadata be part of any formed data subset without need for additional efforts of carrying over its documentation, the metadata should be highly integrated within the data itself, which again avoid separation of metadata and content. Since empirical data itself is not self-explaining, further more human-readable information should be attachable to it. Nevertheless, as to enable advanced machine support, such information should as well be highly machine-processible. One good example for this principle is the linking with descriptors of thesauri, which can be understood clearly by both humans and machines. Furthermore, the model should provide for specification of populations and other important features. In this matter, common practice is to provide so-called code lists, for which standardization processes have been carried out already to facilitate exchange and achieve a higher level of uniformity, whereas particular for the field of economics SDMX[8] is to mention. The Data Documentation Initiative (DDI) [9] therefore focuses on standardizing metadata models for micro data and its curation processes for the social, the behavioral and the economic sciences, which should be mappable onto the model as well.

3.3 The basic model

As a result of these considerations, we decide to use Semantic Web technologies such as RDF[10], which have been particular designed for distributed and formalized knowledge representation that is highly extensible and enables a most powerful referencing infrastructure [2]. In particular, we decide to use the RDF Data Cube Vocabulary (QB)[11] because of its generic yet detailed model which has been carefully

[7] `http://thedatahub.org/`

[8] Statistical Data and Metadata eXchange language, `http://sdmx.org/`

[9] `http://www.ddialliance.org/`

[10] Resource Description Framework, `http://www.w3.org/TR/rdf-primer/`

[11] Still in draft stage, `http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html`

designed, its integration of the above mentioned standards and the increasing acceptance within the community. Research and experiences with similar approaches particularly applied to economic statistics [3] [10] as well as a detailed paper recently published [8] with focus on the social sciences are good readings to get a clearer picture of the modelling details and design decisions, which are very much in line with our argumentation.

3.4 Proxies and empirical models

Empirical research practices, as pointed out in Section 2, involve the use of indicators for making assertions on hidden, not directly measurable variables. We have indicated that such indicators are often combined in a model to yield more meaningful results, as they all are supposed to correlate in some way with the variable that is intended to be measured. While in the earlier example on living standards the target variable cannot be measured, there are other cases where evidence is available for both variables, and correlations can be calculated to evaluate their actual relationship. One example is the popular use of the aluminium price index to make predictions on economic growth. An increase, if interpreted as an increase in demand, indicates that the industry expects more sales in the near future. This idea itself is often questioned and examined in detail within the particular research context [14], yet it is frequently discussed and therefore a useful means. Altogether, we noticed that there are some common practices we might be able to support. As there exist relations among these variables, we could describe them in a model aiming at giving hints to researchers which data could be useful to them depending on the context they are working in.

Although these ideas need more clarification in detail, the efforts to evaluate such relations might be much less than expected. As in the aluminium price example, both related variables can be measured eventually, taking Gross Domestic Product (GDP) for economic growth, their relation can be analyzed with respect to their correlation. Since all data in the model is highly formalized, machines can run carefully designed algorithms on it.

One idea is to simply count frequencies of cooccurrences of combined indicators as they in fact occur in research data sets. In other words, it is possible to run market basket analyses (e.g. applying association rules learners) on the data sets that have been published. One possible application could then be implemented as data suggestions in retrieval scenarios, as is similarly practiced at online shops where one can see what might be interesting on the basis of what other customers have bought.

The other way to achieve such kind of support is on the basis of in-depth analysis. Finding relations among indicators is not a miracle, it is based on clear mathematical properties. On the basis of correlations, analyzing coefficients, variances and other features, relationships could be determined automatically and proposed for useful proxies. This might be unrealistic in difficult cases, where adjustments, shifts or complex formulas would be necessary to detect relations. But it appears feasible for the simple cases in which a few default routines could be used to test for relatedness. To evaluate the usefulness and applicability of these ideas is part of our research agenda.

3.5 Data Publication and Repeatability

When considering data publication strategies and infrasture, the main objectives must be citability, transparency and reusability of published research data. For citation, the discussion on how to apply persistent identifiers[12] still raises questions on implementation details. We argue, however, that most of these questions can only be answered as soon as the overall data model and publication strategies are clear. From a technical perspective, it is crucial that such identifiers can be linked within the data model. A decision on a particular metadata vocabulary is therefore is not subject of discussion in this paper.

To investigate the matter of transparency, we first need to clarify what resources are required for it. Today's level of transparency is mostly limited to a ten-pages description of a research conducted, in which the authors explain their work, a couple of presentation slides and an overall report once the project has finished [9]. This material certainly lays open the claims, the methods applied and the outcome, which is what most people are interested in. Nevertheless, science demands that research results can be replicated, and the researchers we have interviewed agree in the statement that doing so is a very challenging task in their particular scientific domain. This is due to the fact that many steps in preparing the data[13] are very detailed, and the analysis steps involve loads of parameters which are difficult to include in a research paper, as it floods with information not interesting to a reader when the goal is to understand the work - not to evaluate or replicate it. To overcome this problem, we suggest to include such highly detailed information within the data publication.

First and foremost we must clarify what resources are necessary to make empirical research entirely transparent. Therefore, we distinguish three items: a precise reference to the data used, a detailed description of the data cleaning steps, and a detailed description of the analysis steps.

The reference part can be dealt with sufficiently utilizing a simple extension of the semantic model presented in Section 3, whereas we introduce a subclass of `qb:DataSet` to describe a researcher's data publication, containing self-generated data as well as pointers to the exact data items and slices included herein. The exact modelling is rather technical and involves the introduction of one or more classes and predicates only, and we may describe it more thoroughly in later works. Practically speaking, the researcher's data must eventually be captured and annotated, for which there exists prior work to build on [13] [1]. However, it is to mention that such referencing technique requires stable URIs for single data items (observations) for every data set, which is yet to be considered more carefully. There may be other techniques avoiding this issue which therefore may lack in exact referencing, which is also to be examined more deeply

[12] e.g. on Digital Object Identifiers (DOI) or the handle system

[13] also referred to as *data cleaning*, which among others involves plausibility tests, dealing with missing values and all kinds of adjustments

in later works. Altogether we estimate that these efforts pay out, considering the difficulties one has to face in other approaches (see Section 2).

For the detailed descriptions, a textual documentation in plain text formats is maybe the first one may think of. But this demands quite some documentation effort, may lack in clarity and fails in authenticity, as it may not match with what actually has been done, and computers cannot be used for verification due to the inherent implications when using free text. Although quite a daring approach, there is quite good reason for emphasizing on machine-processible descriptions. One is that such procedures are mostly carried out by machines anyway, under the use of scripted algorithms. Basically, most of its steps may probably befound in the command history of the statistics tool of choice, even though quite unorganized as reseachers switch back and forth between different methods and data versions to see which one suits best. Eventually, it is to say that empirical researchers are requested to provide clear mathematical descriptions of the rules and methodologies applied anyway, as the manual manipulation of data by subjective considerations is considered unprofessional among themselves.

And experiences from information technology domain, be it in academic or industry sectors, tools for dependency management and automated build processes are well-established and its use pays off in almost every situations [17]. In particular, we have tools like ANT[14] or Maven[15] in mind to handle language dependend scripts[16] and carry out processes, ranging from download[17] and composition of the referenced research data to the execution of data cleaning scripts as well as the generation of analysis results and maybe even the plotting of graphs and other figures. To make this possible, scripts and environments must be clearly specified. As this all is quite a visionary idea, we also realize that some of the analyses conducted by researchers cannot be performed in one off-the-shelf statistics environment. For some of the research involves quite sophisticated programming and is to be carried out on high performance clusters. Nevertheless, this approach might work for a high number of cases, and considering its potential in providing a highest degree of transparency, we consider it worth following.

As a consequence, a researcher would need to think in terms of build processes and develop scripts in modular fashion. These scripts must then also find their place within the data publication model. There are early thoughts on how such implementation may look like, however, this as well will be subject of later works. This approach eventually not only serves the purpose of transparency. At this stage, we may claim that it also addresses the aspect of *repeatability* as well, which may even be done by machines with the click of a button. Elsevier[18] coined the term *Executable Paper* which intuitively describes this idea quite well. Altogether, the approach again needs evaluation with respect to practicability. From both, theoretical and technical perspective, we are convinced of its feasibility. We call this entire model as a whole, including precise references and build scripts, the *Generating Model* for empirical research results.

[14] `http://ant.apache.org/`

[15] `http://maven.apache.org/`

[16] Plugins would be required for commonly used environments like R, Stata, SPSS and others.

[17] Download is possible for openly available data. In other cases this approach might be enhanced for license handling or is to be examined further for the particular cases of data protection.

[18] `http://elsevier.com/`

4. APPLICATIONS

With the given data representation model, we investigate further ideas for applications in the following.

4.1 Data Access and Retrieval

Common data retrieval use cases imply a query interface for describing the object one is looking for and a presentation scheme for matching results. In the repository context, further means for accessing resources while taking care of access privileges are required as well, and other features like browsing and filtering by categories or attributes are very common, as implemented in typical repository software. These all are fundamental features for a solid system, however, they are not subject of discussion in this section. In the following we investigate further ideas on the grounds of our model and argumentation above.

Let us choose a simplistic research scenario in which we examine the relationship between unemployment and economic growth in the European Union. For a start, we should be allowed to communicate our plans to the system in order to bring it in a position in which it is able to assist us in achieving our goal. Thus, the interface needs one section for entering the two concepts `unemployment` and `economic growth` and another one for describing the population our research is supposed to be based on. Since in many situations, a data table containing statistically significant numbers is the basis for empirical analysis, we chose an empty spreadsheet for the visual design of our interface. Typically, the columns stand for indicators, variables or concepts that are to be compared, while the rows represent the individual data records listed by some sort of dimension attribute like reference periods in time series or countries in transnational comparisons. So we start typing the names of the concepts needed for our research in the column headers while the system assists us per autocompletion suggestions, indicating familarity with the term and availability of related data sets. In our case, both concepts are part of the STW[19] and there exist some related data sets in our demo repository[20]. Finished typing, each column then shows one of the concepts and the number of related and available data sets in its header. A click on the data set counter reveals a listing, and the names and metadata of each available data set can be inspected[21]. However, as the philosophy behind this interface design is to narrow down step by step towards what we need, we first take a look at what the system has to offer at this stage.

To inspect the union over all properties attributed to the datasets of a column, we select the resepctive header. In

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

[20] Made of a few imported data sets from EuroStat and StatsWales

[21] The system already allows for selecting one of these data sets, but we will get to this part later in the scenario

other words, when we select a column header, the panel on the left lists down all properties that can be found for the respective data. For the `unemployment` column, the union of properties consists of geo, time and gender references, which can be inspected in detail by selecting the respective property of interest. It then opens (following the accordeon scheme) and displays the union over the available property values, which in case of `gender` would be `Female` , `Male` and `Total` . Selecting one of the values defines a constraint, meaning that all those data sets that do not provide for that kind will be removed from the listing behind the column, and the counter drops accordingly. This way of successively specifying constraints is what we refer to as narrow-down or drill-down principle. The user may start from a very generic description of what she has in mind and communicate her requirements one after the other, according to her priorities.

As we have claimed earlier, it is important to keep the population within the data consistent[22]. Therefore, the interface allows for selecting multiple columns, offering a way to inspect all properties the columns have in common, namely the intersection over the selected columns' properties. In this situation, all properties and values displayed are shared by some data sets among the columns, and constraining will be applied individually to each of the columns. Moving on in our scenario, we set `Total` for the `gender` constraint in the `unemployment` colum[23] and specify `EU27` for the `geo` constraint for all columns. As we have no further requirements for our data, and the counters have reached small numbers, we now examine the available data sets one by one, and make a selection for each column eventually.

To finally populate the table with actual values, we have to specify the dimension by which the rows shall be generated. This is done by selecting the upper left corner of the table, which colors the entire table, indicating that the values available for dimension must be provided by all of the columns. We decide to choose all available reference periods for our table[24] and the table floods with data, showing date values on the very left, concepts and dataset names in the column headers and the data values in the table cells (Figure 1). Although not yet implemented, the user would then have the option to get a graphical visualization of the results, obtain a URL for this configuration or simply download the table in several formats. Note that the order of specifying constraints, dimension values or choosing data sets does not matter. One may also remove or change any of the constraints or specification. The system consitently fills the table according to its actual state of knowledge and specification.

We conclude the description of our explorative prototype with a few remarks on open issues and further ideas. One is that it has been evaluated briefly by a few economists and social scientists only and a more thorough evaluation is yet to be conducted. It is also to mention that the current version supports macro data only, and the provision of micro data is crucial to make it a powerful tool. This, however,

[22]if not purposely intended otherwise

[23]There is no distinction in gender for economic growth, thus is not available when all columns are selected

[24]This simply is a multiple selection of values, the implementation follows the convention.

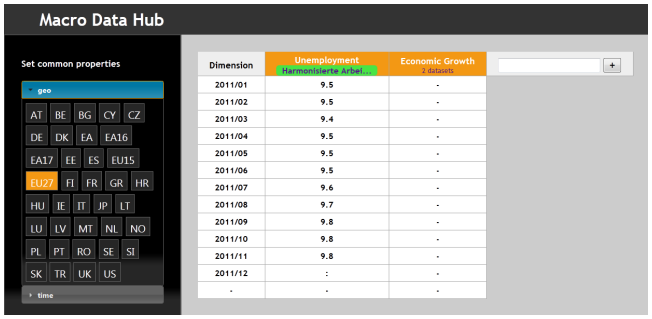

Figure 1: The screenshot displays the scenario at a stage where the data set for the first column is already chosen, while the one for the second column is not. The former therefore displays the name of the data set in a green box, while the latter still shows the counter of available data sets. The two columns are currently selected, as indicated by the orange background color in the column headers, and `EU27` is being selected for the geo property (or constraint) for both of them. Some time references have been chosen for the dimension column, and since the gender property was already specified for the unemployment column, the system is able to determine the data to be displayed in the cells, which is the unemployment rate in the EU at the given points in time according to EuroStat.

requires additional functionality for data cleaning and aggregation, which might complicate the interface and demand a lot of work in the details. Some other difficulties regarding the interface pertain rather small design decisions for dealing with conflicting constraints or user communication in cases where only one constraint value is possible. Another idea we might venture into is an import feature for common models for immeasurable variables that are composed of a set of aggregated indicators.

4.2 Distance Measure

Picking up again the idea of user assistance in finding related data and proxies (Section 3.4), some additional helpful means are required to assess suitability. In addition, besides helping find compatible datasets in data composition use cases, we reverse the thinking and see if we can provide means for evaluating the soundness of composition for a given dataset published with respect to the design described in Section 3.5. As we have already motivated and decided on the data representation, we now want to explore what can be immediately implemented from the current degree of formality in knowledge representation. The discussion motivates a distance metric, which we attempt to formulate subsequently.

To determine whether two data sets describe the same or a similar population, the metric should take into account usual concepts as they occur for example in code lists. Since population is encoded implicitly by means of single properties, there may be some that match and some that do not match when comparing two data sets. For example, assume that in our example above (Section 4.1, unemployment data is based on `EU15` , and the data reflecting economic growth

refers to `EU27` . This mismatch suggests a discrepancy and could be valuable information when composing, re-using or evaluating. In this case, `EU15` and `EU27` are not entirely unrelated, in fact, one includes the other and may hence be considered matching to some extent. Provided this information is found within the linked data cloud, a heuristic algorithm could be applied to produce a more gradual score for the soundness of composition. Another example may include time series data that matches perfectly in terms of the given population. Yet, while both sources provide annual data, one is produced in winter and the other in summer time, and thus may result in a mismatch if both data sets have not been seasonally adjusted.

Population in our case is specified for each single data item in terms of typed relations (RDF properties) to particular values. In the data set depicted in Figure 2, the highlighted

	2004-6		2005-7		2006-8	
	Male	Female	Male	Female	Male	Female
Newport	76.7	80.7	77.1	80.9	77.0	81.5
Cardiff	78.7	83.3	78.6	83.7	78.7	83.4
Monmouthshire	76.6	81.3	76.5	81.5	76.6	81.7
Merthyr Tydfil	75.5	79.1	75.5	79.4	74.9	79.6

Figure 2: Example data set as also used in the RDF Data Cube Vocabulary (QB) draft.

data items is linked per geo-predicate to a resource that represents the place `Cardiff` , it further links to `2005-7` per time-period predicate and also links to `Female` to specify the gender of its population. As the purpose of this discussion is to convey the idea, we keep it simple and build the metric upon the linked resources only, excluding the respective type of predicate on the assumption that the value ranges of these predicates are disjoint and hence produce no clashes, as they are semantically different. Accordingly, every population would be described as a set of URIs, and thus, the population of our example item would be described as { `Cardiff` , `Female` , `2005-7` }.

At present, the requirements for our metric are minimal and include only the typical mathematical ones (non-negativity, identity of indiscernibles, symmetry, triangle inequality). Therefore, Hamming distance, Euclidean distance and the like should be sufficient for our purpose. More sophisticated metrics in the context of semantic models, that include subtyping and other expressions of relatedness, have been discussed thoroughly in prior works and may be evaluated as we proceed with the research [16]. However, the usefulness of such metric is yet to be evaluated and was introduced only to stimulate the debate.

For automatically finding proxies, besides applying learning algorithms like association rules that are based on existing compositions, we can also try to approach it from mathematical considerations. We may therefore assume the following:

- if the population of two datasets is same
- and if the numbers correlate sufficiently
- it is likely that one serves as a suitable proxy for the other.

Given that we are able to test for equality of population[25] for two given datasets, we may suggest them for proxy relation if the mathematical features are strong enough[26]. As we have seen in the Aluminium Price Index example, a proxy relation sometimes involves a shift in time (as for predictions), which should be considered when pursuing this research idea. furthermore, it may be reasonable to include datasets of *similar* populations when searching for proxy relationships in order to equip this search space with a continuous scale, which again motivates a distance measure.

A metric for calculating the semantic distance for populations of empirical datasets has been proposed in order to start a discussion. A deeper examination of its usefulness will be conducted as we proceed with our research. If the metric can be regarded as meaningful, we are then enabled to further examine established data mining and knowledge discovery techniques for finding relevant datasets, which might be an area worth investigating.

5. SUMMARY AND OUTLOOK

Addressing the objective of an organized research data infrastructure, one of our goals was to give insight to the particular challenges in the domain of empirical research. We have illustrated how researchers work with such data, explained its features, and analyzed its practical and legal aspects, which gave clear reason for the data modelling requirements we have formulated. An investigation of existing data representation techniques for empirical data lead us to using the Resource Description Framework (RDF) and the Data Cube Vocabulary (QB) specifically. Based on the fundamental design decisions, we explored further how citation, transparency and re-use of empirical data can be enabled sufficiently and elaborated on a Generating Model. To start a discussion on innovative data access services, we introduced our explorative prototype and proposed a distance metric to allow for recommender applications and user assistance in composing and evaluating data sets on the basis of populations.

Proceeding in our research activities, we are planning to further develop our ideas within the Science 2.0 network[27] that is currently being established by the Leibniz Association[28] in cooperation with affiliated institutes from all over Germany. Among other topics such as scientific communication, societal challenges, the network aims to build a broad community with a strong and interdisciplinary focus on research data management, so that eventually the applicability of our approach will be evaluated from diverse scientific perspectives.

[25] An exact comparison is usually not possible, as for example it may not be clear whether the population refers to citizens or to the chicken in the agricultural sector, just to mention an example.
[26] A simple test could be implemented using thresholds for co-variance and the like.
[27] `http://science20.zbw.eu/`
[28] `http://www.leibniz-gemeinschaft.de/`

6. REFERENCES

[1] M. V. Assem, H. Rijgersberg, M. Wigham, and J. Top. Converting and annotating quantitative data tables. *Proceedings of the 9th International Semantic Web Conferene ISWC 2010*, pages 1–16, 2010.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[3] R. Cyganiak, S. Field, A. Gregory, W. Halb, and J. Tennison. Semantic statistics: Bringing together sdmx and scovo. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[4] S. Dallmeier-Tiessen. Positionspapier Forschungsdaten, 2009.

[5] DFG. Wissenschaftliche literatur- und informationsversorgungssysteme. schwerpunkte der fÃűrderung bis 2015.

[6] M. Feijen. What researchers want - a literature study of researchers' requirements with respect to storage and access to research data, February 2011.

[7] H. Gonzalez, A. Y. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, and W. Shen. Google fusion tables: data management, integration and collaboration in the cloud. In J. M. Hellerstein, S. Chaudhuri, and M. Rosenblum, editors, *SoCC*, pages 175–180. ACM, 2010.

[8] T. Gottron, C. Hachenberg, A. Harth, and B. Zapilko. Towards a semantic data library for the social sciences. In *SDA'11: Proceedings of the International Workshop on Semantic DigitalArchives*, 2011. in Preparation.

[9] J. Gray. Jim Gray on eScience: A Transformed Scientific Method, Jan. 2007.

[10] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. Scovo: Using statistics on the web of data. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. HyvÃűnen, R. Mizoguchi, E. Oren, M. Sabou, and E. P. B. Simperl, editors, *ESWC*, volume 5554 of *Lecture Notes in Computer Science*, pages 708–722. Springer, 2009.

[11] T. Hey, S. Tansley, and K. M. Tolle. Jim gray on escience: a transformed scientific method. In T. Hey, S. Tansley, and K. M. Tolle, editors, *The Fourth Paradigm*. Microsoft Research, 2009.

[12] J. V. D. Hoeven. Insight into digital preservation of research output in Europe Case studies report PARSE . Insight. *PARSEInsight Deliverable D33 Case Studies Report*, 2010.

[13] S. Lynn and D. Embley. Semantically conceptualizing and annotating tables. pages 345–359. 2008.

[14] K. Matthies. Commodity prices remain high. *Intereconomics*, 42(2):109–112, 2007.

[15] M. R. Montgomery, M. Gragnolati, K. A. Burke, and E. Paredes. Measuring living standards with proxy variables. *Demography*, 37(2):155–174, 2000.

[16] R. Oldakowski, C. Bizer, F. U. Berlin, I. Produktion, W. Or, and D. Berlin. Semmf : A framework for calculating semantic similarity of objects represented as rdf graphs pid-58. *World Wide Web Internet And Web Information Systems*, pages 2–4, 2005.

[17] G. Popp. *Konfigurationsmanagement mit Subversion, Ant und Maven: Ein Praxishandbuch für Software-Architekten und Entwickler*. dpunkt, Heidelberg, 2006.

[18] O. Siegert. Speicherung und publikation von forschungsdaten. der beitrag der deutschen zentralbibliothek für wirtschaftswissenschaften. Working Paper Series of the German Council for Social and Economic Data 158, German Council for Social and Economic Data (RatSWD), 2010.

[19] J. Starr and A. Gastl. iscitedby: A metadata scheme for datacite. *D-Lib Magazine*, 17(1/2), January/February 2011.