Official Statistics beyond web 2.0
Developments, rewards and risks to come

Armin Grossenbacher
Swiss Federal Statistical Office FSO

Summary

We are all familiar with the old web of the nineties. It brought about a revolution in publishing by allowing global access to data anytime, anywhere, and gave a tremendous impulse to the global dissemination of official statistics.

We are also aware that a further web (r)evolution has taken place. This new wave has lead users from the traditional "read-only web" to the more revolutionary "read-and-write" web. Everyone has now been given the possibility to voice their opinion and disseminate information to any and all interested parties, creating in the process a social web or, as it is technically called, the web 2.0. The response of official statistics to web 2.0 has been highly cautious.

But meanwhile, yet another web (r)evolution is underway. This one is not about the users anymore but more about access and it is characterised by technological breakthrough and the pivotal use of metadata: this is the semantic web or web 3.0.

This presentation will propose a short overview of the internet's evolution and will especially illustrate its specific impact on the dissemination of official statistics. The focus will be on what is taking place behind this web of linked data (or semantic web) in a world of increasingly globalised data.
What are the challenges, the potential rewards and the risks official statistics will be facing as they are confronted with this new web?

1. Official Statistics and the Real World

The mission of official statistics, dictated by legal and ethical guidelines at national and international levels, is to collect data in an independent manner on the state and evolution of a country or region, using scientific methods and by observing the principle of data protection. Official statistics describe the “real world” with figures and are as such 'an indispensable element in the information system of a democratic society' [1]. A figure or a statistical indicator can express situations and developments fast and concisely, coming to our aid as a powerful, internationally intelligible information medium.

In order to perform this mission to inform, official statistics made early and consistent efforts to render the findings of surveys and analyses available to a wide public. In order to make statistical information useful, for knowledge to be generated from bare figures, data providers must 'disseminate information relevant to people; present it in a way that people can relate to their own interests; use language/tools coherent with those used by people in other contexts.' [2]
The history of these dissemination activities is long. It began with systematic publication in the 19th century and continued to evolve with the use of the newly emerging media. Precisely in the activity of dissemination, official statistics is itself part of the “real world” and what it encounters there are, on the one hand, information technologies which are continually changing at an ever faster rate and, on the other – in particular in recent decades – increasing numbers of suppliers of statistical data outside of official statistics.

Over the course of the last two or three decades, the setting that official statistics have been working in has changed dramatically. Official statistics – and the same is true for other suppliers of information – cannot afford to rest on their laurels. It is essential to observe what is happening in the area of information services, to see what clients of statistical information expect as standard and to react with appropriate action regarding dissemination activities. It is more important and urgent than ever to get

to grips with the new dissemination possibilities and also with the different ways that target groups and their specific information needs can be reached with high-quality official statistics.

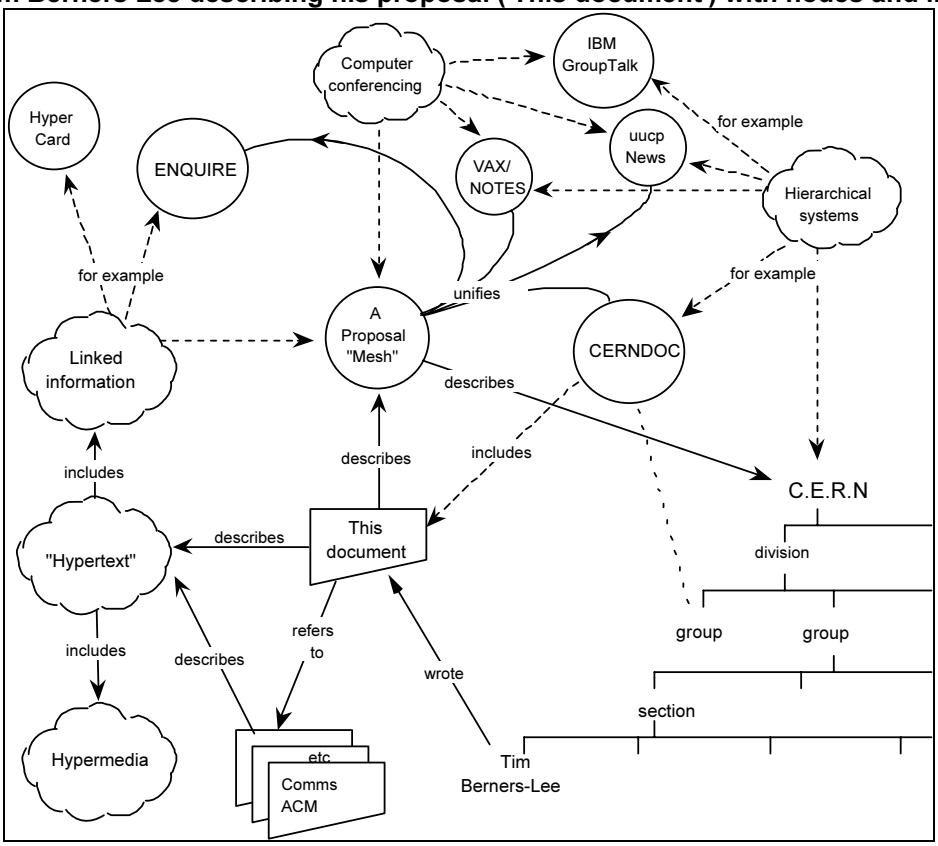
2. Information and Dissemination Services in Web 1.0 and Web 2.0

The formative event in information technology and for information behaviour was the arrival of internet. On the basis of this communications protocol an increasingly complex and networked information offer developed. The web was born and has since developed rapidly from web 1.0 to web 2.0 and now to the much discussed web 3.0.

2.1. Web 1.0

Web 1.0 describes the web's first decade (1989-1999). It arose from a proposal, written by Tim Berners Lee, 'showing how information could be transferred easily over the Internet by using hypertext, the now familiar point-and-click system of navigating through information'[3]
This web includes an enormous collection of documents (texts, pictures, videos, accesses to databanks etc), which can be accessed by a browser with the support of a search engine.

Graph 1 Tim Berners Lee describing his proposal ('This document') with nodes and links [4].



With this system, an information revolution was underway, with information being available from everywhere, at any time and as complete as the supplier allowed. This was, in effect, an information revolution.

2.2. *The fast and innovative development of web 1.0 by official statistics*

Official statistics reacted rapidly with diversified dissemination activities, not least because the statistical agencies were very experienced with informatics in their daily work.

Statistical agencies websites' started by presenting institutional aspects, then moved to a thematic presentation of statistical content, reflecting the growing user-oriented point of view that users were looking for thematic information and rarely for information about the statistical office itself.

In 2001, a report at the 11th Nordic Conference on Information and Documentation, after having examined a handful of National Statistical Institutes' websites (NSIs) concluded that:
'These websites are all very rich in content. … All display basic statistical overviews for their countries free of charge on their websites. … There is a clear trend towards utilizing the website as the main tool for offering access to the output databases of the institutions. [5, p.8]
'The main conclusion of this review is that in the field of official statistics there have been enormous improvements in the dissemination by national statistical institutes, in range, in quantity and in quality. The development of the internet and of internet use has made this possible and been the driving force.' [5, p. 11].

Although the different statistical agencies' websites may differ in appearance, since 2001 a certain standard of functionalities has emerged. Elements of this 'standard' of statistical agencie's websites encompass (not exhaustive):

- Prioritised thematic access and thematic navigation
- Extended search facilities
- Database access (mostly free of charge) with interactive tabulation functionalities
- Extended news services with release calendars and archives of news releases, alerts via RSS feeds
- Freely accessible publications (mostly in PDF format)
- Trend towards presenting data in a user-friendly way, providing diagrams and maps visualising the data
- Story-telling approach by providing textual explanations or even stories
- Offers for special target groups like journalists (newsrooms) and schools
- Statistical literacy programs.

As far as the offer of user-friendly access is concerned, statistical agencies are in the front rank in various areas. For example, individual development of databank queries (such as CANSIM in Canada or STATINF in Switzerland) had already begun before the arrival of the internet and were further developed with widely used, innovative forms of access to the web (such as PC-Axis from Sweden and the north European States or OECD.stat). The diversity of initiatives in the area of visualisation of statistical information is particularly worth noting. Visualisation is seen as the ideal way to make accessible data that are often hard to grasp in pure table form. Static and, increasingly, interactive diagrams and maps which can be modified by users themselves, allow simple and illustrative navigation through data bases. A good insight into this can be found in the new 2009 edition of UNECE's Style Guide on the Presentation of Statistics, in particular on the subject of thematic maps. See also ongoing conversations in the Blog about Stats under the tag "Data visualization" [6].

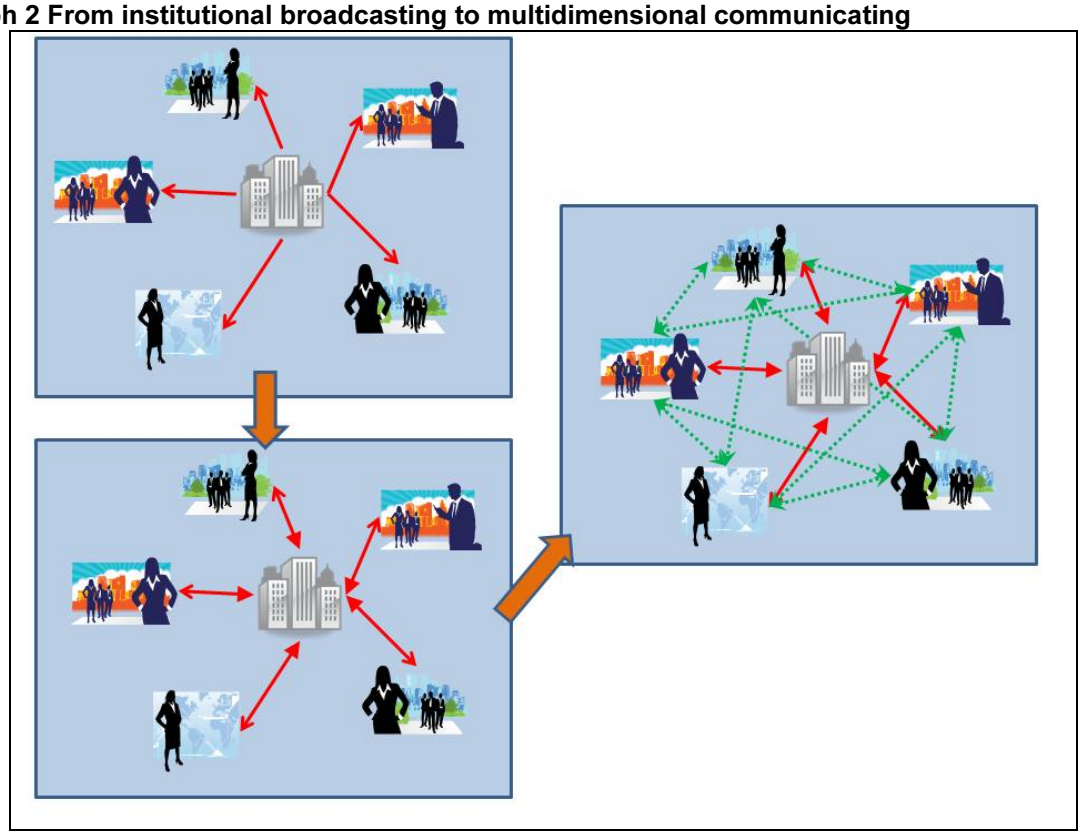
2.3. Web 2.0

Web 2.0 describes the second decade of the web (1999-2009) and brings with it (based upon and in addition to web 1.0) new functionalities and types of use. Web 2.0 leads users from the traditional "read-only web" to the more revolutionary "read-and-write" Web. Everyone has now been given the possibility to voice their opinion and disseminate information to any and all interested parties, creating a social web in the process. 'In particular blogging tools, social bookmarking tools, social networks, social media sites, [wikis] and micro blogging services began to organise the web around people and their relationships.' [7] 'The first ever web browser was also an editor, making the web an interactive medium, the problem was that it only ran on the NeXTStep operating system. With recent phenomena like blogs and wikis, the web is beginning to develop the kind of collaborative nature that its inventor envisaged from the start.' [8]

The traditional way where institutions and professionals broadcast information to an amateur public is fading away. Many-to-many communication is replacing one-way communication. Consumers and

producers meet on the same platform, joining all media types together in digital form. Web 2.0 is a global, social medium which is available everywhere through connections that are becoming more and more rapid (and more often via mobile appliances) and which is becoming cheaper all the time. People are extending their immediate surroundings, saving data on the web (clouds) and entering a flow of continually flowing public and private information in which they themselves can be active. Reality is undergoing a hitherto unknown expansion.

Graph 2 From institutional broadcasting to multidimensional communicating



Multi-dimensional communication is made possible in web 2.0 by various platforms which may be social networks such as Facebook or LinkedIn, photo platforms such as Flickr or Microsoft Live or blog systems such as Wordpress or Bloglines. Active web 2.0 users do not usually limit themselves to just one of these platforms, thus entailing an increasingly complex management. The necessity of bringing all these platforms under one roof is slowly but surely increasing. Latest developments are going in that direction: with Google Wave [9] announced for Autumn 2009, real-time exchange with as many partners as one wishes in various forms and on one platform should be possible. Email, which is now 40 years old, will be replaced by a personal communications centre, where an unlimited number of people can contact one another via a central server (whether this is with Google or elsewhere). Other approaches turn the browser into a server itself and allow people to communicate from browser to browser without having to upload content to third party suppliers. There seems to be no end to the possibilities. [10]

2.4. The official statistics' response to web 2.0: highly cautious

Active **blogging** and commenting facilities are not often found within statistical agencies; discussing disseminated results on their own websites has no visible tradition. However, blogs are observed via specialised search engines to get knowledge on how the public deals with statistical issues, and several statistical agencies offer to embed their content in personal blogs. In particular **microblogging** (such as twitter.com) is used in addition to RSS feeds (which do not seem to be enjoying wide or fast-spreading appeal)) as a tool for spreading statistical news, as for example Statistics Sweden's or Statistics Slovenia's use of twitter since 2009.[11]

Collaborative websites or **wikis** enable anyone to contribute or modify content thereby building a common knowledge. In its strategy to turn statistics into knowledge, the OECD launched what was probably the first 'statistics' wiki (called 'Wikigender') on 7th March 2008. Interested people meet on this platform to exchange information and statistics on gender equality, producers and consumers work together in a close collaboration [12].

A study conducted by Jessica Gardner (UNECE) in August 2008 provides a good insight into the use of web 2.0 by statistical agencies: 'A review of websites suggests that except for RSS, there is limited public use of web 2.0 by statistical offices. This was confirmed by an informal survey of statistical dissemination and communication professionals conducted by the UNECE. … The reasons for limited activity in this area include the need to be cautious in order to maintain public credibility, managing with limited resources and focusing on the upcoming census. The most common applications are usually within organizations, but opportunities clearly exist for interacting with external customers and users of statistical information'. [13, p 1 and 6]

Outside official statistics there is much more use of web 2.0 for providing statistical information. Visualising tools, commenting functionalities and possibilities to share statistics are the star attractions of various private platforms. Specific data oriented social platforms such as YouTube or flickr let people upload statistics and provide some of the mentioned tools to them.
A short list:

- Many Eyes (2004).
- Swivel (December 2005)
- Statista (2007, in German only)
- Gapminder and Gapminder Graphs Community (2005/2007).
- Google motion charts and Google public data (2009)
- Timetric time series analysis [14]

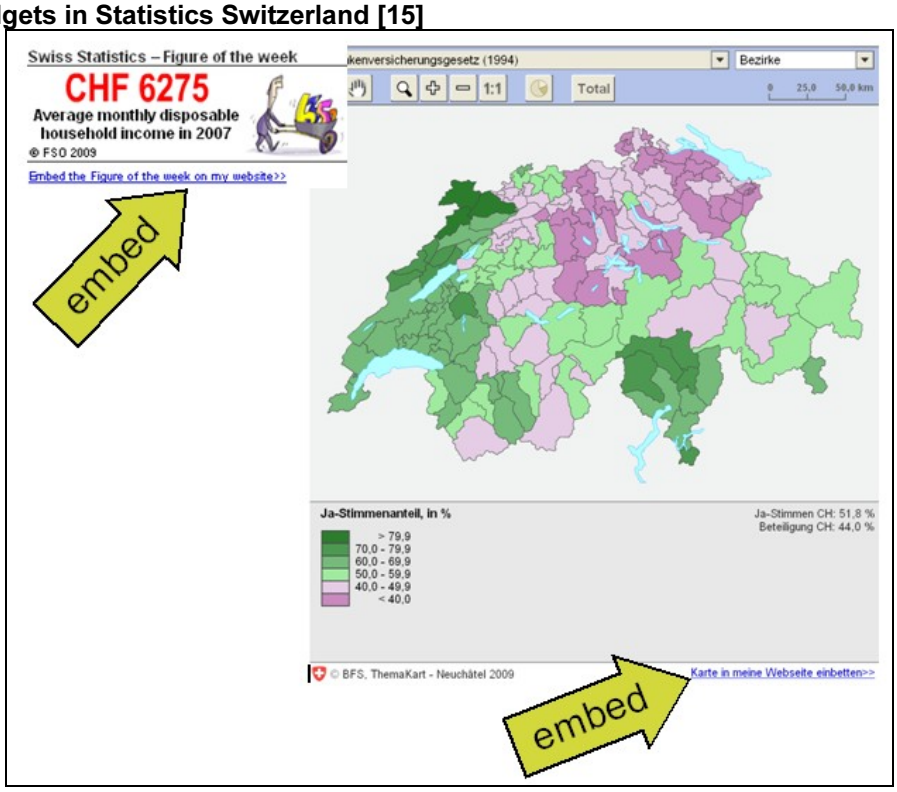
Some of these platforms collaborate with statistical agencies and are helping to disseminate and make statistical data understandable to users. Generally these offers do not seem so far to be in wide demand (except for Gapminder presentations). Statistics stand here alone and are not embedded into topic areas.

3. The Challenge of Globalised Data and Web 3.0

The web is a giant collection of connected websites and documents, but for some time it has also been – in addition to the social aspects of web 2.0 - a lot more than that. Content leaves its original websites to be found embedded in numerous other places. Interactive road maps (e.g. from Google) or book advertisements (e.g. from Amazon) can be found scattered over numerous websites and it is no longer necessary to change websites in order to obtain this specific information. Documented interfaces (application programming interfaces, APIs) to equally open databanks make it possible to embed content from third-party providers and to create so-called mashups with content from different sources: data are more and more globalised. This is partly due to web users wanting to find as much information as possible in the same place on their subject of interest and partly a marketing objective on behalf of suppliers to be as widely present and known as possible with their offer and brand name.

Official statistics have so far made little use of this: RSS feeds are a first step, widgets – small applications embedded in other websites – are still rather rare, as are open interfaces to databanks.
The figure of the week or the interactive maps in Statistics Switzerland are an example of widgets that can be embedded into other websites while maintaining the link to the source, thus allowing content to always be up to date.

Graph 3 Widgets in Statistics Switzerland [15]



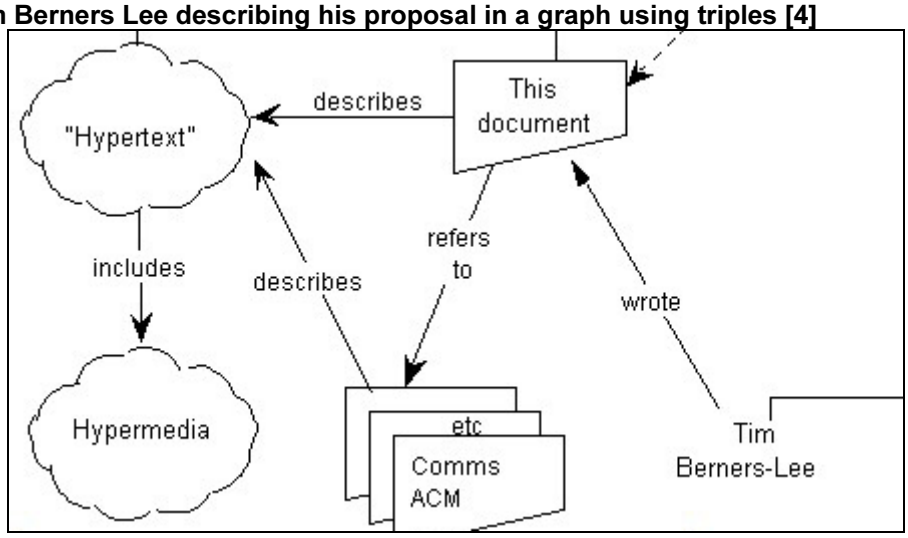
The Neighbourhood Statistics (NeSS) Datastore offered by the UK Office for National Statistics is an example of a statistical databank whose access is open via an API. [16]

Combining content from several sources into mashups by applying the respective rules of these APIs is programmers' work. This is where the semantic web or web 3.0 comes into play. Web 3.0 is not a new web, but a further development of the existing one: through improved description and structure of the contents with metadata, the contents are given meaning which can be "understood" by machines. On this basis, software agents can locate content (resources), take it and embed it in other applications, thus creating mashups. The norms for the description of resources have been established by the World Wide Web Consortium (W3C), in the Resource Description Framework (RDF). Similar to the query language SQL used for relational databanks, SPARQL is an RDF query language for documents which are structured in RDF.
In RDF each content or resource (a book, a person, an indicator) is decribed through subject-predicate-object expressions (called triples). In his above-mentioned proposal, Tim Berners Lee gives an example:

„This document – describes – Hypertext“
Subject – predicate – object

The visual presentation of a sum of triples is a so-called graph.

Graph 4 Tim Berners Lee describing his proposal in a graph using triples [4]

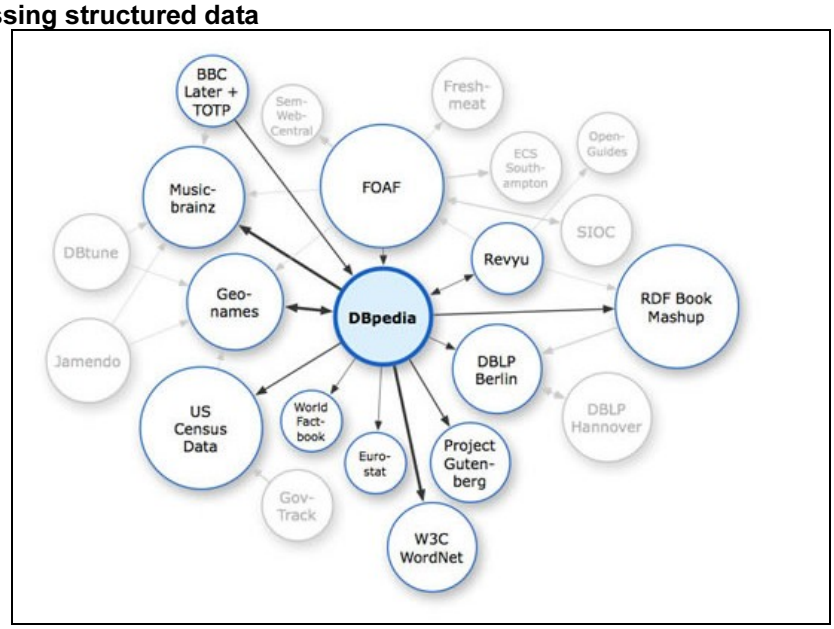


'Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully. Computers can adeptly parse Web pages for layout and routine processing, - here a header, there a link - to another page, but in general, computers have no reliable way to process the semantics. ….The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. …..Adding logic to the Web - the means to use rules to make inferences, choose courses of action and answer questions - is the task before the Semantic Web community at the moment. … The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine-readable Web content and automated services (including other agents) become available.' [17]

The web composed of structured, inter-connected data (and not just documents) that can be read by applications is already a tangible reality today. Numerous areas are documented with metadata. Vocabularies and ontologies exist in order to structure topic fields, for example Dublin Core for the description of documents and publications or Friend of a Friend (foaf) a 'machine-readable ontology describing persons, their activities and their relations to other people and objects.' [18]

Whole data collections have been described in RDF, have been “rdf-ised” and can be queried in the same way as a databank by SPARQL or with a special semantic browser. The embedding of RDF attributes into HTML coded websites is also on offer (RDFa). Dbpedia is an example of this, 'a community effort to extract structured information from Wikipedia and to make this information available on the Web. ... The DBpedia knowledge base currently describes more than 2.6 million things, including at least 213,000 persons, 328,000 places, 57,000 music albums, 36,000 films, 20,000 companies. The knowledge base consists of 274 million pieces of information (RDF triples).'[19]

Graph 5 Accessing structured data

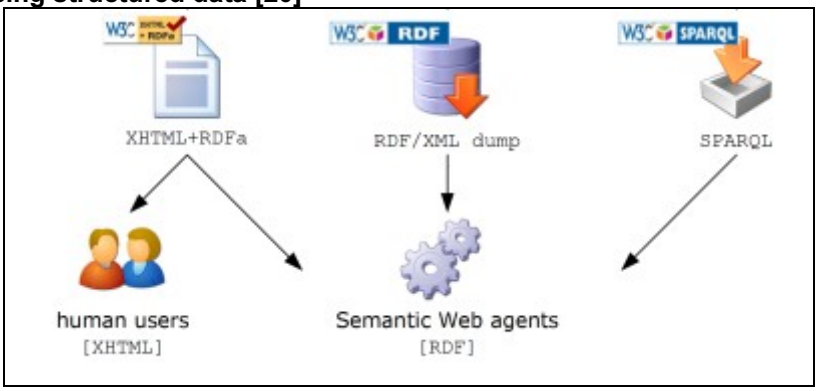


The results of such queries do not provide – as is the case with normal searches in the unstructured web – indications to websites or documents with the information one is looking for, but instead the data itself. Thus queries such as 'Give me all cities in New Jersey with more than 10,000 inhabitants' are possible, the result of which will be a list of towns. The semantic web therefore allows the step to be taken from the web of documents to the web of data. Traditional search engines such as Yahoo and more recently Google place more and more emphasis on the search in structured data, thereby supporting the development of the web of (linked) data; others like sig.ma.com are made for structured data.

Such offers are not yet available in official statistics. Due to their exemplary organisation, statistical data are, however, supremely suitable for a presentation using semantic web technologies. Various universities and research organisations have therefore also been looking into statistical data collections and opened these up in the form of RDF triples. To this end they have downloaded existing data offers – mostly in the form of Excel or csv, transformed them into RDF and made them available via interfaces. A noteworthy example is Eurostat data transformed by Joanneum Research [20]. These rdf-ised Eurostat data are presented via diverse interfaces and made accessible for humans and machines. Three variations are possible:

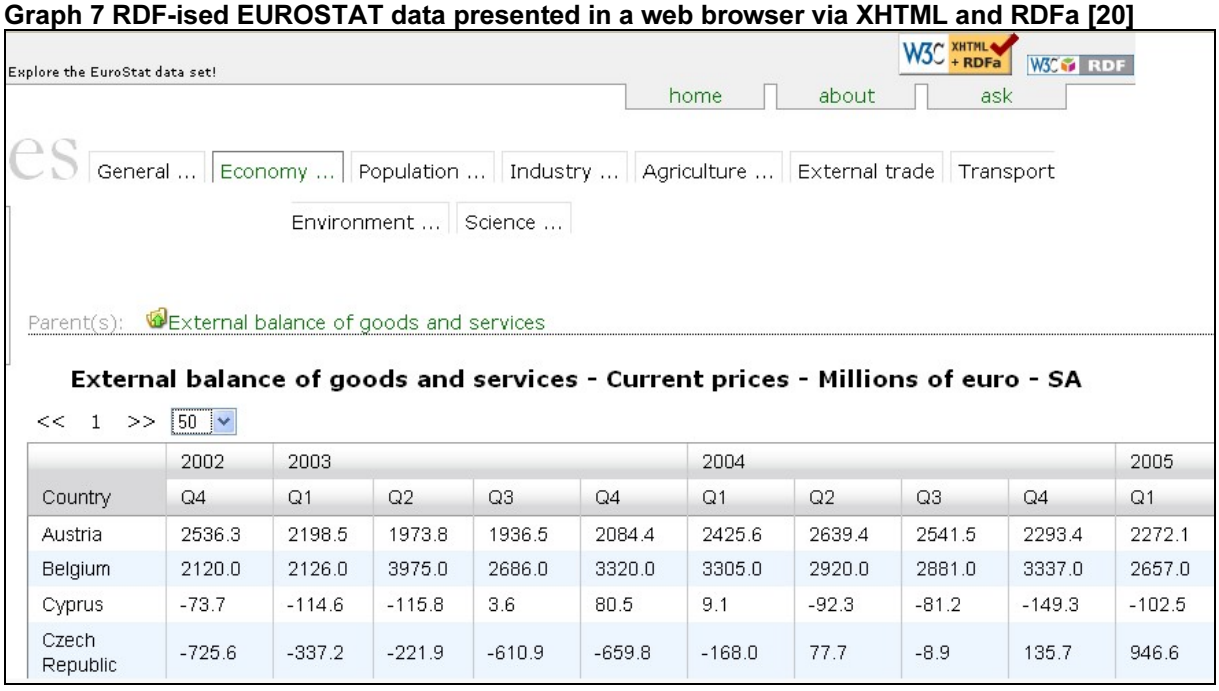
- XHTML with integrated RDF triples (RDFa), is human readable and displayed in a standard browser.
- Pure RDF files, which are readable by applications/machines (semantic web agents) or also with special semantic browsers (such as Tabulator) or accessible via an
- enquiry similar to databank queries (SPARQL queries, SPARQL-Endpoint):

Graph 6 Accessing structured data [20]



The following graphic 7 shows the presentation of structured Eurostat data in a popular browser with embedded RDF triples (RDFa):

Graph 7 RDF-ised EUROSTAT data presented in a web browser via XHTML and RDFa [20]



	2002	2003				2004				2005
Country	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1
Austria	2536.3	2198.5	1973.8	1936.5	2084.4	2425.6	2639.4	2541.5	2293.4	2272.1
Belgium	2120.0	2126.0	3975.0	2686.0	3320.0	3305.0	2920.0	2881.0	3337.0	2657.0
Cyprus	-73.7	-114.6	-115.8	3.6	80.5	9.1	-92.3	-81.2	-149.3	-102.5
Czech Republic	-725.6	-337.2	-221.9	-610.9	-659.8	-168.0	77.7	-8.9	135.7	946.6

4. Conclusion

From web 1.0 to web 3.0, the web has undergone fundamental change in the space of a few decades. With each further development, suppliers of information are confronted with new challenges which they will have to respond to with their information and dissemination policies.

Official statistics met the challenges of web 1.0 rapidly and with innovative methods. Statistical websites are among the most elaborate and attractive suppliers of information: well structured, with a comprehensive visualisation offer, with user-friendly, interactive databank access and, last but not least, the aspiration to promote statistical literacy.

Web 2.0 demands answers to the question of the extent to which official statistics should be present in public discussions and its future role in such discussions. To what extent do we wish to allow commentary on statistical results, to become involved in discussions? The risk of creating new problems for oneself is slight; where data leads to debate, official statistics are called upon anyway. Web 2.0 could be of use in this area.

It is early days for the semantic web 3.0, but it is those who advocate the structured presentation of data, of linked data, who are currently demanding statistical data in an even wider form and its availability to the public on the initiative of the suppliers. [21]
Technically, in view of the well-developed metadata, official statistics are extremely well-equipped for such accessibility. Ontologies exist in various areas (to use semantic web jargon). With the widely recognised SDMX initiative within official statistics an objective is being pursued which corresponds to the semantic web: 'the reliable location of data ... the user could conduct a search and then ask for the data to be delivered as a single HTML page, or as a PDF, or in some other format ' [22]
On the outside as well as within official statistics, "linked data" constitute an intensive endeavour in this direction. Whether and how the initiatives will meet up, remains open.

In the meantime, however, official statistics has to answer various fundamental questions. Questions regarding accessibility by the public of whole databanks for example and the globalisation of statistical data when statistics forsake the homepages of statistical institutions to be integrated everywhere.

There are questions regarding the guarantee of data protection and finally regarding observation of copyright. Because official statistics are ultimately responsible for the quality of data that is publicly available and is used by third party applications.

In the face of web 2.0 and web 3.0, official statistics must ask itself what presence, if any, it should have in the evolving information standards and cultures and whether it can afford to lag behind. The risk of being excluded from new forms of data presentation and also the risk of being marginalised to a certain extent by innovative data suppliers outside of official statistics must be weighed against the rewards of an active and extensive presence on the part of official statistics and their high quality information offer thanks to the use of new technologies.

References

1 United Nations Economic Commission for Europe, *Fundamental principles of official statistics in the region of the Economic Commission for Europe,* 15 April 1992, http://www.unece.org/stats/archive/docs.fp.e.htm

2 Enrico Giovannini, Chief Statistician OECD, *the Role of Communication in Transforming Statistics into Knowledge*, OECD. Seminar on Innovative Approaches to Turning Statistics into Knowledge, 26-27 May 2008, Stockholm http://www.oecd.org/document/54/0,3343,en_21571361_31938349_39800246_1_1_1_1,00.html

3 *The website of the world's first-ever web server*, http://info.cern.ch

4 Berners Lee, Tim, *Information Management: A proposal*, 1989. http://www.w3.org/History/1989/proposal.html

5 Snorrason, Hallgrímur and Gylfadóttir, Jóra, *Trends in the dissemination and pricing of official statistics*. Paper presented at the 11th Nordic Conference on Information and Documentation, Reykjavík, Iceland, 30 May 2001

6 UNECE, *Style Guide on the Presentation of Statistics* (http://www.unece.org/stats/documents/ece/ces/ge.45/2009/wp.24.e.ppt)
Blog about Stats (http://blogstats.wordpress.com,
tag data visualization: http://blogstats.wordpress.com/category/031-data-visualization/).

7 Spivack, Nova, *Blog Minding the planet* http://novaspivack.typepad.com/nova_spivacks_weblog/2009/05/the-next-generation-of-web-search-search-30.html, and http://www.twine.com/item/128lryv9z-46/welcome-to-the-stream-the-next-phase-of-the-web.

8 *Tim Berners-Lee's original World Wide Web browser*, http://info.cern.ch/NextBrowser.html

9 *Google Wave*, http://wave.google.com/

10 Opera with opera unite (http://unite.opera.com) or Codelathe with Tonido (http://www.tonido.com) are suppliers of this communication philosophy: 'We believe that a handful of companies controlling everybody's data is not healthy in the long-term and could lead to serious privacy issues and reduced freedoms. Tonido is a step against this trend, empowering customers and giving them back the control and ownership over their applications and data'. http://www.codelathe.com-

11 http://twitter.com/scb_nyheter
http://twitter.com/StatSlovenia

12 http://www.wikigender.org and Seminar on Innovative Approaches to Turning Statistics into Knowledge, 26-27 May 2008, Stockholm Sweden: http://www.oecd.org/dataoecd/30/32/40293021.pdf

13 Gardner, Jessica (UNECE), *Blogs, wikis, and official statistics: new perspectives on the value of Web 2.0 by statistical offices*, International Marketing and Output Database Conference (IMAODBC), Finland, 2008

14 *Many Eyes Website* http://services.alphaworks.ibm.com/manyeyes/home
Swivel Website www.swivel.com
statista Website http://de.statista.org/
Gapminder Websites www.gapminder.org and http://www.gapminder.org/graphs.html
Google public data, http://www.google.com/support/websearch/bin/answer.py?answer=144522&&hl=en and http://www.google.com/publicdata?ds=uspopulation#met=population&idim=state:06000
timetric time series analysis http://timetric.com/

15 Statistics Switzerland, *Figure of the Week Widget* http://www.bfs.admin.ch/bfs/portal/en/index.html
Interactive Map Widget

16 NeSS Data Exchange, http://neighbourhood.statistics.gov.uk/dissemination/Info.do?page=nde.htm
The NeSS Data Exchange is a new online service designed to make it easier for you to interrogate and engage with the Neighbourhood Statistics (NeSS) Datastore. Using machine to machine technology it will speed up and simplify your data requests and still allow you access to over 300 geographically referenced datasets, free of charge.

17 Berners-Lee, Tim, Hendler, James and Ora, Lassila, *The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.* Scientific

American Magazine, May 2001, http://www.scientificamerican.com/article.cfm?id=the-semantic-web&p...
18 Dublin Core http://dublincore.org/documents/dcmi-terms/#terms-subject,
Friend of a Friend (foaf), http://en.wikipedia.org/wiki/FOAF
19 DBpedia, http://wiki.dbpedia.org/About and http://wiki.dbpedia.org/Interlinking.
20 JOANNEUM RESEARCH (one of the largest non-university research institutions in Austria), http://www.joanneum.at/en/jr.htm, RIESE=**R**DFizing and **I**nterlinking the **E**uro**S**tat Data Set **E**ffort). http://riese.joanneum.at/about.html
More on Official Statistics and Semantic Web in Blog about Stats:
http://blogstats.wordpress.com/2009/04/25/semantic-web-and-official-statistics/
21 Berners Lee, Tim *Raw Data Now!* http://blogstats.wordpress.com/2009/04/09/raw-data-now-tim-berners-lee-and-hans-rosling/ and
http://www.ted.com/index.php/talks/tim_berners_lee_on_the_next_web.html
22 *SDMX user guide, version 2009*.1, January 2009, pages 39/40
'SDMX Registries are designed to support focused queries on a known set of data and/or metadata, utilizing their knowledge of the structural metadata found in DSDs and MSDs. Thus, the reliable location of data becomes possible. …... If the SDMX Content-Oriented Guidelines are also followed, then registry searches can be even more powerful, as all of a common set of concepts will share a name and a representation. Another ubiquitous catalogue functionality is data retrieval. An SDMX Registry contains information about the location and formatting of the data and metadata. This makes it possible for the catalogue application to retrieve one or more files at the user's request, and pre-process it so that only the data or metadata of interest is retrieved. If the data is coming from a variety of sources – a set of static files, and maybe from some real-time database queries – the catalogue application can hide all of the complexity of assembling the data into a useable set expressed in a standard XML format. … Thus, the user could conduct a search and then ask for the data to be delivered as a single HTML page, or as a PDF, or in some other format. '