

Work Session on Statistical Metadata  
(Geneva, Switzerland, 18-20 February 1998)

Item 6 of the provisional agenda

## **INTERNET ACCESS FOR STATISTICAL DATA AND METADATA**

Submitted by the Berlin Statistical Office, Germany <sup>1</sup>

---

<sup>1</sup> Prepared by Rudolf Frees.

# Internet access for Statistical Data and Meta-data

*Rudolf Frees • Berlin Statistical Office*

The rise and great success of the World Wide Web is mainly based on the technical standards used:

- Hypertext Markup Language (HTML) offers easy to be used possibilities to give a document a formal structure, which is easy and clear visualized,
- Different kind of media (text, graphs, sound, video etc.) can easily be used with one common application,
- Various documents that are related in content one to each other can easily be linked using the technique of hyper-link references.

And above of all, such kind of multimedia, clear structured and linked documents can be read and used by applications running on all major Operating Systems (Windows, NT, MacIntosh, UNIX etc.). And the TCP/IP based transport protocol for html-files – http – makes it possible, that computers with different operating systems can communicate one with each other in a way that users do not have to know complicated commands and difficult syntax, and so exchange all that kind of information.

This very user friendly environment caused a tremendous flood of information, which is worldwide accessible. The problem that now occurred was (and somehow still is), to find quick and efficient that information within this oversupply, that really meets the user's needs. To solve this problem different kind of search engines had been developed. With the help of one (or several) such search engine, one can find quiet easy specific information, that may be stored on servers, one never knew before.

However all the advantages of the above described, platform independent techniques had a certain price. From the perspective of software development HTML and HTTP are causing some restrictions. Things like complex transactions, interface design etc. stay behind the possibilities given by development environments and programming languages used to develop software that runs on stand alone PCs or within local area networks.

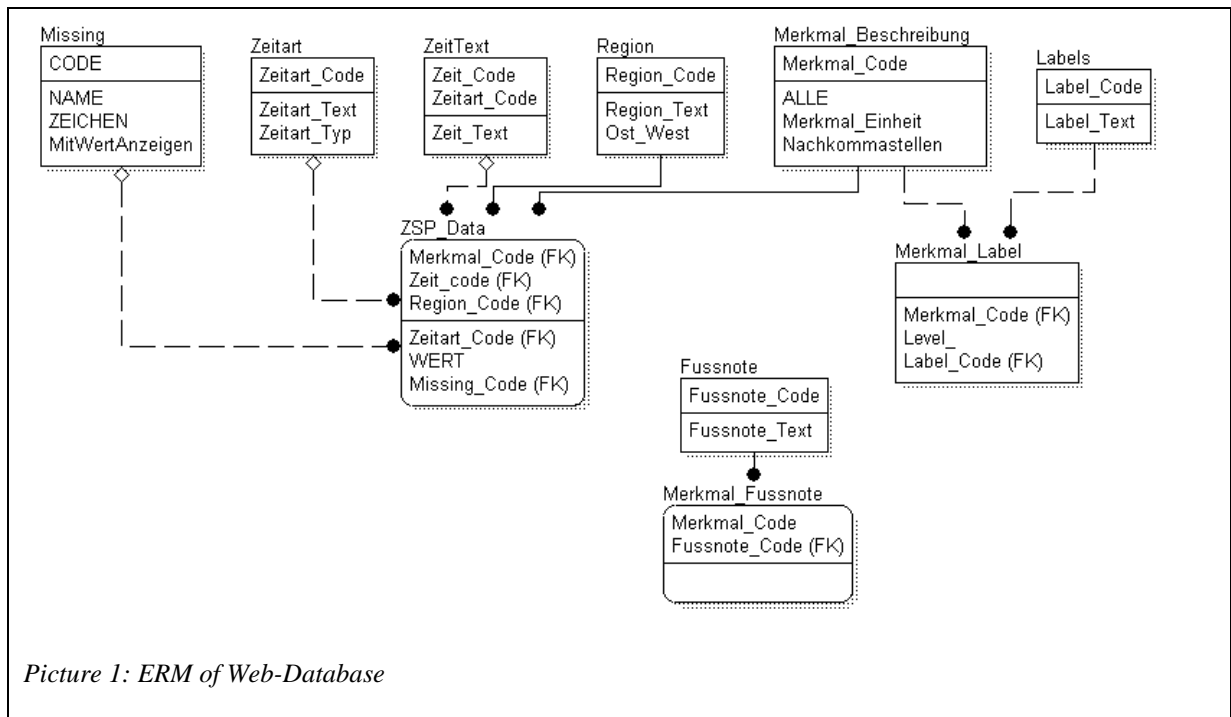
During the last one or two years things improved a lot. Gateways between Web-Servers and Database-engines are nowadays available, that makes it easily possible to provide html-based front-ends to use query-techniques to explore databases very flexible and to dynamically generate result sets of a query as html formatted output. Stuff like JAVA-Script, JAVA, ActiveX, dHTML etc. increases the possibilities to develop more interactive user interfaces.

Among those new facilities the use of an efficient database gateway seemed for the Berlin Statistical Office the most important to design and develop it's own WebSite. Mainly two reasons were responsible for this decision:

1. Many visitors of a official statistics WebSite are basically interested in specific data. They want to select and compile data, they need to describe and/or analyze trends or structures. For those kind of users it is not very efficient to offer just precompiled and pre-selected data-sets. An easy to be used instrument to select and compile "their" data-sets out of our comprehensive database should meet exactly those users needs.
2. To keep the data of more than 300 surveys up-to-date is a tremendous work for our staff. Those limited resources we have will never be enough to update all the static html-files we would need, to present comprehensive reports on the Internet. Only with an architecture where we have just to update tables in a database and the most recent figures will be automatically available for the WebUsers, makes it able to solve this supply-problem.

## Current design of the Berlin Web-Database

When we started to design the "Web-able" database we concentrated on the questions, how to store the empirical data in a way, that they can be very flexible used to generate html-based query interfaces and output. The following picture gives a rough impression of the main part of this basic design.



The main table of the ERM is **ZSP\_DATA**. It contains the data (WERT), codes indicating the variable (Merkmal\_Code), time (Zeit\_code), region (Region\_Code), periodicity (Zeitart\_Code) and Missing-values (Missing\_Code). Labels and other meta-data related information are stored in the other tables of the ERM. Time- resp. Period-labels in the table "Zeitart" and "ZeitText", region-labels in "Region" and Variable-labels in "Merkmal\_Label" and "Labels". Information about measurement-units are stored in "Merkmal\_Beschreibung". The table "Fussnote" contains footnotes and the many-to-many relationship between footnotes and variables is contracted via coding of footnotes and the field "Merkmal\_Code".

A set of procedures (some running in the context of the database server as stored procedures and views, some running in the context of the WebServer) is getting data from the database to build as well the retrieval interface as the final output of a retrieval. The retrieval has three stages: At first the user can select a domain of data like population, labor market, manufacturing industries, wages etc. He has the choice to see only a list of available domains or to get the domains together with an overview of some descriptions (like variables and keyword, regions available, periodicity and time-period covered). If he wants to, he can also look at some more detailed descriptions explaining methodological aspects, legal basis, definitions etc.

The following picture shows stage 1 of the Retrieval-Web-Frontend (Bevölkerung=population, Arbeitslage=labor market, Verarbeitendes Gewerbe=Manufacturing, Energieversorgung=energy supply etc.). What seems to be a normal list of hyper links is indeed not a static html-file but is dynamical



Picture 2: HTML-based database retrieval • stage 1

generated. That means, whenever a new domain is loaded on the database it will be appear automatically on this page without a need to update an html-file.

The following picture shows stage 1 of the retrieval, enriched with some descriptive meta-data.

Datenbestands-Übersicht:						
Sachgebiet	Untergliederung/ Stichwörter	Verfügbare Regionen			Periodizität/ Zeitraum	Anzahl der Zeit- reihen
		Berlin	Ost/West	Bezirke 1		
<u>Bevölkerung</u> Meta	Bevölkerungsstand, Ausländer, Geschlecht Bevölkerungsbewegung, Eheschließungen, Lebendgeborene, Gestorbene Wanderungen, Zuzüge, Fortzüge	X	O/ W	X	monatl./ Jan. 94 bis Sep. 97	31
<u>Arbeitslage</u>	Arbeitslose, Ausländer, Geschlecht, Alter, Arbeitslosenquote Kurzarbeiter, Offene Stellen, Arbeitsvermittlung, ABM, Bildungsmaßnahmen	X	O/ W	X	monatl./ Aug.93 bis Nov. 97	17
<u>Verarbeitendes Gewerbe</u> Meta	Betriebe, Beschäftigte, Grundstoff-, Vorleistungs-, Investitions-, Gebrauchs-, Verbrauchsgüterproduzenten, Arbeiter, Angestellte, Arbeitsstunden, Löhne, Gehälter, Umsatz, Energieverbrauch, Auftragseingangs-, Produktionsindizes	X			monatl./ Jan 94 bis Nov. 97	61
<u>Energieversorgung</u>	Strom, Erzeugung, Verbrauch, Gasverbrauch	X			monatl./ Aug. 93 bis Sep. 97	3

Picture 3: HTML-based database retrieval • stage 1 with descriptions

On stages 2 and 3 the user can select one or several variables and some or several periods and regions. The selectable items are presented within listboxes, which are common html-objects. The items are filled into the listboxes by procedures, which retrieve the items from the database. So it can be assured, that only variables, periods and regions are selectable for which entries are really available, and - for example after an upload of new variables and/or periods – users automatically get the information, that these new entries exist. No html-file or hyperlink has to be updated, to bring the information about the extended contents of the database to the front-end. The specification of a query - working with traditional query-languages like SQL a more or less difficult action – can so be done interactively with some mouse clicks on common html-controls (listbox-items and transmit buttons), dealing with understandable “labels” instead of cryptic fieldnames. At the end of a retrieval server sided procedures fetch the resultset from the database and “wrap” it with html-code. In this sense the html-page is dynamical generated, instead of being statically stored within the filesystem of the WebServer.

The following two pictures show how the dynamically generated result of a retrieval not only varies in the number of table rows (depending on how many variables the user has chosen) but also the number of printed footnotes varies. The pictures also show, that descriptions of missing values are automatically generated.

## Datenbank mit statistischen Monats- bzw. Quartalszahlen

Tabelle mit den Ergebnissen Ihrer Auswahl aus dem Sachgebiet  
**Öffentliche Sozialleistungen**

Merkmal	Einheit	Region	1/96	2/96	3/96	4/96
Sozialhilfe der Bezirksämter - Abteilung Sozialwesen - ..Personen außerhalb von Anstalten ....Empfänger laufender Hilfen zum Lebensunterhalt .....Haushaltsvorstände <sup>1,2</sup>	Zahl	Berlin	117.415	124.359	129.569	141.437
Sozialhilfe der Bezirksämter - Abteilung Sozialwesen - ..Personen außerhalb von Anstalten ....Empfänger laufender Hilfen zum Lebensunterhalt .....Hilfempfänger insgesamt .....Veränderung gegenüber dem Vorjahreswert <sup>1,2</sup>	%	Berlin	x	x	x	x

Zeichenerklärung

x Tabellenfach gesperrt, weil Aussage nicht sinnvoll

Fußnoten

1 ohne Kriegsopferfürsorge

2 ab II Quartal 1994 ohne Zentrale Sozialhilfestelle für Asylbewerber

Picture 4: HTML-formatted result of retrieval • two variables selected

## Öffentliche Sozialleistungen

Merkmal	Einheit	Region	1/96	2/96	3/96	4/96
Sozialhilfe der Bezirksämter - Abteilung Sozialwesen - ..Personen außerhalb von Anstalten ....Empfänger laufender Hilfen zum Lebensunterhalt .....Haushaltsvorstände <sup>1,2</sup>	Zahl	Berlin	117.415	124.359	129.569	141.437
Sozialhilfe der Bezirksämter - Abteilung Sozialwesen - ..Personen außerhalb von Anstalten ....Empfänger laufender Hilfen zum Lebensunterhalt .....Hilfempfänger insgesamt .....Veränderung gegenüber dem Vorjahreswert <sup>1,2</sup>	%	Berlin	x	x	x	x
Sozialhilfe der Bezirksämter - Abteilung Sozialwesen - ..Personen außerhalb von Anstalten ....Aufwand insgesamt <sup>1,2,3</sup>	1000	Berlin	153.511	152.158	153.259	168.243
Sozialhilfe der Bezirksämter - Abteilung Sozialwesen - ..Personen außerhalb von Anstalten ....Aufwand insgesamt .....Veränderung gegenüber dem Vorjahreswert <sup>1,2,3</sup>	%	Berlin	x	x	x	x

Zeichenerklärung

x Tabellenfach gesperrt, weil Aussage nicht sinnvoll

Fußnoten

1 ohne Kriegsopferfürsorge

2 ab II Quartal 1994 ohne Zentrale Sozialhilfestelle für Asylbewerber

3 für einmalige und laufende Hilfen zum Lebensunterhalt sowie für Hilfen in besonderen Lebenslagen

Picture 5: HTML-formatted result of retrieval • four variables selected

## **Strategies for extension and integration of Meta-data**

There is no doubt, that Web-Sites of statistical offices should meet different needs of different users. Two main groups of users can be distinguished: Users with more analytical needs and users, that are more interested in ready made compilations. On the Web-Site of the Berlin statistical office one can find different kind of data offers. Ready-made compilations that provide an overview over recent figures and information (e.g. press releases, base-data and html versions of printed brochures; ready made, static html-files can be also generated automatically using the database) and Database front-ends to provide the possibility for a user, to retrieve data from time-series and for the different regions of Berlin among a bright variety of subjects. Besides those two main information pillars we also offer some back-ground information upon official statistics in Germany, the Berlin office, the legal base of our work (law on official statistics in Berlin, "Landesstatistikgesetz") etc.

However there are not enough meta-data available on the WebSite providing information like exact definitions of variables, description of surveys, methodological aspects etc. Current activities to improve this situation are oriented along three guidelines:

- Meta-data or meta-information shall be highly integrated with data. That means, all documents (press-releases, precompiled tables, database retrieval and -output etc.) have to be linked with meta-data, so that they are available for all relevant aspects exactly at the time, a user wants to get them ("only a mouse-click away").
- Visitors of the Site shall be informed about the existents of meta-data and meta data should be easy accessible, but users shall not be forced to read meta-information before they are allowed to get empirical data.
- Meta-information shall be also useful as a stand-alone information service; that means, visitors that have already the data they need (from what ever source) shall be able to get definitions, questionnaires, methodological information, comments etc. quite easy.

To meet all these requirements, meta-data can be organized in two different ways:

### **1) Meta-data organized in static HTML-files**

There exist a lot of regular publications containing various kind of meta-data: The statistical yearbook, monthly (quarterly or annual) reports for important statistics, questionnaires and many internal documents. Fortunately many of those are already in machine-readable form (indeed we are running the pre-print processes for many publications on Personal Computers since we introduced a DTP-System in 1992). So it is quite easy to convert these publications (or at least those parts of them containing meta-data) to Browser readable files. In many cases this is HTML itself, because the layout of these documents is not very complex and can be similar reproduced by the layout features of HTML. Questionnaires have normally a more complex layout. Therefore we are converting them into PDF-files. Based on these converted documents we are now generating a new section on our Web-Site, where one can investigate Meta-data. The structure of this metadata information section is made upon the various domains and subjects. For a quick search within these meta-data, we implemented an index server and specific search routines on our site. So users will be able to search for any topic, word or phrase among all this descriptive stuff. All these HTML-files are hyper linked and we will generate hyper links from all other static documents to the meta-data documents.

## **2) Meta-data organized under the DBMS**

Though it seems to be necessary, to make meta-data available in form of static html-files through which one can navigate via hyper-link references (not every user likes to use flexible query-techniques to get data), a more efficient way is to organize meta-data in relational databases. And in deed there is no reason, why meta-data could not be organized the same way, empirical data are.

For this purpose we are designing an expanded ERM for our database. An additional entity (database table), which contains the meta-data, has to be specified. It has a key for each meta-data entry, a key indicating what kind of meta-data the record contains and the meta-data itself. Another table contains for each meta-data key the variable key, it is relevant for. For we don't want to store blobs (binary large objects) in our database, we decided not to store pdf-formatted objects (e.g. questionnaires) themselves, but only a reference to the file, stored in the file system.

With this architecture we think, we will be very flexible, in presenting meta-data. For example the user will be able to decide, within a database retrieval, if and which kind of meta-data shall be included within the html-page presenting the resultset. He shall be able easily to request additional meta-data after he sees the resultset on his screen. So the user will be enabled to decide whether he wants meta-data included in the output page or he wants to get only hyper links, which enables him to decide later, which part of the information he wants to be presented.

We are convinced that at the end the strategy to organize all meta-data under the umbrella of a DBMS is the only practicable way to manage a "network of links" between data and meta-data. And such a network is the only architecture that enables users to get easily meta-data for data they already found, or to find appropriate data, they can only find via the meta-data. To meet all that kind of demands is the main reason, why we build up WebSites.

## **Monitoring the demand**

Running our WebSite we lay great emphasis on the permanent exploration of users demands. Therefore we are logging each request on our Site. E.g. regular http requests (gets, posts etc.) are logged by the httpd. To monitor the information on how many requests for pages, how many visits and visitors come to our Site, which navigation paths they are following etc. we installed and configured various kind of reports. One is automatically generated each day and another one once a week.

To observe exactly which series, variables, time points and regions are selected in our databases we developed some scripts, to store all that kind of information within the database. To examine this valuable information, we have written some specific reports too.

Based on all these statistical information upon the overall demand and demand profiles of our visitors we will also be able to examine the demand for meta-information. One information we have to catch is, what words, phrases and strings are users looking for, when they use the search facilities of our site, especially what are they looking for, that is not to be found.

Along with all these monitors we will be able to provide a demand driven site tomorrow instead of the supply driven perspective we are in nowadays. This is what we really will need, when we try to sell our data and information the day after tomorrow.