Semantic Web 1 (2012) 1–5
IOS Press

The Digital Agenda Scoreboard: An Statistical Anatomy of Europe's way into the Information Age

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Michael Martin [a,*], Bert van Nuffelen [b], Stefano Abruzzini [c] and Sören Auer [a]
[a] *University of Leipzig, Institute of Computer Science , AKSW, Postbox 100920, D-04009 Leipzig, Germany*
E-mail: {lastname}@informatik.uni-leipzig.de
[b] *TenForce, Haachtsesteenweg 378, 1910 Kampenhout, Belgium*
E-mail: bert.van.nuffelen@tenforce.com
[c] *European Commission, Directorate-General for Information society and Media, B-1049, Brussels, Belgium*
E-mail: Stefano.Abruzzini@ec.europa.eu

Abstract. Evidence-based policy is policy informed by rigorously established objective evidence. An important aspect of evidence-based policy is the use of scientifically rigorous studies to identify programs and practices capable of improving policy relevant outcomes. Statistics represent a crucial means to determine whether progress is made towards policy targets. In May 2010, the European Commission adopted the *Digital Agenda for Europe*, a strategy to take advantage of the potential offered by the rapid progress of digital technologies. The Digital Agenda contains commitments to undertake a number of specific policy actions intended to stimulate a circle of investment in and usage of digital technologies. It identifies 13 key performance targets. In order to chart the progress of both the announced policy actions and the key performance targets a scoreboard is published, thus allowing the monitoring and benchmarking of the main developments of information society in European countries. In addition to these human-readable browsing, visualization and exploration methods, machine-readable access facilitating re-usage and interlinking of the underlying data is provided by means of RDF and Linked Open Data. We sketch the transformation process from raw data up to rich, interlinked RDF, describe its publishing and the lessons learned.

Keywords: European Commission, Information Society, Scoreboard, Internet, Data Cube, RDF, Linked Open Data

1. Introduction

Evidence-based policy is policy informed by rigorously established objective evidence. An important aspect of evidence-based policy is the use of scientifically rigorous studies to identify programs and practices capable of improving policy relevant outcomes. Statistics represent a crucial means to determine whether progress is made towards policy targets. In May 2010, the European Commission adopted the *Digital Agenda for Europe*[1] (DAE), a strategy to take advantage of the potential offered by the rapid progress of digital technologies. The DAE is part of the overall *Europe2020 strategy* for smart, sustainable and inclusive growth.

The Digital Agenda contains commitments to undertake 101 specific policy actions (78 actions to be

*Corresponding author (martin@informatik.uni-leipzig.de).

[1] http://ec.europa.eu/information_society/digital-agenda/

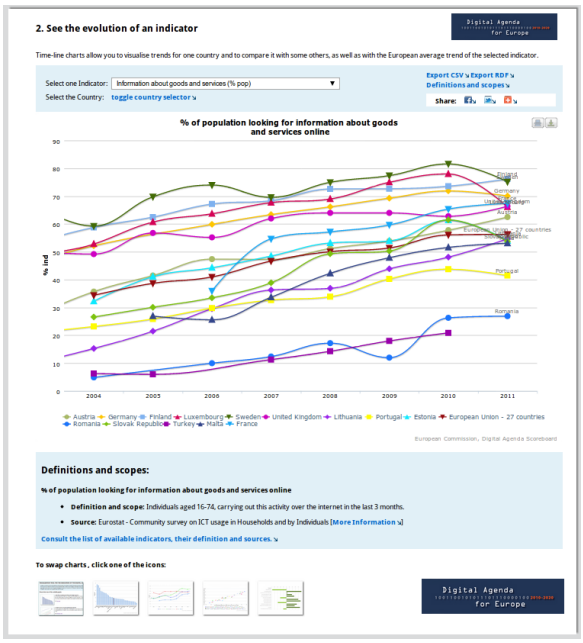


Fig. 1. Screenshot of the Digital Agenda Scoreboard tool.

taken by the Commission, including 31 legal proposals, and 23 actions proposed to the Member States) intended to stimulate a virtuous circle of investment in and usage of digital technologies. It identifies 13 key performance targets to show whether Europe is making progress in this area. In order to chart the progress of both the announced policy actions and the key performance targets, the DAE calls for the publication of an annual scoreboard, supported by a large set of statistical indicators allowing monitoring and benchmarking of the main developments of information society in European countries.

As an outcome, the visualization tool of the *Digital Agenda Scoreboard* (DAS) was published in June 2011, which is exemplary depicted in Figure 1. This application was developed for interested citizens and professionals (e.g. journalists) providing them with the possibility to browse statistical data with suitable visualization and interaction features. In addition to these human-readable access methods, machine-readable access facilitating re-usage and interlinkability of the underlying data in a dereferencable way is provided by means of RDF and Linked Open Data.

In this article we sketch the transformation process from raw data up to the resulting RDF in Section 2. Descriptions about publishing the RDF representation are given in Section 3. A summary of related work and our lessons learned while dealing with statistical data is presented in Section 4 and 5.

2. Scoreboard Data Transformation

The indicator's statistical data is collected by the *European Commission services* by combining data from different sources (cf. Table 2). From this information a subset with the most important indicators is selected for the publication within the digital agenda scoreboard. The subset goes then through an additional quality and information extension process. Namely, the data is consolidated in a standard exchange structure which forms the input of the RDFization process. The core statistical information of each indicator is extended with contextual, provenance and metadata information. This includes in particular labels for each indicator (long and short ones), structural information (such as the group to which the indicator belongs) as well as comments and definitions on the concepts being used. These extra pieces of structured knowledge are necessary for creating dynamic views accessible to humans and for facilitating the reuse of the data by 3rd parties. Before entering into the RDFization process the whole data is checked for completeness.

In the sequel we sketch the transformation process flow from the given raw data representation up to the resulting RDF.

2.1. Analyzing the original dataset

The original data were encoded as spreadsheets containing the following four sheets: (a) a sheet describing the indicators (b) a sheet listing the groupings of indicators, (c) a sheet containing provenance information of the data and (d) a sheet containing the data / observation itself. In (a), an indicator is described through a set of methodological information about the source, the scope and the definition (such as variable, breakdown, unit, label etc.). An overview about indicator groups and the respective counts of indicators as well as the usage of them as part of observation descriptions are listed in Table 1. In sheet (c), information about the provenance (source) of data is given by a short and a long label as well as a link and its extraction date. Every indicator references to a specific source and furthermore every observation is specified by a specific indicator, which offer the possibility to identify the source of every observation as well. A summary of the sources and its respective count of indicators and observations is listed in Table 2. Sheet (d) contains the observation data, which are defined by a variable, the breakdown and the unit used to identify the respective indicator. The measurement units applying to the

Table 2
Sources of observations.

Id	Source	Indicators	Observations
S1	eGovernment Benchmarking Report, Study for the EC realized by Capgemini (2001-2010 data)	2	398
S2	Electronic communications market indicators for the Communications Committee (COCOM) - January reports.	18	1903
S3	Eurostat - Community survey on ICT usage and eCommerce in Enterprises	17	2918
S4	Eurostat - Community survey on ICT usage in Households and by Individuals	51	9248
S5	Broadband Coverage in Europe, Studies for the EC realized by IDATE (2005-2010 data) and by Point Topic (2011 data)	5	545
S6	EC database of ICT research projects under the EU's FP7 (all projects from the Cooperation program excepting the projects from Capacities program)	15	2635

Table 1
Scoreboard indicator groups.

Indicator group	Indicators	Observations
Telecom sector	3	168
Broadband	15	2258
Mobile	9	746
Internet usage	6	1258
Take up of internet services (% of internet users)	13	2384
Take up of internet services (% of population)	13	2414
eGovernment	7	1348
eCommerce	13	2790
eBusiness	7	712
ICT Skills	7	933
R&D abs	4	704
R&D rel	11	1931
Overall 12 Groups:	108	17646

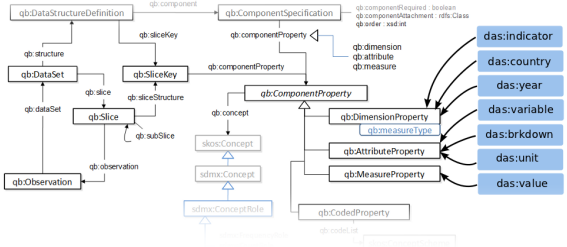
observations are encoded in a separate value column. In order to organize the data in spatial and temporal dimension every observation is tagged with a country code and the year of measurement (between 2001 and 2011).

2.2. The Scoreboard DataCube schema

The RDF representation of the DAS is based on the DataCube vocabulary [4], which defines terminological resources[2] for structuring and representing statistical information. Statistical information is represented as resources of type `qb:Observation` which are described by using component properties (c.f. Figure 2) such as the observation measure

[2]namespace: `http://purl.org/linked-data/cube#`

Fig. 2. DAS DataCube Schema (extended from [4]).



(property of type `qb:MeasureProperty`) as well as contextual information using properties of type `qb:DimensionProperty`.

Based on the analysis presented in Section 2.1, we integrated the dimension properties `das:country`, `das:indicator` and `das:year` enabling the definition of the DAS observation context. Furthermore, we integrated the attribute properties `das:brkdown`, `das:variable` and `das:unit` used to encode definitions of the indicator and as illustrated in Figure 2 the measure property `qb:value`, which is used to encode measure information of the observation. As an example, a DAS observation resource is illustrated in Listing 1, representing the measure (`das:value`) of individuals aged between 16 and 74, carrying out this activity over the internet in the last 3 months observed in year 2010 in country Iceland.

```
<http://data.lod2.eu/scoreboard/obs/Iceland/2010/                    1
    i_igovrt_IND_TOTAL_ind>
  a qb:Observation ;                                                 2
  das:value "0.5049785991"^^xsd:float ;                              3
  das:indicator ind:i_igovrt_IND_TOTAL_ind ;                         4
  das:year <http://data.lod2.eu/scoreboard/years/2010> ;             5
  das:country country:Iceland ;                                      6
  das:unit "% ind" .                                                 7
```

Listing 1: Example of a DAS observation

To create unique URIs for observation resources the item namespace[3] were combined with the variable, breakdown, unit, country and year information of the respective observation. The indicator relation references to one out of the 108 indicator resources. Every indicator resource is located in the indicator namespace[4] and is uniquely identifiable by its localname consist of the variable, breakdown and unit. As being illustrated in Listing 2, an indicator contain further information such as an `rdfs:label` and an `rdfs:comment` for textual information as well as `dcterms:source` and `dcterms:publisher` to encode provenance information.

```
1 <http://data.lod2.eu/scoreboard/indicators/
      i_igovrt_IND_TOTAL_ind>
2   rdfs:label "% of population sending filled forms to public
        authorities, over the internet, last 3 months" ;
3   rdfs:comment "Individuals aged 16-74, carrying out this
        activity over the internet in the last 3 months." ;
4   das:brkdown "IND_TOTAL" ;
5   das:unit "% ind" ;
6   das:variable "i_igovrt" ;
7   dcterms:source <http://epp.eurostat.ec.europa.eu/portal/
        page/portal/information_society/introduction> ;
8   dcterms:publisher "Eurostat-Community survey on ICT usage
        in Households and by Individuals".
```

Listing 2: Example of a DAS indicator

In order to complete the DataCube we added further structural resources, which are of type `qb:DataSet` and `qb:DataStructureDefinition` used to explicitly encode components of the dataset. For every component property (year, country, indicator, value) a definition resource of type `qb:ComponentSpecification` is integrated in the DataCube as well.

2.3. Transformation workflow

To generate charts on the basis of user selected filter in DAS, the underlying RDB is queried by using SQL. In addition to the HTML output DAS is offering an CSV and RDF representation of the same selected set of information which can be downloaded by users. To offer that functionality, DAS is generating the output only by using different templates which are rendering the data (observations, used component properties and referenced object resources such as indicators) within a loop. To export the complete RDF DataCube an administrative function is integrated into DAS offering a selection query without any filter. Using that function all observations, component properties, indicators, countries and years are rendered as RDF in turtle notation. As an outcome, the actual resulting RDF file contain 126981 triples.

To finalize that transformation workflow a resource of type `owl:ontology` is added to integrate meta data about the dataset itself such as the following set of information:

- a label using `rdfs:label`,
- a description using `rdfs:comment`,
- the last modification date using `dct:modified` (actual: 14. of May 2012),
- the license using `cc:license` (published under CC-BY-3.0)
- the project homepage using `doap:homepage`,
- publisher using `dct:publisher` as well as
- the contributor using `dct:contributor`.

If a new version of the raw data is delivered as spreadsheet, the DAS RDB is being updated and the described transformation workflow above will be restarted.

2.4. Scoreboard interlinks in the LOD cloud

Officially, the scoreboard is not directly linked to any other dataset in the LOD cloud. The main reason for the absence was the absence of RDF datasets which are approved by EUROSTAT. As the source of most data originates from EUROSTAT, it is natural to link the published RDF with RDF resources that are EUROSTAT controlled. This ensures coherency in the semantics of the data.

For instance, at the same level of the countries often the EU27 average is being used. The semantics of the average is not simply the average over all countries, but it is a weighted formula.

`http://eurostat.linked-statistics.org/`, a result from the LATC project [5], provides an RDF snapshot of EUROSTAT data. Using SILK [19] to create links between the scoreboard data and this snapshot is rather straightforward. The linking is based on exact matching of the `rdfs:label` for the countries and time dimension. The result of this linking operation is summarized in Table 3 and published as a separate dataset[6].

The link property used is `skos:RelatesTo`. It is not an identity relation as `owl:sameAs` because each

[3] `http://data.lod2.eu/scoreboard/items/`
[4] `http://data.lod2.eu/scoreboard/indicators/`
[5] `http://latc-project.eu`
[6] `http://data.lod2.eu/scoreboard/links/`

Table 3
Created links.

Linktype	Source	Target	Links
Countries	*Scoreboard*	*Eurostat*	
skos:RelatedTo	qb:Observation	skos:Concept	16325
owl:sameAs	das:Country	skos:Concept	27
Year	*Scoreboard*	*Eurostat*	
skos:RelatesTo	qb:Observation	skos:Concept	17102
owl:sameAs	das:Year	skos:Concept	11
Overall:			33465

statistical observation is a piece of information about a country at a given time. Note that due the special cases like EU27, there is no perfect match based on the labels for the countries. To relate the countries and the time dimension elements with the LOD cloud, the identity relation holds and hence `owl:sameAs` is used.

3. Dataset Publishing

The visualization application of the Digital Agenda Scoreboard[7] (DAS) is an PHP application obtaining the data from an relational database. On the one hand the raw CSV data was transformed into a relational scheme as described in Section 2.3. This application creates visualizations of the data using different chart types and various data selection options. In addition to the HTML/JSON representation of the data, users are able to obtain the selected data as RDF. On the other hand the complete dataset was transformed into RDF and loaded into the public RDF store[8] based on OntoWiki [14]. OntoWiki provides in addition to an exploration interface RDF export functionality, a SPARQL endpoint and a Linked data interface facilitating the dereferencability of stored RDF resources.

All DataCube resources which are requested by users of the DAS application are encoded equivalently to them published in the RDF store. This setup provide users alternatives to select subsets of the data according their technical skills and possibilities. After publishing the Scoreboard RDF DataCube as described above, we added an entry[9] to theDataHub.org metadata repository for Open Data.

[7] `http://ec.europa.eu/information_society/digital-agenda/scoreboard/`

[8] `http://data.lod2.eu/`

[9] `http://thedatahub.org/dataset/scoreboard`

4. Related Work

Related work can be roughly divided into other RDF triplification approaches, statistical data publishing and linked governmental data applications.

Triplification. Currently most of the work in the area of triplification focuses on generating RDF from relational database content. There is a wide range of approaches developed in this regard ranging from very simple scripts such as *Triplify* [5] over standalone solutions such as *D2R* [7] up to integrated tools such as *Virtuoso RDF Views* [12]. Under the auspices of the W3C, the *RDB2RDF working group* is currently standardizing the *R2RML* mapping language for the mapping and transformation of relational data to RDF. One of the few works in the area of transforming statistical data to RDF is [15], which explores the opposite direction to our approach, i.e., the transformation of statistical Linked Data for use in OLAP systems.

Statistical Data publishing. Statistical Data and Metadata eXchange (SDMX, [2]) is an initiative started in 2001 to foster standards for the exchange of statistical information. The SDMX sponsoring institutions are the Bank for International Settlements, the European Central Bank, Eurostat, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations Statistics Division and the World Bank. The SDMX message formats have two basic expressions, SDMX-ML (using XML syntax) and SDMX-EDI (using EDIFACT syntax and based on the GESMES/TS statistical message). Experiences and best practices regarding the publication of statistics on the Web in SDMX have been published by the United Nations [1] and the Organisation for Economic Co-operation and Development [3].

The representation of statistics in RDF started with SCOVO [13,8] and continued with the successor RDF Data Cube Vocabulary [4]. The Data Cube Vocabulary is closely aligned with SDMX [8]. Examples of statistics published as RDF adhering to the Data Cube vocabulary and visualized for human consumption include the EC's INFSO Digital Agenda Scoreboard[10] and the LOD2 Open Government Data stakeholder survey [17].

[10] `http://ec.europa.eu/information_society/digital-agenda/scoreboard/`

Linked Governmental Data. Several governments started to publish governmental data on the Web. Tim-Berners Lee discussed a set of Design Issues [6] on how to publish governmental information in a reusable way. One of the first Linked Data providers publishing governmental data is UK government with `http://data.gov.uk/`, hosting information about different governmental sectors of Great Britain including transport, legislation and finance [18]. A further provider of governmental data is `http://data.gov/` which is hosting information about the USA. Due to the fact that this information was not made available as Linked Data, external groups started to transform and publish the information according the Linked Data principles [9]. Recent research work also aims to facilitate government data ecosystems through specialized portals [10] and distributed dataset catalogs [11]. Another important issue, which is particularly tackled by this paper for the statistics domain, is enabling interoperability of government data catalogs [16].

5. Conclusion and Lessons Learned

The Digital Agenda Scoreboard is one of the first datasets publishing statistical information according the DataCube vocabulary [4] published by a non-research institute. In the past year, the RDF structure went through several revisions. The changes were driven by changes in the provided raw data or by needs of the visualization. The DataCube vocabulary turned out to be flexible enough to capture them. A future extension of the current dataset structure will be the support for dataslices, a notion available in the DataCube vocabulary.

As mentioned before interlinking the dataset is technically not so difficult. The major stumble stone is the absence of an official RDF formatted publication of the data where is it based upon. Having that available the usability of the scoreboard dataset is boosted.

This dataset is actively maintained and will grow over the coming years. Regularly new data will be uploaded for the existing indicators. Also more indicators will become available. In the future Digital Agenda will open up also new datasets covering other aspects of the digital society. To make this sustainable an improved publishing process will be setup so that the Digital Agenda becomes an even more reliable and up-to-date source of information about the European digital society.

References

[1] Guidelines for statistical metadata on the internet. Technical report, United Nations, Economic Commission for Europe (UNECE), 2000.

[2] Statistical data and metadata exchange (sdmx). Technical report, Standard No. ISO/TS 17369:2005, 2005.

[3] Management of statistical metadata at the oecd, 2006.

[4] The rdf data cube vocabulary. Technical report, 2010. http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html.

[5] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller. Triplify: Light-weight linked data publication from relational databases. In *WWW*. ACM, 2009.

[6] T. Berners-Lee. Putting Government Data online. W3C Design Issue, 2009. `http://www.w3.org/DesignIssues/GovData.html`.

[7] C. Bizer and R. Cyganiak. D2r server - publishing relational databases on the semantic web. Poster at ISWC, 2006.

[8] R. Cyganiak, S. Field, A. Gregory, W. Halb, and J. Tennison. Semantic statistics: Bringing together sdmx and scovo. In *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[9] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, D. L. McGuinness, and J. Hendler. Twc data-gov corpus: incrementally generating linked government data from data.gov. In *WWW*, pages 1383–1386. ACM, 2010.

[10] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, and J. Hendler. Twc logd: A portal for linked open government data ecosystems. *J. of Web Semantics*, 2011.

[11] J. S. Erickson, E. Rozell, Y. Shi, J. Zheng, L. Ding, and J. A. Hendler. Twc international open government dataset catalog. In *Proceedings of the 7th ICSS*, I-Semantics '11. ACM, 2011.

[12] O. Erling. Automated Generation of RDF Views over Relational Data Sources with Virtuoso, 2009.

[13] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. Scovo: Using statistics on the web of data. In *ESWC*, volume 5554 of *LNCS*. Springer, 2009.

[14] N. Heino, S. Dietzold, M. Martin, and S. Auer. Developing semantic web applications with the ontowiki framework. In T. Pellegrini, S. Auer, K. Tochtermann, and S. Schaffert, editors, *Networked Knowledge - Networked Media*, volume 221 of *Studies in Computational Intelligence*, pages 61–77. Springer, Berlin / Heidelberg, 2009.

[15] B. Kämpgen and A. Harth. Transforming statistical linked data for use in olap systems. In *I-SEMANTICS 2011*, 2011.

[16] F. Maali, R. Cyganiak, and V. Peristeras. Enabling interoperability of government data catalogues. In *Proc. of the 9th IFIP*, EGOV'10, 2010.

[17] M. Martin, M. Kaltenböck, H. Nagy, and S. Auer. The open government data stakeholder survey. In *OKCon*. OKFN, 2011.

[18] J. Sheridan and J. Tennison. Linking uk government data. In *WWW2010 Workshop on Linked Data on the Web (LDOW)*, 2010.

[19] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *ISWC*, 2009.