Distr.
GENERAL

WP.19
2 May 2012

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2012)
(Washington, DC, 21-23 May 2012)

Topic (iv): Collaboration

Eurostat data as open data: experience with Google and with the open data community

Invited Paper

Prepared by Chris Laevaert, Eurostat, Luxembourg

I. Introduction

1. EUROSTAT, the Statistical Office of the European Union, compiles statistical data that are, for the most part, collected by member states and it adds value by providing statistics at European level that enable comparisons between countries and regions and by disseminating these data, free of charge, in consolidated format, via publications and on-line data bases.

2. Since 2004, free access to and re-use of data is a cornerstone of Eurostat's dissemination policy. By promoting the widest possible use of Eurostat data, it helps to establish official European statistics as the preferred source of data on European societies and economies, and it enables European citizens and businesses to make use of the data which they have provided and paid for.

3. As a general principle, and with few exceptions, Eurostat statistics can be downloaded from the Eurostat website and they can be re-used for any purpose, including commercial purposes, so long as Eurostat is mentioned as the source of the data. No charge is made for data and organisations which are re-using data are not required to sign any licence agreement.

4. Through the website, Eurostat makes it possible both to view and extract data using the visualisation tools developed by Eurostat; The Data Explorer which focuses on displaying the data in a user customisable way and the Table, Graph and Maps tool which offers additional graphing and maps functionality.

5. However, users may also download complete raw data files in several formats using the bulk download facility. This is the preferred way of fetching data for users who regularly download big chunks of data to store it in their own databases or who want to re-use the data with the tools of their choice.

6. The combination of freely accessible data and lack of restrictions on re-use of data means that Eurostat is already in line with the principles of the open data movement and with the European Commission's initiatives to open up access to public sector information (PSI) in the EU, for which a Directive , adopted in 2003, is currently under review.

II. The power of open data

A. Cooperation with Google

7. Cooperation with Google started in 2009. Google considered Eurostat as a main reference data source, and was extremely keen on the fact that Eurostat offers its data free of charge and that it can be downloaded in bulk. To underline this, Google clearly mentioned Eurostat as the reference in the Government Summit 2.0 in Washington DC in September 2009.
http://www.youtube.com/watch?v=Bf-inv6IccE

8. After the acquisition of Gapminder's Trendalyzer in 2008, Google started working on creating a new service that makes lots of data instantly available for intuitive, visual exploration.

9. Google started to pick up U.S. data first. As an example, typing in [unemployment rate] or [population] followed by a U.S. state in google.com, you will see the most recent estimates in the following way:

10. Once a user clicks the link or the graph, he will go to an interactive chart that lets him add and remove data for different geographical areas. This search feature will pave the way for public data to take a more central role in informed public conversations.

11. The tool is targeted to statistically inexperienced users.

12. Google expanded this service internationally and therefore contacted Eurostat to explore the possibility of putting Eurostat data behind this feature because the Eurostat data are harmonised across Member States and also freely available through the bulk download facility.

13. In a first step, based on a list of most common search terms used in the Google search engine, we identified together 11 Eurostat datasets which matched one or more of these commonly used search terms. We then worked together to get the data for three datasets curated for the first release of the Public Data Explorer. The HICP, the minimum wages and the monthly unemployment in the EU. The tool was officially launched by Google in March 2010, in Google Labs.
See http://www.google.com/publicdata/home

14. In a second step, Eurostat worked with Google to integrate and make some datasets directly available via Google search. It concerned unemployment rates, government debt, minimum wage and broadband penetration. We provided all the required meta information on the datasets. In doing so, Google translated table titles, definitions, footnotes, labels in 34 different languages. Google also made changes to its search algorithm to ensure that appropriate searches led directly to these datasets.

15. This search feature called "OneBox" was released with 4 Eurostat datasets in October 2010. This implies that after a search for e.g. [minimum wage Belgium] on google.com, the first search result is a graph as shown below. Eurostat will be mentioned as "Source" and from within the Public Data Explorer there are two links back to the Eurostat website. The "More information…" within the visualisation tool deep-links back to the relevant dedicated section on the Eurostat website, and another one points to a page which re-groups all relevant meta-information about the dataset.

16. To see this in action, here are some examples of searches:
- Arbeitslosenrate Deutschland
- Smic France
- Deuda publica españa

17. The traffic impact to the Eurostat website from Google Public Data Explorer is important. The number of extractions on these datasets doubled or tripled after integration in the Google search.

18. In the meantime a European policy blog was launched by Google explaining the reasons and the public data explorer.
See http://googlepolicyeurope.blogspot.com/2010/03/statistics-for-changing-europe-google.html

19. A couple of weeks after the opening of the Public Data Explorer it was possible to see concrete examples of its usage by different bloggers with the European data.

20. As of March 2012, there are nine "data cubes" based on Eurostat data which can be accessed using the Google Public Data Explorer:

- Unemployment in Europe (monthly)
- Harmonized Index of Consumer Prices in Europe
- Minimum Wage in Europe
- Broadband penetration in Europe
- Government Debt in Europe
- Road transport in Europe
- Food supply chain monitor
- Eurostat, Tourism
  Demography

See http://www.google.com/publicdata/directory

Other data cubes on GDP and external trade are in preparation.

21. Statistical data from other European National and regional Statistical Institutes are available, in particular from the Italian National Institute of Statistics (Istat) and the Statistical Institute of Catalonia (Idescat)

22. Due to the importance of SDMX as a standard in the statistical domain Eurostat developed a converter tool from SDMX-ML to the Dataset Publishing Language (DSPL), the format used by the Public Data Explorer. See also
http://circa.europa.eu/Public/irc/dsis/stne/library?l=/x-dis/tools/sdmx_converter/converter_tool/2011&vm=detailed&sb=Title

23. With the phasing-out of Google Labs in September 2011, described by Google as an effort to "put more wood behind fewer arrows", the Public Data was elevated to full Google status and the eliminated the initial fear of observers that the project would not be sustained.

B. Cooperation with open data community and other data redistributors

(a) Eurostat Hackday

24. Within the open data movement, one regular type of event is the so-called hackday , where people interested in building innovative applications around publicly available data are encouraged to meet, brainstorm and build prototypes.

25. The Open Knowledge Foundation and the Digital Enterprise Research Institute (DERI) at the National University of Ireland, Galway organised a Eurostat Hackday in December 2010.

26. Eurostat was not involved in the decision to set up the Hackday, nor in its organisation. The Hackday was an initiative which tended to confirm that people see Eurostat as one of the most important and accessible data sources; Eurostat cooperated with the event by offering a helpline during the day and remaining in contact with participants via Internet Relay Chat (IRC).

27. A flavour of the results of the hackday are discussed on the OKN blog. The hackday had also a wide media coverage in Europe.

28. Nevertheless, the open data community is somewhat anarchic in its approach: many projects are launched but don’t advance very far; some projects achieve short-term visibility but don’t appear to have any kind of long-term plan or user base. However, even these ad hoc proof of concept projects lead to ideas for

better ways to present our data. It will probably take several years before we can see, in the open data landscape, which projects have real staying power. But it is exactly the competitive approach which pushes the open data community to innovate freely

(b) Other examples of re-use of Eurostat data

- The Guardian newspaper (UK) has on its website Europe by numbers: the complete interactive guide
- The magazine Capital (FR) republishes economical indicators
- the Swedish company NComVA uses Eurostat regional statistics to demonstrate its World eXplorer data visualisation tool
- the Portuguese public data portal Pordata.pt
- some people from the german Karlsruher Institut für Technologie (KIT) setup a project which provides a complete mirror of Eurostat data expressed in RDF (Resource Description Framework)
- City Forward , an IBM-sponsored project to create a portal for public data on cities, gives users data visualisations based on Eurostat Urban Audit data
- EU unemployment infographic http://www.iiea.com/blogosphere/the-eu-unemployment-infographic
- European Commission's Open Data Portal which is currently under construction and will be made available later this year

III. The supporting infrastructure

A. Eurostat website

29. The Eurostat website is a trilingual portal (in English, French and German) to give access to European Union statistics (http://ec.europa.eu/eurostat). It is used to support the Internet dissemination policy of Eurostat. It is intended to be the single entry point to Eurostat's statistical information and services made accessible to the public audience.

30. The Eurostat website is built using Oracle Portal 10g and runs on the Oracle Application Server (OAS) on Solaris 10. The infrastructure is made of two middle-tiers, an Oracle repository (database and OID) and a single sign-on layer. It makes use of load balancing and cluster groups. Hosting is assured at the Data Centre of the European Commission.

31. Apart from so-called dedicated sections, which regroup all relevant information belonging to a specific statistical domain or topic, the website gives access to three main types of content:

- Tables and datasets: these are numerical tables of data. The site differentiates fixed 2-dimensional tables and multi-dimensional datasets. Each type has its own interface. The Data Explorer which focuses on displaying the data in a user customisable way and the Table, Graph and Maps tool which offers additional graphing and maps functionality. Currently the website hosts around 1150 tables and 4200 datasets.
- Statistical metadata: these are standardised explanatory texts about the data documenting data collection methods, quality, relevance, accurancy, comparability etc. Metadata accompanying the data is disseminated in accordance with the Euro SDMX Metadata Structure as defined by the European Statistical System (ESS). Currently the website hosts around 360 metadata files.
- Statistical publications on various statistical topics in pdf format. Currently the website hosts around 7000 publications.

All website content is classified according to 9 main statistical themes and about 70 subthemes.

32. Disseminating statistical information is a large undertaking. The average consultation figures for 2011 show that website had 3.1 million visitors and 2.8 million page views per month. There were 1 million dataset/table extractions, 1.5 million datasets downloaded from the bulk download facility and some 500,000

publication downloads in .pdf format per month. The site is among top 5 visited websites of the European Commission.

B. Bulk download facility

33. Fetching data automatically is being facilitated by the bulk download facility http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing which is a separated area of the website. The bulk download allows users

- to interactively browse the datasets and then download a complete dataset with a single click
- to automate the downloads if larger or more frequent updates are required via machine to machine interaction

34. Data in the bulk download area is available in tsv, sdmx-ml and dft formats. The sdmx-ml files are self explanatory and contain also the metadata. In order to complete the tsv files with the required metadata, a separate directory of the bulk download hosts all dictionary files used.

C. Web services

35. A beta version of a SOAP based web service is also available. An initial phase with a number of National Statistical Institutes as early adopters is ongoing. However, the current version still has a number of weaknesses with respect to the SDMX standards and web service guidelines and it is not available on Representational State Transfer (REST). More work will be carried to make it fully SDMX 2.1 and REST compliant this year. The web service supports

- a method providing complete list of public datasets available
- a method detailing the complete definition of a given dataset
- a method to download an entire dataset
- a method to extract specific data from a dataset by using sdmx queries

36. The bulk download area also offers a possibility to easily automate the download of entire datasets in a stepwise approach. The cornerstone is the Table of Contents file in xml format which contains all required metadata about all available datasets. This file is updated twice a day, at the same time when the data is being refreshed. The file contains for each dataset the following information:

- title of dataset in 3 languages
- short description of the table in 3 languages
- the dataset code which uniquely defines the dataset
- the date on which the dataset was last updated
- the date on which the dataset had the last structural change
- the start date of the data series in the dataset
- the end date of the data series in the dataset
- the total number of values in the dataset
- the unit value
- the URL to the ESMS metadata in html format
- the URL to the download location of the dataset to each of the available formats tsv, sdmx-ml and dft

```
- <nt:leaf type="table">
    <nt:title language="en">Life expectancy at birth, by gender</nt:title>
    <nt:title language="fr">Espérance de vie à la naissance, par sexe</nt:title>
    <nt:title language="de">Lebenserwartung bei der Geburt, nach Geschlecht</nt:title>
    <nt:code>tps00025</nt:code>
    <nt:lastUpdate>29.03.2012</nt:lastUpdate>
    <nt:lastModified>29.03.2012</nt:lastModified>
    <nt:dataStart>1999</nt:dataStart>
    <nt:dataEnd>2010</nt:dataEnd>
    <nt:values>842</nt:values>
    <nt:unit language="en">Years</nt:unit>
    <nt:unit language="fr">Années</nt:unit>
    <nt:unit language="de">Jahre</nt:unit>
    <nt:shortDescription language="en">The mean number of years that a newborn child can expect to live if subjected
      throughout his life to the current mortality conditions (age specific probabilities of dying).</nt:shortDescription>
    <nt:shortDescription language="fr">Nombre moyen d'années qu'un nouveau-né peut espérer vivre s'il se trouve tout
      au long de sa vie dans les conditions de mortalité du moment (quotients de mortalité par
      âge).</nt:shortDescription>
    <nt:shortDescription language="de">Mittlere Zahl der Jahre, die ein Neugeborenes voraussichtlich lebt, wenn die zu
      diesem Zeitpunkt herrschenden Sterbebedingungen während seines ganzen Lebens bestehen bleiben
      (altersspezifische Sterbewahrscheinlichkeit).</nt:shortDescription>
    <nt:metadata>http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/en/demo_mor_esms.htm</nt:metadata>
    <nt:downloadLink format="tsv">http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?
      file=data/tps00025.tsv.gz</nt:downloadLink>
    <nt:downloadLink format="dft">http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?
      file=data/tps00025.dft.gz</nt:downloadLink>
    <nt:downloadLink
      format="sdmx">http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListing?
      file=data/tps00025.sdmx.zip</nt:downloadLink>
  </nt:leaf>
```

Extract of table_of_contents.xml

37. Based on the information available in the Table of Contents file anyone with minimal programming knowledge can setup a procedure with limited effort to automate the download of the required datasets based on parsing the xml file and fetching the relevant information. This facility is heavily used by our big data consumers.

IV. Conclusion

38. Free access to and re-use of data is a cornerstone of Eurostat's dissemination policy, and it is precisely the reuse of its data—in all kinds of commercial and non-commercial projects—that gives Eurostat much higher visibility than it could achieve solely through its own dissemination products

39. As an example, working with Google resulted not only in data's being featured on the Google Public Data Explorer but also in the integration of data into Google search with Onebox. The Google search integration makes data sets searchable in 34 languages and ensures the highest ranking in search results. Currently, four Eurostat data sets have been integrated, which has significantly improved the overall visibility of its data.

40. Free access to and re-use of data is a cornerstone of Eurostat's dissemination policy, and it is precisely the reuse of its data—in all kinds of commercial and non-commercial projects—that gives Eurostat much higher visibility than it could achieve solely through its own dissemination products

41. Making data available in machine-understandable formats using open standards and metadata also enables the media or other data redistributors to easily pick up the data and integrate it into their own specific visualization tools for further dissemination. This enhances the visibility of the data and allows a statistical agency to reach a much broader audience with tools specifically targeted for such audiences.

42. We clearly see a shift in the way people are consuming statistics, they are becoming more and more demanding. People will make their own decisions on what they want to use our data for, and how they want to

present it. As a data producer we will not always be able to respond to this increasing demand for visualisation of statistical data. Therefore we will have to focus more on supplying timely data in formats that can be easily re-used.