Hyperlinked Eurotrends

*by Lorenz Gräf**

Abstract
Hyperlinked Eurotrends are a new groundbreaking approach to present codebook data on the World Wide Web. The author created the so-called Eurobarometer Trends Codebook System (EUTRECS). Nearly all available material about Eurobarometers is presented in a clear and userfriendly way. EUTRECS gives fast and easy access to all metadata and datasets. It is the first comprehensive attempt to use internet technology to serve the most basic needs of secondary analysis researchers. EUTRECS is based on four principles: 1) Hyperlinks connect all interrelated variables; 2) A comprehensive search engine gives access to full text retrieval in all Eurobarometer Codebooks; 3) An index of over 1,000 keywords and a classification scheme of over 400 trend variables present an easy way to browse through Eurobarometer questions; 4) datasets and codebooks are available for immediate download. It will greatly enhance the dataservice of the archives. This paper describes the basic features of this system.

Scenario: What a researcher could think

Suppose you are a social scientist and you want to know which questions were asked in the Eurobarometer surveys. You have access to the Internet and you are looking for a database containing question wording of Eurobarometer questionnaires. After having visited all WWW search engines in vain, a friend of yours gives you a hint. You finally get through and a welcome page invites you: "Please type in your query". But how can you know what you are looking for, if you don't know what is in those Eurobarometer studies.

Let's assume, you know already what you are looking for. You are familiar with the topics of the Eurobarometer, but you have never worked directly with Eurobarometer data. Now you want to do some analysis about attitudes towards the common currency. You go to that question database, ask for "currency" and find the items you are looking for. But you are still without data. If your institution is a member of ICPSR, you are lucky and get the desired data quick via Internet. But otherwise you will have to wait until your request for data has reached your home archive and has been processed there. However, you need the data just at this very moment. It would be better for you if you knew someone who has these data already. Send him an e-mail and you will have your data within short time.

Let's think further. You have got the data and have run an analysis. But you are puzzled by your findings. You rerun the analysis, the findings remains the same. After a while you will be wondering if there is anything wrong with that question: "What was the exact wording of question Q34_a in EB 40.1?" Just two days ago you had received the printed codebook . But unfortunately it's in your office and you are on weekend some hundred miles away. So if you are lucky and you know a friend who has specialized in European politics and analysis of Eurobarometers, go to the telephone and give him a call. But otherwise?

Let's think positively. You have taken the codebook with you and resolved your question wording problem. Now you have a new idea: it could be very interesting to compare the attitudes towards the common currency between various professions and diverse levels of occupational prestige. You find a variable concerning "occupation" but you have to construct a prestige scale for that variable. After a few minutes you remember that there was an article concerning this issue some months ago. But who was the author and in which journal has it been published? It would be great if you could use his work and not having to redo the construction work for a second time.

The EUTRECS answer

The Eurobarometer Trend Codebook System gives a solution for all these problems. Those who are not familiar with Eurobarometers can browse in a keyword and trend index. They first scroll through subjects or concepts. If they find something interesting, they click on a link to that variable and have question text and marginals right in front them. Is that question asked more often than once they simply click on another link to follow that item over time. Browsing through codebooks that way, you may find an interesting keyword and you are wondering what else has been asked concerning that topic. So you follow the keyword link of that concept, finding yourself in the keyword index and having a list of questions with similar content in front of you.

Those researchers who are familiar with Eurobarometer studies find a fully searchable codebook database. Every single word contained in a codebook can be found. All the internal linking between similar variables has been preserved. From within EUTRECS users can download every dataset except those which are currently under embargo. This service is open only for non-commercial use and academic purpose. You have to register and accept online the terms of use agreement. Once registered researchers get an username and a password. After having selected one or more datasets they get emailed a transaction id. Within minutes you can have the data. The hyperlinked Eurobarometer web pages provides space for user communication. In the moment there is a list of nearly 800 working papers all about Eurobarometers. If users contribute to that pages there will be much more content in the next future.

EUTRECS in detail

Overview

Eurobarometer are surveys conducted on behalf of the European Commission in the member countries of the EC. They focus on topics concerning the EU. Normally they are conducted in spring and fall every year. Sample size in each country is about 1,000 persons. Data are gathered using face to face interviews. Besides this standard series, there are several so-called flash Eurobarometers. They are conducted by telephone, on a smaller sample size (about 500) and focus mainly on one topic. Since the breakdown of the former soviet imperium every year a survey similar to the standard Eurobarometer is carried out in the

Central and Eastern European countries. Two years after the fieldwork the data is normally available for academic researchers. Those who want the data earlier need a special permission of the European Commission. The series of Eurobarometer surveys started in the early seventies. Up to now, there have been conducted more than 70 standard, 50 flash surveys and 10 Central and Eastern European surveys.

The Eurobarometer site built for the Central Archive (ZA) is based upon the codebooks created by ICPSR, SSD and ZA itself. So far the codebooks are only produced for standard surveys which do not underlie the embargo restrictions. Background information about survey characteristics and topics covered is available for all Eurobarometer surveys. Datasets of all type of Eurobarometer surveys (standard Eurobarometer, flash Eurobarometer and Central and Eastern Eurobarometer) can be downloaded, but this service is restricted to registered users. EUTRECS consist of 32,000 HTML-pages, comprises 1,000 keywords and over 400 trend variables. It is built on basic usability principles and aims to encompass user needs.

Hyperlinked structure

The Eurobarometer codebooks

make use of the core web technology. Hyperlinks between variables make it easy for the user to explore survey content and to follow his associations meanwhile browsing through the hyper codebook. What does this mean? Suppose you are
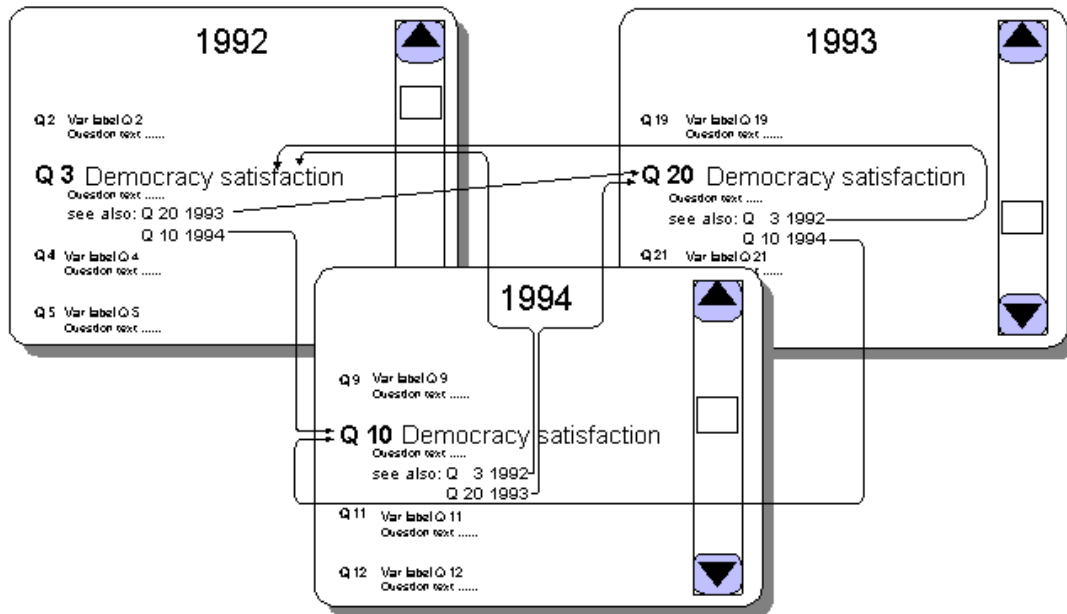
interested in attitudes towards *technology*. You look for the item *technology*. Some of the items you find deal also with *computer*. In EUTRECS you can follow the link underneath the item *computer* and find a list of all variables containing this term. By browsing these variables you discover a variable containing information about the diffusion of computer in diverse European countries. You can see that at the beginning of the nineties in the Netherlands and Great-Britain computers were
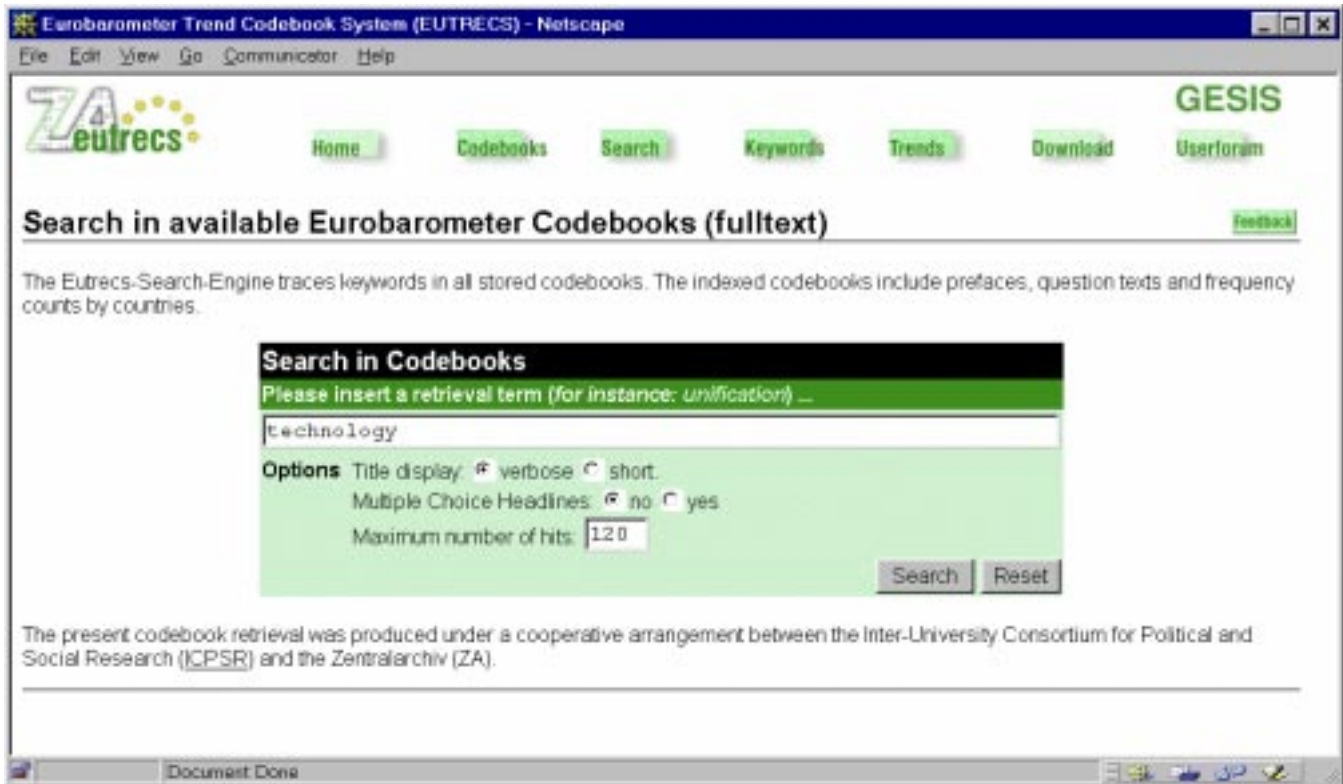
present in a third of all households. At the same time in Germany and France only every fifth household possessed a computer. Now you are interested to know how these figures have changed over time. In other information systems on the web you have to go back to your query result and click on the next item. So if you want to follow a fairly large trend you are forced to go forward and backward some nasty long time. EUTRECS shows on every variable output all other variables which contain the same trend. Therefore it is easy to follow the same variable over time by only clicking on the variable name.

EUTRECS is based on the already available ASCII codebooks. When creating HTML pages EUTRECS needs a continuity table to link identical or nearly identical questions. This table is built in two ways. First, variables with identical labels will be identified and inserted in the table. The second way consists in preparing this table outside of EUTRECS and using the internal codebook information to check the applicability of the pre-given information. As shown in the next figure EUTRECS enlarges the ASCII codebooks and adds links.

Fulltext search engine
EUTRECS has a built-in search engine based on Freewais-SF. All relevant codebook content can be searched using a fully featured search engine. It is the same engine that serves the CESSDA database. What is unique to EUTRECS is that it preserves all the internal links during processing in the search engine. So, if a user puts a query to locate all variables containing the term *'technology'* on the display of results he finds links to similar concepts, links to other variables of the same trend and a link to the complete online codebook enabling him to explore the context of this variable in the original questionnaire.

Keyword and trend index A web based information system should present as much information as possible in html files. Following this way, researchers can use the web site like a book. In EUTRECS browsing in codebooks is possible.
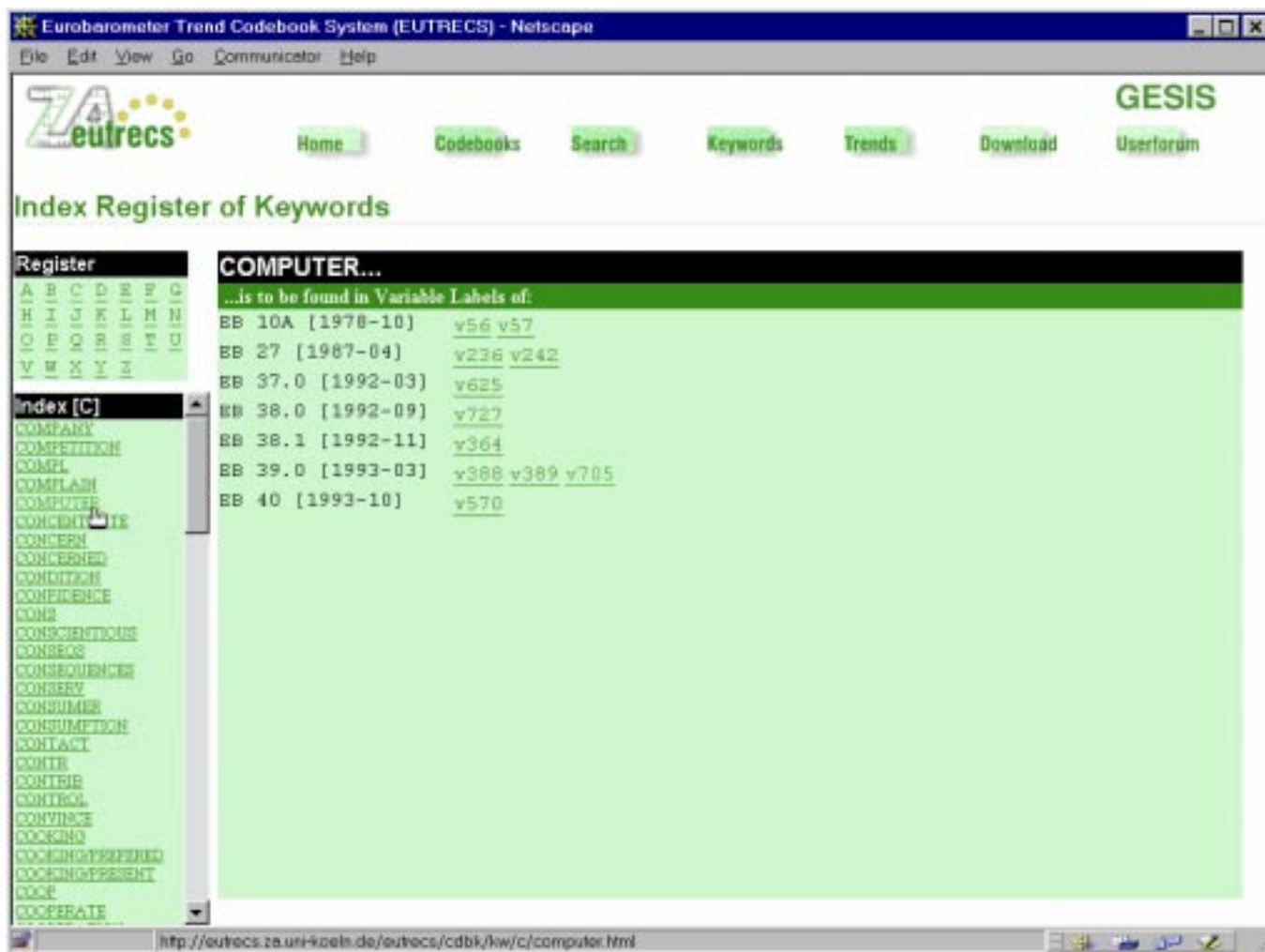
Comparing EUTRECS to books, the single online codebooks are the chapter of the book. Like in books, identical or similar concepts are dealt with in different chapters. To facilitate navigation, EUTRECS provides not only a free text search engine but also a keyword and a trend index. The keyword index groups variables with similar subjects together. The trend index combines identical or nearly identical questions in different surveys.

Keywords are extracted from variable labels. Doing so EUTRECS can take advantage of a quasi-controlled vocabulary and is not bound to the question wording. But it heavily depends on the quality of the variable label wording. In Eurobarometers this labeling should be improved. Only codebooks produced in the last years contain really good labels but standardization is already on the way. It would be better if we had real descriptors for each variable. But for the time being we have to be content with the material as it is. From the variable labels of the available codebooks we extracted over 8,000 tokens. As a first step1 we implemented a stop-word list and ended up with about 1.000 different keywords. EUTRECS present the keywords in alphabetical order entirely on one screen. You can see in the following figure how easily and fast navigation is. With only three clicks the user gets to the desired variable. A comprehensive list of all keywords is available but it is nearly 2 MB big.

A series of thematically similar surveys offers great opportunities for time oriented secondary analysis. To foster such analysis researchers ought to know which question was asked more than once and in which surveys. The EUTRECS trend index gives the most comprehensive trend register of the Eurobarometer surveys. It contains trend information identified at ZEUS (Mannheim), and at ZA (Köln). But unlike ZEUS we defined trends strictly on a single variable level. In our definition a trend is every single variable that can be traced over time. It is this data that can be compared between different points in time. Now what is the difference? At ZEUS trends are identified, named and differentiated on a concept level. In EUTRECS only questions with identical wording or identical subject have got the same name. Let's give an example. ZEUS presents each item of the question "Which of the following aereas of policy do you think should be decided by the (NATIONAL)
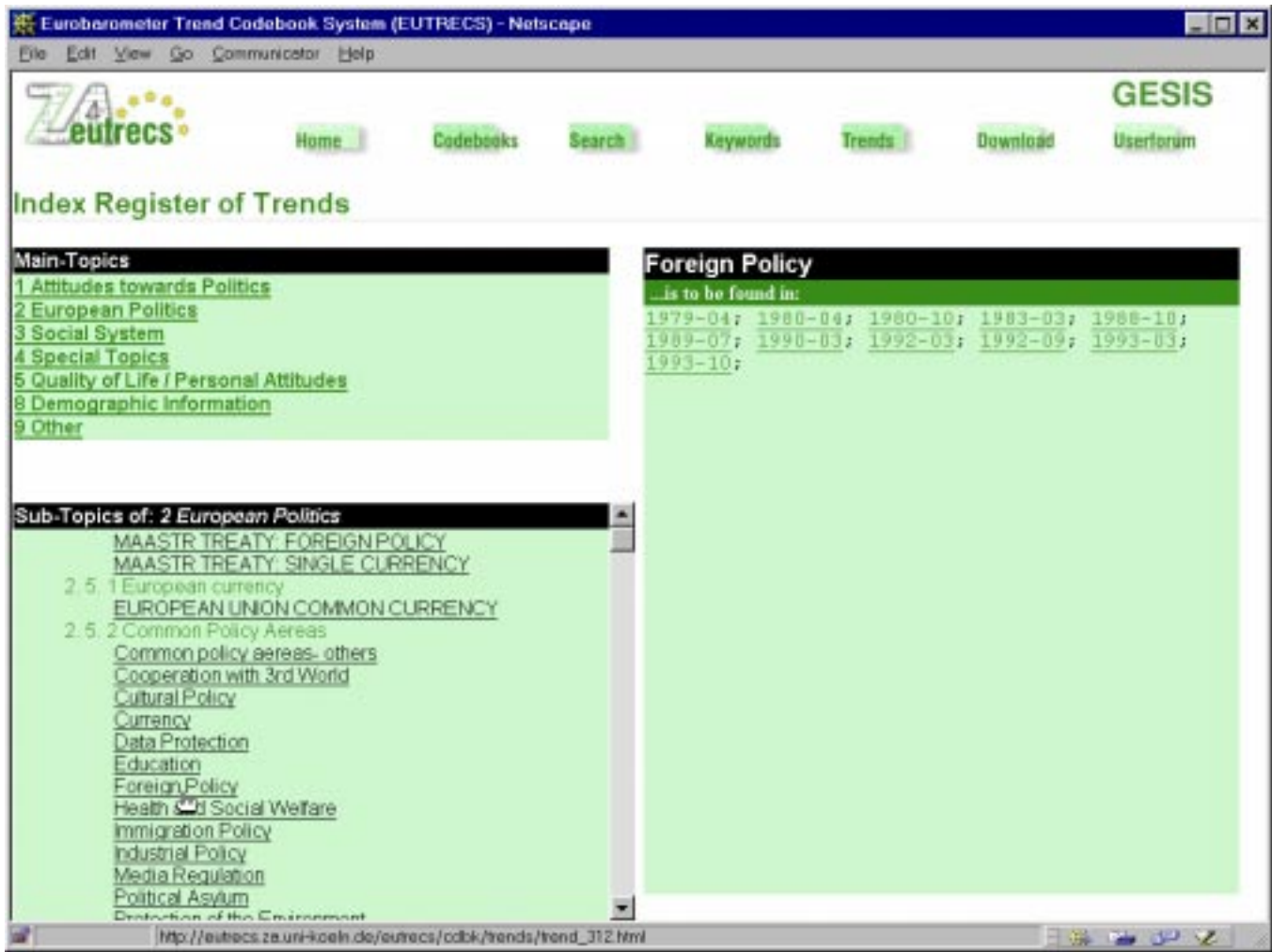
government, and which should be decided jointly within the European Community?"2 under the label 'Common policy aereas'. Whereas in EUTRECS we name the concept 'Common policy aereas' but give the trends the name of the stimulus object, i.e. 'foreign policy', 'currency' or 'education'. By doing so we can provide links between all variables which have the same question wording. Clicking on these links will guide the user through all instances of the trend. As EUTRECS is built out of machine readable codebooks together with each question text full marginal information is displayed. In this version we were not strict on question wording and did not differentiate questions with minor deviations in question wordings. We leave it up to the researcher to decide whether questions were identical or not.

Until now we have identified over 400 single trend variables. To facilitate finding topics we classified the trends in a new classification scheme. In our opinion it was the best method to group similar topics together and avoid getting lost in a mere alphabetical ordering. This classification scheme includes three levels. The third level describes the concept level and is comparable to ZEUS trend names. Beneath that level we have classified the variable trends which we could identify. So we can distinguish between the concept and a measurement level. In the following figure you can see how trends are presented in EUTRECS.

Download
Within EUTRECS the direct download of SPSS datasets and of codebook material is possible. Bandwidth on the Internet is normally small. Therefore we divided the big codebooks in smaller parts of about ten or fifteen variables. But for printing purposes many users want to have entire codebooks. For offline use we give access to all codebooks in html- and postscript format. Keyword and trend indices can be accessed in one-file lists.

Access to datasets is given via an entirely online procedure. We tried to find the easiest and fastest way to deliver data over the Internet without having to compromise to the vital interests of the archives. So we ended up with a three step procedure. In the first step researchers register at the ZA in Cologne. All they need is a functional e-mail address. After registration they get an account and a password via e-mail. Provided with username and password in step 2 they can select datasets. After having accepted the *terms of use agreement* and having paid a fee (depending on the usage conditions of the archive) they get a transaction id by e-mail. With this id they go directly to the download page and get their material immediately.

User forum
The web site we intend to create should be as interactive as possible. It should be the most efficient site for all researchers

doing secondary research with Eurobarometer surveys. We pursue three aims: First, this site should be a platform for the communication between users and the archive and between users themselves. It would be nice, if the archives use this forum to announce any news concerning Eurobarometers (bug reports, announcement of availability and declarations about archive policy). Second, we try to stimulate a shared knowledge forum. In this forum every researcher should announce his findings in the Eurobarometer data. The site now contains a list of working papers done with Eurobarometer material. We would appreciate if researchers uploaded their newest paper for the communication with the scientific community. And we encourage researchers to share their operationalisations and measurement attempts of the data. It would be a really common good to have samples of SPSS or SAS statements there for often needed recoding of Eurobarometer variables. Third, we will give EUTRECS users the necessary software tools to use materials of this site, like ghostview to read ps-files.

What comes next
EUTRECS is going to be developed in a multistage process. In stage one it was important to explore possibilities to present large and complex survey metadata on the web. In the next step we need the help and feed-back of the users to add consistency to the pages. We have to find and correct miscoding, misleading labeling and false classification. It has to be tested how usable the site is and how navigation could be improved. Also the gap in the codebook production should be closed as soon as possible. In the last step we want to incorporate in the data itself trend information and other measurement suggestions of the research community.

1 In future versions a better suited dictionary will be used to improve building of keywords.

2 Exact question wording (EB33 Q30): ”Some people believe that certain areas of policy should be decided by the (NATIONAL) government, while other areas of policy should be decided jointly within the European Community. Which of the following areas of policy do you think should be decided by the (NATIONAL) government, and which should be decided jointly within the European Community?”

- Foreign policy towards countries outside the European Community

- Education

WWW: http://infohttpsoc.uni-koeln.de/graef http://solix.wiso.uni-koeln.de/~graef/

* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

Lorenz Gräf, University of Cologne, E-Mail: lorenz.graef@uni-koeln.de