Building Linked Data For Both Humans and Machines*

Wolfgang Halb†
Institute of Information Systems & Information Management
Graz, Austria

Yves Raimond‡
Centre for Digital Music
London, UK

Michael Hausenblas§
Institute of Information Systems & Information Management
Graz, Austria

ABSTRACT

In this paper we describe our experience with building the riese dataset, an interlinked, RDF-based version of the Eurostat data, containing statistical data about the European Union. The riese dataset (`http://riese.joanneum.at`), aims at serving roughly 3 billion RDF triples, along with millions of high-quality interlinks. Our contribution is twofold: Firstly, we suggest using RDFa as the main deployment mechanism, hence serving both humans and machines to effectively and efficiently explore and use the dataset. Secondly, we introduce a new way of enriching the dataset with high-quality links: the User Contributed Interlinking, a Wiki-style way of adding semantic links to data pages.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Linked data, Semantic Web, XHTML+RDFa, User Contributed Interlinking

1. MOTIVATION

The goal of the "RDFising and Interlinking the Eurostat Data Set Effort" (riese)[1] is to offer a Semantic Web version of the public accessible data provided by the Eurostat data source. riese has been initiated as part of the W3C SWEO Linking Open Data (LOD) project and aims at being useful for both humans and machines.

Existing linked datasets such as [3] are slanted towards machines as the consumer. Although there are exceptions to this machine-first approach (cf. [13]), we strongly believe that satisfying both humans and machines from a single source is a necessary path to follow.

We subscribe to the view that every LOD dataset can be understood as a Semantic Web application. Every Semantic Web application in turn is a Web application in the sense that it should support a certain task for a human user. Without offering a state-of-the-art Web user interface, potential end-users are scared away. Hence a Semantic Web application needs to have a nice outfit, as well.

Further, the interlinking algorithms found in current LOD datasets are largely based on templates. This means that a huge number of interlinks can be generated, however, the quality of these links in terms of their respective 'semantic strength' is somewhat limited. It is well known that humans are good at associations, so we basically propose in the following to let humans do the hard part of the interlinking.

The paper is structured as follows: Section 2 discusses related efforts, then in section 3 we introduce the Eurostat dataset and state the requirements. In section 4 we describe the riese, and discuss in section 5 the current implemented version of the system. Finally, in section 6, we conclude on the current results and outline future steps.

2. RELATED WORK

Statistical data on the (Semantic) Web. Looking at related work reveals that there is actual demand for new solutions to disseminate statistical data using semantic technologies. As reported by Assini [2] the European Union funded a research and development project called NESSTAR in 1998, with the aim of bringing the advantages of the Web to the world of statistical data dissemination. Another project that is entirely situated on the Semantic Web is the U.S. Census data [20] where 1 billion RDF triples containing statistical information about the United States were published in 2007.

An earlier attempt to publish Eurostat is known from the FU Berlin[2], using a very small subset of country and region statistics. Stuckenschmidt [19] has reported on translating and modelling the European fishery statistics in ontologies. [10] recently pointed out issues with translating the Swiss

LDOW2008, April 22, 2008, Beijing, China.

†JOANNEUM RESEARCH Forschungsges. mbH, email: `wolfgang.halb@joanneum.at`

‡Queen Mary, University of London, email: `yves.raimond@elec.qmul.ac.uk`

§JOANNEUM RESEARCH Forschungsges. mbH, email: `michael.hausenblas@joanneum.at`

[1] `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/EuroStat`

[2] `http://www4.wiwiss.fu-berlin.de/eurostat/`

statistics to an RDF basis. A somehow related approach is the Rswub[3], a package for handling statistical data, based on RDF and capable of handling ontologies.

RDFa. As RDFa [1] is turning into a W3C Last Call document at the time of writing of this paper, the penetration is expected to dramatically increase in the next couple of months. Although not yet a standard, there exist a number of smaller-sized deployed datasets, such as those listed at `http://rdfa.info/rdfa-in-the-wild/`. It has been reported that for example Joost plans to offer RDFa-enriched content[4] and we have recently proposed to use RDFa as a base for multimedia metadata deployment [11]. However, to the best of our knowledge there exists no other linked-data set deployed in RDFa.

3. REQUIREMENTS AND ISSUES

3.1 The Eurostat data

This section provides a short description of the Eurostat data, which served as the primary input for the riese dataset.

Eurostat provides detailed statistics for the entire European Union as well as additional statistics for major non-European countries. The Eurostat data is arranged along the following themes:

- General and regional statistics
- Economy and finance
- Population and social conditions
- Industry, trade and services
- Agriculture and fisheries
- External trade
- Transport
- Environment and energy
- Science and technology

Three main data sources are being provided by Eurostat for public download[5], namely (i) the statistical data itself, (ii) a table of content, and (iii) dictionaries.

The statistical data is provided as dump download of approximately 4,000 single tab-separated values (TSV) documents, having a total size of approximately 5GByte, and containing some 350 million data values. This data is updated twice a day. Only limited semantic exploitable information is contained in these TSV documents, hence it is inevitable to use other available information sources.

A table of content (TOC) provides a hierarchical overview of the datasets—organised in so called themes—allowing to identify the structure and content of a dataset.

[3] `http://www.biostat.harvard.edu/~carey/hbsfin.html`

[4] `http://rdfa.info/2007/08/23/joost-using-rdfa-on-website/`

[5] `http://europa.eu/estatref/download/everybody/`

Dictionaries are especially valuable as they contain all information for resolving the nearly 100,000 data codes used in the statistical data. These data codes refer to dimensions such as time, location, currency, etc. The data codes also contain an implicit hierarchy, which can be used for further classification. However, various schemas have been used requiring individual processing for extracting classifying features. For example, in order to refer to locations, the Nomenclature of Territorial Units for Statistics (NUTS)[6] is in use. This basically allows to extract information about the structure of administrative divisions of countries. For each of the dictionaries a different terminology is used.

Most of the data is represented in time series with varying granularity, ranging from annual to daily data. Each single data item can be identified using the corresponding dataset and various dimensions as the following example illustrates: The population of the European Union can be seen as one single data item valued at 497,198,740 (contained in the dataset 'Total population'), having as time-dimension the year 2008, as indicator-dimension 'Population on 1. January', and as geo-dimension the 'European Union (27 countries)'. Additionally the data is flagged as provisional and Eurostat estimate.

3.2 Requirements

In a first phase we have analysed the Eurostat data. We have identified the implicit semantics present in the TOC and the dictionaries and gathered a number of issues. Firstly, the Eurostat data set is highly heterogeneous; the data sources formats vary (TSV, HTML) and are not machine-processable per se. Another issue is the modelling of temporal data, more specific how to represent time intervals. Further, the schemas in the dictionaries form a multidimensional space that somehow has to be linearised in order to be represented in a URI format. We have also identified data provenance (and trust) issues, which are currently only handled on a global level.

Based on the analyses given above we state the following requirements for a linked dataset that is designed to serve both humans and machines:

- The system must serve both humans and machines in an adequate way by applying the don't-repeat-yourself (DRY)[7] principle;
- To allow both humans and machines to reveal more information, the follow-your-nose[8] principle must be applied.
- To be a useful (real-world) Semantic Web application, the system must be able to scale to the size of the Web;

Additionally we want to point out that we aim at providing high-quality interlinking. Hence, the sheer template-driven generation of global interlinks is certainly not sufficient.

[6] `http://ec.europa.eu/comm/eurostat/ramon/nuts/`

[7] `http://skimstone.x-port.net/node/272`

[8] `http://www.inkdroid.org/journal/2008/01/04/following-your-nose-to-the-web-of-data/`

4. LINKED DATA FOR HUMANS AND MACHINES

In order to demonstrate how to address the issues raised earlier in this paper, we have implemented the riese dataset (`http://riese.joanneum.at`) as a Semantic Web application. This section describes how the mapping—from the available, relational data into RDF form—has been done, explains the interlinking mechanisms applied, and finally introduces the riese system architecture.

4.1 Data, Schemas and Mapping

This section explains the schemas utilised in riese and discusses the mapping to RDF.

The data used in riese is a snapshot of the data available for bulk-download taken on 9 Jan 2008. Depending on the type of data, three formats are used by Eurostat: HTML or plain text for the TOC, and TSV for the dictionary files and the actual data tables.

In Fig. 1 the riese core schema is depicted. Currently the riese core schema is modelled using RDF-Schema [4] rather than OWL [14] based and comprises three main classes: `riese:Dataset`, `riese:Item` and `riese:Dimension`. A dataset is the logical container of either more sub-datasets (related via `skos:narrower`) or data items. An item represents one single data value (like 497,198,740 for the population of the European Union) with all accompanying metadata about the containing dataset and the dimensions used. A dimension semantically describes the value of a data item in terms of, e.g. time, location, unit, etc. In listing 1 an exemplary snippet of an item is shown.

```
data:eb040_infl_2006_at a :Item ;
  dc:title "Inflation rate Austria 2006" ;
  rdf:value "1.7" ;
  :dimension dim:geo_at ;
  :dimension dim:time_2006 ;
  :dataset data:eb040 .
```

Listing 1: An single data item.

Additionally, the following schemas are used or have been extended:

- Dublin Core (DC) Elements [7] and Terms [6]
- Geonames [9]
- Simple Knowledge Organisation Systems (SKOS) [18]
- Description of a Project (DOAP) [8]
- the event ontology [16]

We decided to model a flat schema for the following reasons:

- Queries can be constructed with very little a-priori knowledge about the structure of the dataset;
- Additional Eurostat datasets can easily be added without changing the schema (and are instantaneously integrated in the hierarchy, hence available to all users regardless of the access method);
- Dimensions can be added without any changes to the schema;
- Finally, it is possible to formulate very flexible queries.

Other approaches, such as the U.S. Census data [20] use a more complex schema, where for example a new property for every possible description is introduced. This yields properties such as `population15YearsAndOverWithIncomeIn1999`, which do not offer any additional semantic information.

Querying data using these properties can get very cumbersome, as the user would have to know about the exact terms beforehand. We believe that our flat approach, where every value can be identified by the corresponding dataset and dimensions, enables fairly flexible queries.

```
SELECT *
WHERE
{ ?item riese:dimension dim:geo_at.
  ?item riese:dataset ?dataset.
  ?dataset dc:title ?ds_title
  FILTER regex(?ds_title, "food",i)}
```

Listing 2: A query in riese.

The example in listing 2 demonstrates this. All items for Austria are returned that belong to a dataset with 'food' in the description[9].

4.2 Interlinking

Leaving the mapping of the Eurostat data into RDF apart, it is equally important to apply the follow-your-nose principle, hence creating interlinks to other datatsets. For creating interlinks in riese we have basically used the following approach:

1. Restrict the source dataset to possible candidates for interlinking to the target dataset;
2. For each qualifying item in the source dataset look up the label or another identifying feature in the target dataset;
3. Restrict the results by appropriate classifications or identifiers;
4. Create the interlink.

For example the interlinking between country descriptions in riese and Geonames is done using the ISO-3166 alpha2 country codes (AT) instead of the label (Austria) assuring that exactly the same resource is addressed in both datasets.

[9] with default namespace `http://riese.joanneum.at/schema/core#`

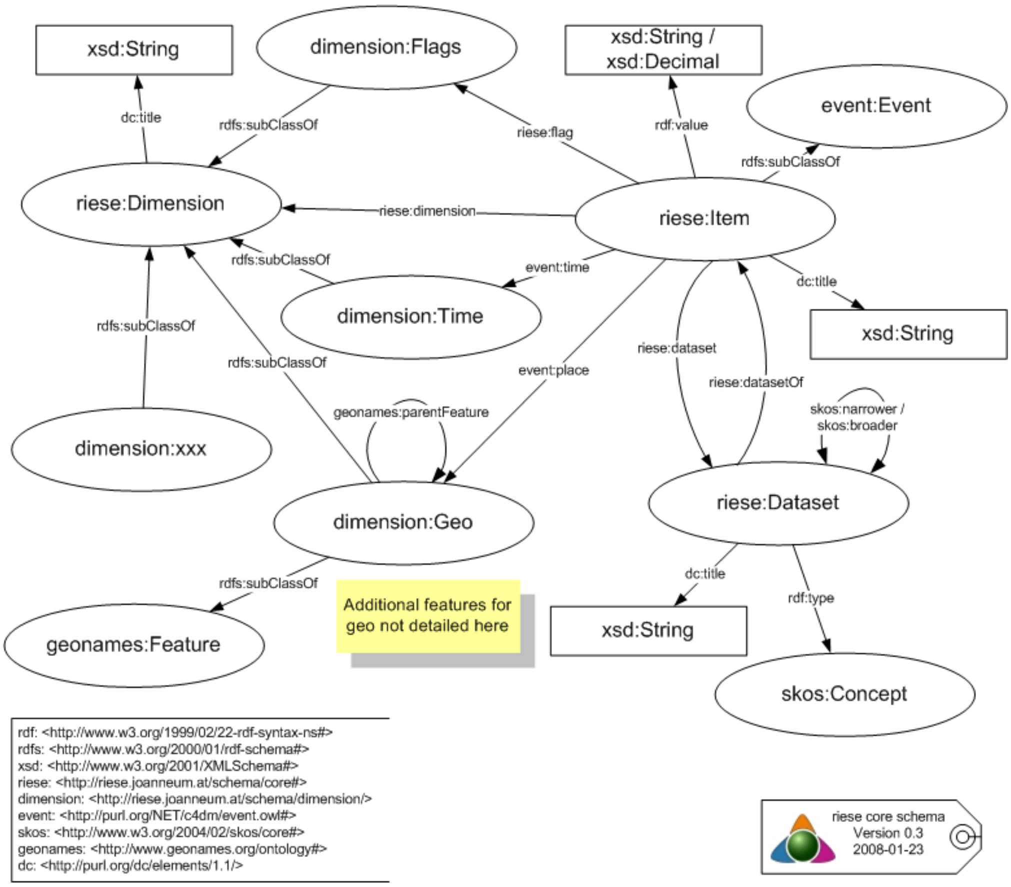

Figure 1: The core schema.

Note that using ISO-3166 codes for identifying country descriptions in different datasets was already used by Voss [21] and others.

In the practical implementation this means that first of all the source dataset is restricted to only geographical features. According to the nomenclature used it is also possible to identify country descriptions in the source dataset. Then the Geonames search Webservice (i.e. the target dataset) is queried using the standardized codes. The result from the target dataset is then further restricted to return only countries, i.e. entries having a specified Geonames feature code (A.ADM1). Finally all the matches are being interlinked by inserting a new triple into the source dataset which relates the resources using owl:sameAs. In this case it is possible to create exact matching high-quality interlinks.

Further candidates for interlinking the Eurostat data are Geonames (more geographical features), DBpedia, CIA Factbook and Wikicompany. By introducing these interlinks users of riese will not only benefit from a larger interlinked dataspace but especially for the geographic features also by being able to produce even more flexible and powerful queries.

As already mentioned above, the pure pattern-based approach is believed to be not sufficient for high-quality interlinks. This is why we additionally allow users to add their own links, a new feature called 'User Contributed Interlinking' (UCI). The idea behind is applying the WikiWiki approach to LOD: Users can add semantic links to other datasets on their own. Currently three different types are supported: `rdfs:seeAlso`, `owl:sameAs` and `foaf:topic` (cf. also [5]).

4.3 System Architecture

Based on the lessons learned from [12] we have developed the riese Web application. It comprises:

1. An (offline) module, being responsible for converting the Eurostat data into an RDF representation and creating the global, pattern-based interlinks (RDFising & Interlinking), and a
2. Web server including a scripting environment that fills predefined templates with the values from the (static) RDF/XML representation in order to generate an RDFa representation of the themes and the data tables.

The Fig. 2 depicts the riese system architecture and shows as well the interfaces with the environment (in and out ports).

The riese Web application supports the following tasks:

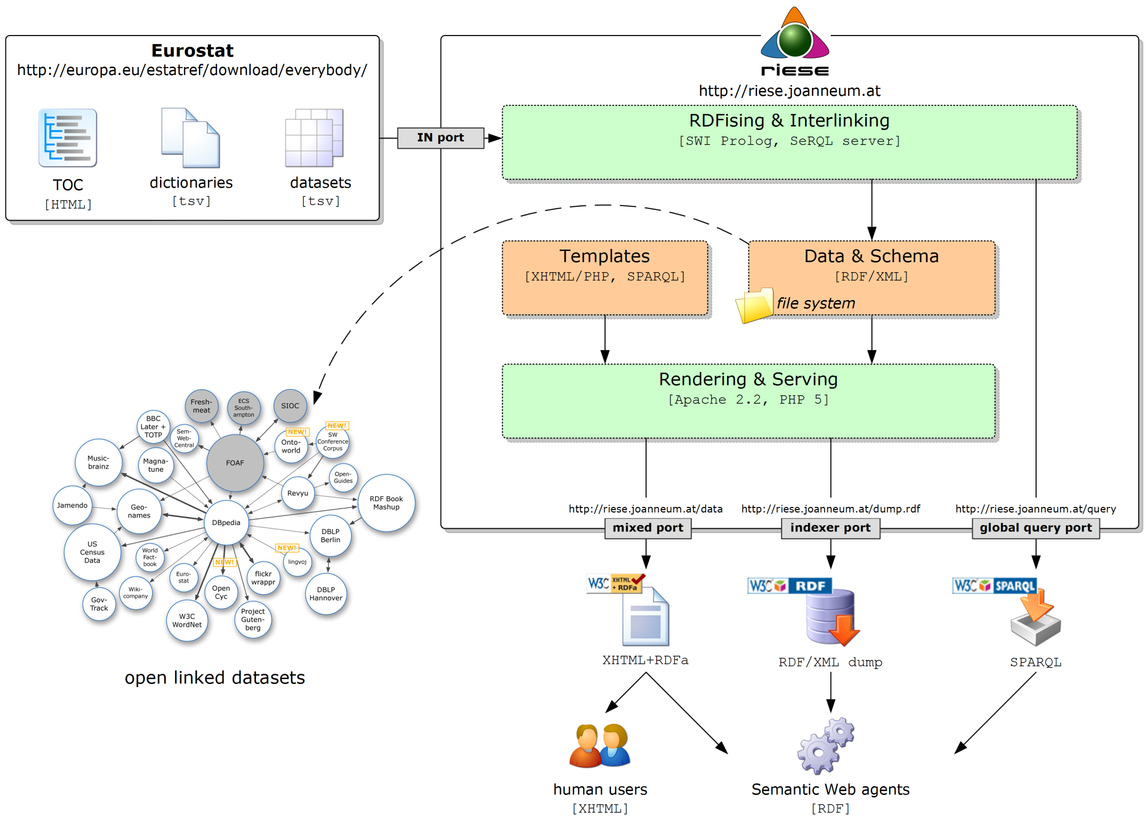

Figure 2: The system architecture of riese.

- Human users: Users can navigate the dataset provided in XHTML+RDFa;
- Semantic Web agents:
  - single item query—XHTML+RDFa per page allows the exploration of the dataset and the query of a single data item (FYN);
  - global query—To allow an efficient query of the entire dataset, a SPARQL-endpoint is provided;
  - indexer: to allow semantic search engines (indexer) an effective processing, the entire dataset is offered as a dump and an according description using the semantic crawler sitemap extension protocol[10] is offered.

For creating the RDF representation from the original Eurostat files, SWI-Prolog scripts are used. The SWI-Prolog Semantic Web Library provides an infrastructure for reading, querying and storing semantic web documents. Additionally the Prolog-2-RDF (p2r) modules[11] and individually defined mappings are used for translating the input data to RDF.

The resulting RDF can be accessed via a SPARQL endpoint and it is further possible to consume a dump of the entire data. We have created one large dump containing all triples, and also store the triples according to their URI directly into the file system in RDF/XML.

[10] `http://sw.deri.org/2007/07/sitemapextension/`
[11] `http://moustaki.org/p2r/`

The latter approach is currently used for 'Rendering & Serving' where the PHP scripts looks up the files in the file system and renders a RDFa representation. Beside the data that originates from Eurostat (the official statistical data), the UCI module stores the user-contributed triples in a separate document. This physical separation is mainly due to being able to replace parts of the data without too much additional effort.

5. USING RIESE

In the following we show how riese can both satisfy the human user, as well as the machine (Semantic Web agents). Please note that the alpha version of the riese system is available at `http://riese.joanneum.at/`.

Both human and machine users would presumably start at the top-level page in order to get an overview of the available data. In Fig. 3 the hierarchical rendering of a selected Eurostat theme (the 'Economy' theme) is depicted. A machine accessing the same page would have another view, namely focusing on the embedded RDF, exemplary shown in example 3.

Note that although both humans and machines access the same resource, different parts are relevant. This is made possible through the deployment in XHTML+RDFa. The

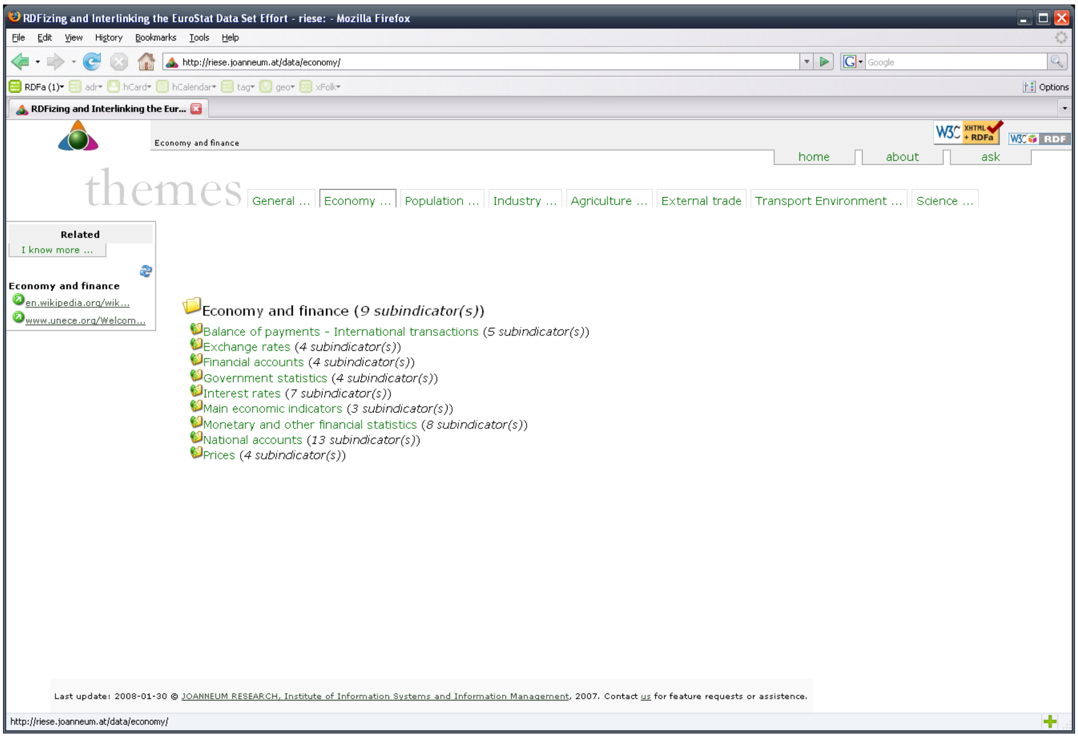

Figure 3: The Eurostat theme 'Economy' viewed by a human user.

```
<body
 about="http://riese.joanneum.at/
        data/economy"
 instanceof="riese:Dataset">
...
<div id="main-ind">
 ...
 <a href="http://riese.joanneum.at/
          data/bop"
    rel="skos:narrower">
 Balance of payments -
 International transactions
 </a>
</div>
```

Listing 3: The Eurostat theme 'Economy' viewed by a machine.

browser will render a nice GUI, the machine gets what it deserves: triples.

Further, a single table may be explored; this is depicted in Fig. 4.

However, till now the user was passively consuming the information. But riese offers more: Users can provide their own links using the UCI (cf. Fig. 5).

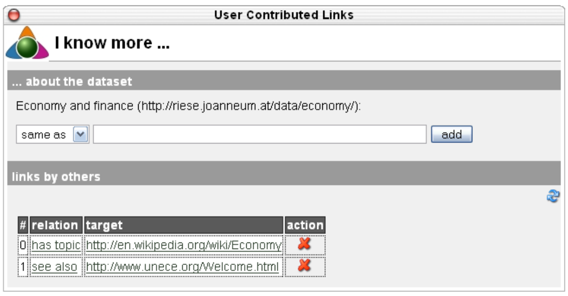

Figure 5: The UCI module—users can provide own links.

The UCI module enables the user to add (and remove for that matter) additional links to a certain data page. As the user must specify the type (cf. the drop-down box in Fig. 5) it is ensured that only valid triples are introduced to the system—the subject of the RDF statement is always the page where the 'Related' box is on; the predicate is de-

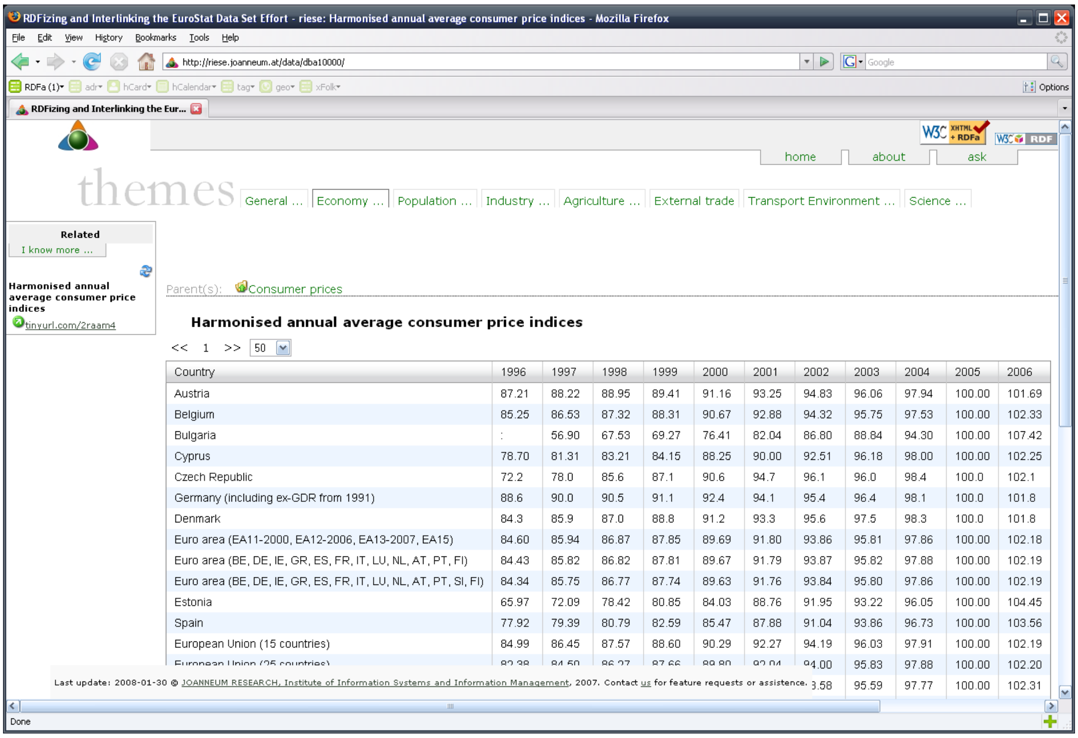

Figure 4: A single data table in XHTML+RDFa.

termined through the type selection. The object (named target in our context) is the only variable we are not able to control. However, we rely on the community effect, i.e. we expect that 'wrong' links will be removed. A REST-based interface for adding UCI-triples automatically is available as well. Regarding the acceptance of the UCI, i.e. enabling users to contribute semantically typed links, we refer to the success story of Wikipedia [17] and strive for considerable community involvement. In riese , we therefore try to implement many of the success factors of Wikipedia, such as openness or ease of editing. However, UCI may need to be applied to other datasets with more 'appealing' data—compared to statistical one—in order to properly evaluate its uptake.

6. CONCLUSION

In this paper we have presented the riese dataset containing statistical data from Eurostat. We have shown how to RDFise and interlink this data, hence making it possible to expose it onto the Semantic Web. The benefits of supplying data for both humans and machines have been explicated and a WikiWiki approach for adding user contributed interlinks has been introduced.

We have also identified some issues and bottlenecks when deploying datasets of such enormous size. Generating a static file-structure with small RDF files requires quite a lot of time. This is due to our current way of storing the data items in the file system. Because in riese several hundred millions of folders and files have to be created, the bottleneck is somehow obvious. Moreover, when accessing datasets (tables) containing thousands of items (cells) in individual files this yields thousands of file access operations for simply parsing them. Regarding the file system we came across another limitation: reserved names on the MS Windows operating systems (as it turned out, it is not possible to create files or folders named 'con', 'aux', etc. [15]).

When modelling the representation of time related to a certain statistical information we encountered some challenges as the raw data from Eurostat is sometimes ambiguous and can only be resolved by analysing the corresponding document. For example the statement `time\2007` can stand for the value over a period of time (e.g. entire year) or at the end of the reporting period (e.g. 31 Dec). In our future work we will focus on resolving these issues.

The future work roughly comprises a thorough analysis of the current bottlenecks, as well as gathering feedback from end-users of the system. We are planning to use a solution based on an triple-store (such as SESAME or Virtuoso) allowing us to generate triples at a faster pace—currently it would take us several weeks to RDFise the entire Eurostat data set. Using a dedicated store will likely improve the performance serving the data to both human and machine users.

Finally, as Eurostat updates their data twice a day, we aim at updating the data on riese continuously. One of the issues to be solved in this respect is how to deprecate the data when updating the items. From a UI point-of-view we also want to address navigational issues (using maps and timelines[12]) to further enhance the user experience.

[12] http://simile.mit.edu/timeline/

7. ACKNOWLEDGMENTS

The research leading to this paper was carried out in the "Understanding Advertising" (UAd) project[13], funded by the Austrian FIT-IT Programme, and was partially supported by the European Commission under contract FP6-027026-K-SPACE.

The authors would like to thank the Linking Open Data community and the RDFa folks. Additionally we would like to credit all the people that made available the following magnificent technologies: SWI-Prolog, Apache, PHP, RAP - Rdf API and YUI. We wish to thank Danny Ayers for his early comments on modelling issues, Giovanni Tummarello for feeding sindice (allowing advanced queries), and Jan Wielemaker for his superb SWI-Prolog support.

8. REFERENCES

[1] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton. RDFa in XHTML: Syntax and Processing. W3C Working Draft 18 October 2007, W3C Semantic Web Deployment Working Group, 2007.

[2] P. Assini. NESSTAR: A Semantic Web Application for Statistical Data and Metadata. In *International Workshop Real World RDF and Semantic Web Applications, 11th International World Wide Web Conference (WWW2002)*, 2002.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 722–735, 2007.

[4] D. Brickley and R. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, RDF Core Working Group, 2004.

[5] D. Brickley and L. Miller. FOAF Vocabulary Specification. `http://xmlns.com/foaf/0.1/`, 2004.

[6] Dublin Core Metadata Initiative. DCMI Metadata Terms. `http://dublincore.org/documents/dcmi-terms/`, 2008.

[7] Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1. `http://dublincore.org/documents/dces/`, 2008.

[8] E. Dumbill. Description of a Project (DOAP) vocabulary. `http://usefulinc.com/ns/doap`, 2005.

[9] Geonames. Geonames Ontology. `http://www.geonames.org/ontology/`, 2007.

[10] A. Grossenbacher. Semantic Web: Basics, RDF, DC and the description of a statistical site. `http://tinyurl.com/2d5gta`, 2007.

[11] M. Hausenblas, W. Bailer, and H. Mayer. Deploying Multimedia Metadata in Cultural Heritage on the Semantic Web. In *First International Workshop on Cultural Heritage on the Semantic Web, collocated with the 6th International Semantic Web Conference (ISWC07)*, Busan, South Korea, 2007.

[12] M. Hausenblas, W. Slany, and D. Ayers. A Performance and Scalability Metric for Virtual RDF Graphs. In *3rd Workshop on Scripting for the Semantic Web (SFSW07)*, Innsbruck, Austria, 2007.

[13] T. Heath and E. Motta. Revyu.com: a Reviewing and Rating Site for the Web of Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 895–902, 2007.

[14] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. W3C Recommendation, OWL Working Group, 2004.

[15] Microsoft. Naming a File. `http://msdn2.microsoft.com/en-us/library/aa365247.aspx`, 2008.

[16] Y. Raimond and S. Abdallah. The Event Ontology. `http://motools.sourceforge.net/event/event.html`, 2007.

[17] L. Sanger. The Early History of Nupedia and Wikipedia: A Memoir. In C. DiBona, M. Stone, and D. Cooper, editors, *Open Sources 2.0: The Continuing Evolution*. O'Reilly, 2005.

[18] Semantic Web Deployment Working Group. SKOS Simple Knowledge Organization System Reference. `http://www.w3.org/TR/swbp-skos-core-spec/`, 2008.

[19] H. Stuckenschmidt and F. van Harmelen. *Information Sharing on the Semantic Web*. Springer, 2005.

[20] J. Tauberer. The 2000 U.S. Census: 1 Billion RDF Triples. `http://www.rdfabout.com/demo/census/`, 2007.

[21] J. Voss. Encoding changing country codes in RDF with ISO 3166 and SKOS. In *International Conference on Metadata and Semantics Research (MTSR07)*, 2007.

[13] `http://www.sembase.at/index.php/UAd`