



Eau de Web

R4 Final Report – September 2015

Support services for the Digital Agenda Scoreboard Website SMART 2012/0103

Description of the work carried out, of the results, relevance and future challenges concerning data visualisation, data linking and interactive functionalities

15 September 2015

Table of contents

1. Summary	5
2. Outcomes of the project	5
3. Evolution of the subordinated website	5
3.1. Documentation and deliverables	6
3.2. Activities and contractual tasks	6
4. Roadmap and future challenges.....	7

1. Summary

This document represents the final report delivered under the contract 30-CE-0508604/00-20, for the project “Support services for the Digital Agenda Scoreboard website - SMART 2012/0103”. This document includes the comments made during the final handover meeting (M36).

The aim of this report is to summarize the work carried out and the results obtained for each contractual task. It contains an executive summary of the most relevant aspects and indications to other documents where the mentioned topics are addressed in much more detail.

2. Outcomes of the project

During this three-year contract, the following outcomes have been achieved:

- A interactive dissemination tool for digital data and indicators in the Information Society area, in particular related to the Digital Agenda targets;
- A tool that can convert statistical data available in various non-semantic formats to linked statistical data;
- A tool that can be easily used to explore data sets expressed in the RDF Data Cube Vocabulary¹, create attractive charts and facilitate the reuse of data and metadata in open and machine-readable formats;
- A deeper understanding of linked open data and underlying semantic technologies by the Commission Services
- Opportunities for establishing contacts, sharing knowledge and alignment with other organizations, projects and tools related to the statistical linked data domain

The tools have been developed in line with the recommendations for publication of linked open data; the data upload and visualisation tools work directly with the semantic format (stored in a non-relational data model). While this architectural design has many advantages – interoperability being the most prominent – there are also a few technical challenges when it comes to data maintenance, use of SPARQL by non-expert users, stability and performance of the visualisation tool. An implicit objective of the DAD tool suite is to hide the underlying complexity and address these technical challenges through visual interfaces and process automation, providing step-by-step guidance for the administrative tasks.

Until the date, DG Connect has published so far three datasets related to benchmarking the performance of European countries:

- the Digital Agenda Key Indicators (137 indicators, 518k observations, 4.2 M triples)
- the Digital Economy and Society Index (53 indicators, 6k observations, 51k triples)
- the Lead Indicators for DG Connect policy priorities (31 indicators, 35k observations, 280k triples)

The datasets are modelled using the RDF Data Cube vocabulary (a W3C standard focused on the publication of multi-dimensional statistical data on the web).

3. Evolution of the subordinated website

The contract consisted of two main phases:

- A. Takeover and relocation of the previous website and data portal (a custom PHP application featuring four predefined visualisation scenarios loading data from a relational databases, OntoWiki - a semantic content authoring tool and Virtuoso semantic database). This website was hosted on a new server under the domain <http://digital-agenda-data.eu> until June 2013;

¹ <http://www.w3.org/TR/vocab-data-cube>

- B. Since June 2013, a new suite of applications was released, based on Plone (visualisation tool, replacing the PHP application), Content Registry (data maintenance tool, replacing OntoWiki) and Virtuoso (data storage).

The new server is hosted in a fully featured hosting environment, located in Bucharest, Romania (GTS Telecom's Data Centre). Technical details about the hardware and networking environment are included in deliverable R1 – Inception Report. Regular backup, system monitoring and maintenance was implemented. There are currently three environments: production, test and development, all hosted on the same physical server.

The new system was developed using open-source libraries, the entire source code being publicly available online on GitHub (details in chapter 1.5 of the deployment manual). The major software components are:

- The visualisation tool, developed in JavaScript and Python on top of Plone CMS, initially reusing some of the EEA Daviz² modules and later converted to a standalone package
- The data maintenance tool, an adaptation of the EEA Semantic Data Service³
- Virtuoso Open-Source Edition⁴, a free triple store maintained by OpenLink Software

In terms of web statistics, the website has 2000-3000 unique visitors each month, with occasional peaks up to 6000 unique visitors (e.g. February and March 2015, when the Digital Economy and Society Index has been published).

3.1. Documentation and deliverables

The documentation in PDF format is uploaded online, in the documentation page (<http://digital-agenda-data.eu/documentation>). The most relevant documents available in this page are:

- The latest technical report (latest update: July 2015): <http://digital-agenda-data.eu/documentation/technical-report-m30>, describing the architecture of the website, its interactive functionalities, the administrative features, usage statistics, the data cube reference model and the data linking experiments
- Deployment manual (<http://digital-agenda-data.eu/documentation/deployment-manual>) detailing the complete process that would allow a third party to replicate the website in a new environment
- Dataset upload manual: <http://digital-agenda-data.eu/documentation/dataset-upload-manual>, explaining the data maintenance procedure

Based on the deployment manual, under the Open Data Support contract⁵, another project funded by DG CONNECT, a set of installation scripts has been developed, published on GitHub at <https://github.com/tenforce/vagrant-digital-agenda-scoreboard>. The setup is fully automated and uses Vagrant to easily run a virtual machine containing a complete clone of <http://digital-agenda-data.eu/> and its subordinate tools.

3.2. Activities and contractual tasks

According to the specific objectives and tasks from the tender specifications, the new system was continuously improved with new features (as detailed and planned during the interim meetings), such as:

- Task 2: redesign the data upload/maintenance process, allowing conversion and import of data from various sources (Excel spreadsheets, MS Access files published by Eurostat, SDMX web services or other RDF sources) and further maintenance of data and metadata using Content Registry
- Task 3: further development of the visualisation tool, including various new types of visualisations, using parameters configured by administrators
- Task 4: conducting experiments concerned with data linking with other SPARQL endpoints (detailed in Annex 2 of the latest Technical Report)
- integration with new data sources (e.g. Eurostat) and data publishers (e.g. EU Open Data Portal)

² <http://eea.github.io/docs/eea.daviz>, <https://plone.org/products/eea.daviz>

³ <http://semantic.eea.europa.eu>

⁴ <http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main>, <https://github.com/openlink/virtuoso-opensource>

⁵ <https://joinup.ec.europa.eu/community/ods/home>, <http://data.opendatasupport.eu/>

- Task 5: interactive functionalities (collection of user comments, embedding charts in external websites)

Along with hosting, regular system maintenance, bug fixing and user support, a chronological overview of the new features developed is listed below:

- 2013-06 Initial launch of the new website, followed by a stabilization period
- 2013-10 Documentation page (<http://digital-agenda-data.eu/documentation>)
- 2013-12 New chart type: Compare the evolution of two indicators; added Croatia & EU28
- 2014-01 Chart embedding feature
- 2014-02 Activation of HTTPS and SSL certificates
- 2014-04 Support for user comments
- 2014-04 User interface for maintenance of combinations used in country profile
- 2014-04 Create a new dataset by copying existing visualisations from another dataset
- 2014-05 Activation of stacked bar charts
- 2014-07 Configuration of a new web analytics tool (Piwik)
- 2014-09 Compare the evolution of two indicators from different datasets
- 2014-12 Automatic generation of metadata for Open Data Portal
- 2015-01 Interactive composite charts for DESI
- 2015-03 SDMX extraction and conversion to RDF DataCube
- 2015-05 Knowledge sharing workshop on how to use the DCV for dissemination of statistical data
- 2015-06 Redesign and improvement of the visualisation website on small devices
- 2015-07 Radar chart for alternative country profiles
- 2015-08 Handover preparation activities

All the above are described in more detail in the latest technical report and in the dataset upload manual.

4. Roadmap and future challenges

During the current contract, a number of possible improvements to be addressed in a future contract have been identified:

- Maintain the conformity of the reference data model to new versions of the Data Cube Vocabulary and other relevant vocabularies like SKOS, DCAT, STAT DCAT-AP, etc. Since the initial development of the DAD reference model, the RDF DCV standard was revised (new properties, recommendations and integrity rules have been added, such as qb:HierarchicalCodeList), other work in the same domain has been made, new code lists and metadata schemes might be more appropriate (e.g. RAMON – Eurostat's Metadata Server or vocabularies uploaded by the Publications Office of the EU in open-data.europa.eu).
- Revise the conceptual data model, the indicator and breakdown dimensions and the dataset-indicator relation. In the current DAD, a dataset may cover several indicators or even hundreds. However, it has been recognized that it would be more intuitive and less burdening for the users to obtain DAD datasets on a per-indicator basis (especially when discovering DAD datasets via Open Data Portal). Such an approach would also be more in line with the same approach taken by EUROSTAT.
- Add a search functionality, in case a large number of datasets is published. The search should also address data inside each dataset (e.g. search in name and definition of each indicator)
- Ensure interoperability with future developments of EUROSTAT dissemination tools, in particular new API for publishing datasets and metadata in SDMX format
- Improve the support for generic data structure definitions and remote endpoints, in order to provide better interoperability with other data publishers (for instance, to compare the evolution of two indicators for the same country – one from the DAD triple store and the second from an external publisher). Some technical limitations and possible improvements are listed in Annex 2 of the latest Technical Report

- Improve the linking between indicators subject and thematic domains, using well-established semantic vocabularies, for instance, using DCAT:Theme and dct:subject (as recommended by DCV⁶) to create links between datasets and EuroVoc or other concept schemes.
- Share the findings and possible improvements related to the Data Cube Vocabulary with the RDF DCV working group
- Further share knowledge and expertise with other initiatives/projects, by providing technical support in the modelling and validation of data cubes, in order to improve the governance and the adoption of best practices related to modelling semantic statistical data
- Identify possible synergy with other tools in the statistical linked data ecosystem
- Address various performance, stability, SEO and usability improvements (related to regular data maintenance operations)
- Upgrade the system components (most importantly: CentOS operating system, Virtuoso, Java, Tomcat and Plone) to latest versions
- Add new chart types and new parameters to the chart configurator:
 - ✓ improve the support for multi-dimensional and multi-dataset charts, such as the comparison of two indicators from different datasets, add support for charts using federated queries across different SPARQL endpoints,
 - ✓ add support for scatter/bubble/timeline charts with different dimensions (left vs right side),
 - ✓ add new possibilities of selecting and composing chart titles and labels, e.g. when two vertical axis are enabled,
 - ✓ add new configurable parameters (colours, ranking criteria, data rounding, hyperlinks),
 - ✓ manually specify “special” dimensions such as time, unit measure and indicator, when they cannot be identified automatically from the dataset structure.
- Address the existing technical debt, bugs and remaining issues (listed in the latest Technical Report)

⁶ <http://www.w3.org/TR/vocab-data-cube/#metadata-categorization>