STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 10
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Geneva, Switzerland, 22-24 September 1999)

Topic (ii): Responsibility for the management, control and nurturing of statistical metadata

PROVIDING GLOBAL ACCESS TO DISTRIBUTED DATA THROUGH METADATA STANDARDISATION – THE PARALLEL STORIES OF NESSTAR AND THE DDI

Submitted by The Norwegian Social Science Data Services[1]

Invited paper

I. INTRODUCTION

1. The paper tells the parallel and highly interlinked stories of two initiatives originating in the world of social science computing, the DDI and NESSTAR. The first of these acronyms represents a concerted effort among data archives and data providers in USA, Canada and Europe to develop a metadata standard for social science data resources (the Data Documentation Initiative). The second belongs to a European-based software development project that uses this emerging standard as a platform to provide on-line access to huge amounts of distributed data over the Internet (Networked Social Science Tools and Resources).

2. Despite the fact that the ideas were conceived more or less at the same time and by persons and organisations belonging to the same international community, the two projects lived their infant lives in relative isolation. However, as the contact and affinity grew, the dependencies between the ventures became increasingly evident. For a software project aiming at seamless integration between a broad range of locally controlled data holdings, the need for a generally accepted metadata standard is obvious. That metadata standards need software support to reach general acceptance might be less evident but is nonetheless true. Without software tools that can prove its usefulness and efficiency, any standard is doomed to die.

3. To understand the rationale behind and the direction of the DDI and NESSTAR projects, let us start with a short description of the environment from which they both originated.

II. SOCIAL SCIENCE DATA ARCHIVES

4. Within the academic sector social science data archives and data libraries have been established to provide researchers and students with data for secondary analysis. Some of these institutions have been in existence for 2-3 decades and house the largest collections of accessible computer-readable data in the social sciences in their respective countries. The primary goals of the archives and libraries have been to safeguard the data and to make them as easily accessible as possible for teaching and research independent of whether the users are able to pay for the services or not.

5. The social science data archives are rarely engaged in the collection of primary data, but serve as brokers between various data providers and the academic community. Their holdings contain data from

[1] Prepared by Jostein Ryssevik.

the public sector (statistical agencies, central government etc), the commercial sector (opinion and market research companies) and academic research. The archives do not only preserve data for future use but also add their own value to the collections:

- data received by the archives goes through a variety of checks and cleaning procedures to ensure their integrity;
- any system or software dependency is stripped away to make sure that data can be read at any time in the future;
- comprehensive computer-readable metadata are developed;
- data from various sources are often integrated and harmonised in order to produce easy-to-use information products (on-line databases, CD-ROMs etc.);
- data are catalogued and made accessible through electronic search and retrieval systems;
- in order to encourage the use of statistical data among students, teaching packages and interactive statistical laboratories are developed.

6. Due to the extensive refinements of the data sources, as well as a long-standing reputation of responsiveness to users' needs, data and related services from the archives are frequently requested by non-academic users. This includes users from the public sector, as well as from the mass media and private companies. To the extent that services to non-academic users do not run counter to the agreements with the data depositors, access is usually granted.

III. TOWARDS A METADATA STANDARD FOR SOCIAL SCIENCE DATA RESOURCES

7. The characteristics of the user communities go a long way to explaining the high priority that the archives have given to the development of metadata:

- users of archived data have rarely been engaged in the creation of a dataset;
- archived data will frequently be used for research purposes other than those intended by the creators (secondary analysis);
- archived data will frequently be used many years after they were created;
- academic users often compare and combine data from a broad range of sources (across time and space).

8. The common denominator of the four characteristics is an emphasis on the relative distance between the end-users of a statistical material and the production process. Whereas the creators and primary users of statistics might possess "undocumented" and informal knowledge, which will guide them in the analysis process, secondary users must rely on the amount of formal meta-data that travels along with the data in order to exploit their full potential. For this reason it might be said that social science data are only made accessible through their metadata. Without human language description of their various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end-users. The metadata provide the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

9. The metadata is also the pivotal point for any resource discovery system (paper-based as well as digital). Academic users will frequently shop for the most relevant data that might be used to shed light on a topic, substantiate a theory, test a hypothesis, etc. Although basic catalogue information might provide information about the overall content of a data-source, detailed metadata on source- as well as variable-level are needed to increase the precision of the resource discovery process. This includes detailed information about concepts and definitions, methodologies and procedures, exact and complete question text (for survey-based studies), relationships to other sources and studies, etc.

10. The conveying of the meaning of data through links to the production process and the facilitation of efficient and high-precision resource discovery can consequently be seen as the two most important reasons for the social science data archives to engage in metadata development.

11. A third reason, which also should be mentioned, is the potential role that metadata might play as a bridge between the data, the users and the intellectual production of the users. By embedding references and hyperlinks to the reports and scientific studies written on the basis of a dataset, as well as references and links to the researchers and institutes that have been responsible for the research, the metadata might become an important communication node and a vehicle in the process of knowledge accumulation. Too often the dataset or the table report is seen as the terminal station of the statistical production line. By including the (secondary) research activities in this process, an extension of the metadata concept along the lines described above, seems quite reasonable. This argument will be developed further in the concluding section of the paper.

12. Over the years many initiatives have been taken within the data archive movement to create metadata standards. None of these have, however, reached the level of acceptance that is needed for a standard to be successful. The majority of social science data archives have documented their holdings according to a standard study description agreed in the mid 1970's by an international committee of data archivists. Unfortunately many local "dialects" of this standard have evolved and the archives have adapted their metadata holdings to fit the requirements of different storage and retrieval systems. As a consequence the level of standardisation across archives is rather low.

13. In order to improve this situation, a new international committee, the Data Documentation Initiative (DDI) was established in 1995 to create a universally supported metadata standard for the social science community. The committee was initiated and organised by the Inter-University Consortium for Political and Social Research (ICPSR). The members were coming from social science data archives and libraries in USA, Canada and Europe and from major producers of statistical data (like the US Bureau of the Census, the US Bureau of Labour statistics, Statistics Canada and Health Canada). Information about the work of the DDI-committee can be found at: http://www.icpsr.umich.edu/DDI/.

14. The original aim of the DDI was to replace the old-fashioned and obsolete standard study description with a more modern and Web-aware format. The first version of the new standard was consequently expressed as an SGML DTD. In 1997 it was translated to XML (Nielsen, 1997) where it have stayed since. This was just a few months after the World Wide Web Consortium (W3C) released the very first working draft for this new language which according to the visions of the creators would add a new dimension to Web-publishing, especially related to resource discovery and metadata.

15. In the period between March and August 1999 the proposed DDI-standard has been subjected to intensive beta-testing by thirteen organisations from both sides of the Atlantic. The beta-test activities include production of marked-up metadata for different types of data, development of software that can assist the mark-up process or produce the new standard from existing sources, comparisons of the DDI-standard with other relevant documentation standards etc. The recommendations, documents and software from the beta-test process can be found at: http://www.icpsr.umich.edu/DDI/codebook/testers.html

16. A Working Group is currently assessing the various proposals that resulted from the beta-test process. The final decisions concerning the content and structure of the standard will be taken at a DDI Committee Meeting in Ann Arbour; Michigan in the middle of October. If agreements can be reached, Version 1.0 of the DDI-standard will eventually be published by the end of 1999. Accompanying the standard will be a tag library explaining the use of each single element in the tag-hierarchy. A preliminary version of this tag library can be found at: http://www.icpsr.umich.edu/DDI/codebook/codedtd.html.

17. An XML DTD (Document Type Definition) provides the rules for applying XML to a document of a specific type. The DTD defines the elements that the document is composed of, the attributes of these elements, and their logical relationships to other elements. The elements will usually be arranged in a hierarchical or tree-like structure. The DDI-tree contains five main branches, or sections:

- **The document description**, which describes the metadata document and the sources that have been used to create it (this section can thus be looked upon as a kind of metadata for the metadata , or meta-metadata if you like).
- **The study description**, which contains information about the entire study or data collection (content, collection methods, processing, versioning, sources, access conditions etc).
- **The file description**, which describes each single file of a data collection (formats, dimensions, processing information, missing data information etc.)
- **The variable description**, which describes each single variable in a datafile (format, variable and value labels, definitions, question texts, imputations etc.)
- **Other Study-Related Materials**, which can include references to reports and publications, other machine readable documentation that is relevant to the users of the study (referenced by URI's) etc.

18. Each of these main branches is divided into a finer hierarchy of sub-branches. A graphical description of parts of the methodology branch, which is a sub-branch of the study description branch is shown below:

2.3 method* (ATT == ID, xml:lang, source)
---- 2.3.1 dataColl? (ATT == ID, xml:lang, source)
---- 2.3.1.1 timeMeth* (ATT == ID, xml:lang, source, method)
---- 2.3.1.2 frequenc* (ATT == ID, xml:lang, source, freq)
---- 2.3.1.3 sampProc* (ATT == ID, xml:lang, source)
---- 2.3.1.4 deviat* (ATT == ID, xml:lang, source)
---- 2.3.1.5 collMode* (ATT == ID, xml:lang, source)
---- 2.3.1.6 resInstru* (ATT == ID, xml:lang, source, type)
---- 2.3.1.7 sources? (ATT == ID, xml:lang, source)

19. As an example of the use of the DDI-standard, the timeMeth element (2.3.1.1 in graph above) is supposed to contain a description of the time method or time dimension of the data collection. The element text can be used to give a human language description of the method, whereas the method-attribute can include a controlled vocabulary, which more easily can be understood by a software system. An example of how this structure can be used in concrete mark-up is shown below:

<method><dataColl><timeMeth method='panel'>The study is a panel survey where 50% of the sample are replaced at each subsequent.....etc. </timeMeth></dataColl></method>

20. The example demonstrates one of the basic strengths of XML – the ability to bridge the gap between human and machine readability. The language itself does, however, not guarantee that a marked-up document can be read and understood as easily by a digital process as by a human being. In order to achieve this, the structure of the document (the DTD) must be designed with this dual aim in mind.

21. Another positive trait of XML DTDs is the ability to specify content as well as syntax or format. Whereas the existence of an element informs the metadata author about the type of information that is expected to be supplied, the element- and attribute structure as well as the basic rules of the XML-language proscribes a format.

22. In sum, the choice of XML as a platform for the DDI standard seems to have been the right one. Since the freezing of the XML specification in February 1998, we have witnessed an explosive growth of web-oriented XML technology. This includes everything from XML-editors and tools for publishing of XML-documents, to XML data servers, XML search engines and XML databases. Even more important, an increasing number of organisations, communities, domains and sciences are looking towards XML as a platform for domain-specific mark-up as well as vendor-neutral data exchange. The effort to move the not too successful UN/EDIFACT family of standards for electronic data interchange to an XML-based

platform is a prominent example of this trend. Information about the European XML/EDI Pilot Project can be found at http://www.cenorm.be/isss/workshop/ec/xmledi/isss-xml.html.

23. A number of important add-ons and extensions to the XML-platform is also published or in the pipeline for publication. Among these are:

- **XSL** (eXtensible Style Language) – a stylesheet language for the display of XML-documents on the Web.
- **XLink** (the XML Linking Language) – that allows efficient and advanced hyperlinking across XML-resources.
- **XFDL** (Extensible Forms Description Language) – a language for digital representation of complex forms.
- **XQL** (XML Query Language) – one of several proposal for an XML-based parallel to SQL

24. Concerning metadata standards, the most interesting extension to the XML-language is undoubtedly RDF (the Resource Description Framework). RDF is an application of XML that provides the foundation for metadata interoperability across different resource description communities. This is achieved through the publication of domain specific metadata vocabularies that enables the semantics of objects to be expressible as well as exploitable. The utilisation of this framework will be a natural next step for the DDI-standard. Further information about RDF can be found at: http://www.w3.org/RDF/.

25. The take-up of the proposed DDI-standard among the community of data archives and libraries has so far exceeded the expectations. At an increasing number of sites from all parts of the World efforts are being made to convert existing holdings of metadata to the new standard. The work of the DDI have also served as an instrument to revitalise the co-operation and sharing of know-how among the archives, as well as strengthening the ties to the data producers. The process of developing software that supports the new standard is also well under way.

26. Before we return to a discussion of the shortcomings and future directions of the DDI-standard, the story of the most ambitious of these software projects, NESSTAR, ought to be told.

IV. TOWARDS THE "SOCIAL SCIENCE DREAM MACHINE"

27. What if a representative group of users of statistical information (researchers, students, planners, decision-makers) where given a chance to describe their "dream machine" – the technological environment which could serve as an enlarged, intimate supplement to their own brain, enhancing their productivity as well as the quality of their work. We are confident that their combined list of requirements would include at least the following items:

- all existing empirical data available on-line
- an integrated resource discovery gateway and search-system that could help to identify and locate these resources
- extensive amounts of metadata available (multimedia, hyperlinked and totally integrated with the data as such)
- the ability to browse, analyse and visualise data on-line
- the ability to convert the data in one of a number of formats and copy, with the metadata, to a local machine
- "active research agents" (knowbots) mining the net and informing the user when new data within their special field of interest are made available
- efficient hyperlinks from the data sources to every scientific publication ever produced on the basis of a dataset
- ditto e-mail/web addresses to all relevant researchers, departments etc.
- an efficient feedback system to the body of metadata allowing the user to add to the collective memory of a dataset

28. The aim of the NESSTAR project has been to realise as many of these dreams as possible. The project was initiated by the Web-revolution and a drive to develop a common Internet-based gateway to the various data holdings of the European data archives (the archives belonging to the Council of European Social Science Data Archives - CESSDA). Three steps on this route can be identified:

The CESSDA map: A common Web-page including a "clickable" map with hyperlinks to the Web-sites of the various archives (1994). The map, which also include a section on data libraries in North America can still be found at http://www.nsd.uib.no/cessda/europe.html

The CESSDA IDC: An integrated on-line data catalogue based on WAIS-sf technology (1995). The integrated Data Catalogue can be viewed at http://www.nsd.uib.no/cessda/IDC/

NESSTAR: A seamlessly integrated data discovery, analysis and dissemination system based on Java, XML and the DDI (1997-99).

29. Whereas the first of these steps might be seen as a simple portal to separate and non-standardised services, the second offers an integrated interface to semi-standardised catalogues. The ongoing NESSTAR project is taking this to a more advanced stage by linking the fully standardised XML metadata as well as the data to the system. As a consequence, users of NESSTAR will be able to:

- locate multiple data sources across the holdings of several data repositories
- browse detailed metadata describing these data
- analyse and visualise the data online, and
- download the appropriate subsets of data in one of a number of formats

30. NESSTAR is a joint development project between the Norwegian Social Science Data Services (NSD), UK Data Archive and the Danish Data Archive (DDA). In addition to the three main contractors, several associate contractors provide the links to the end users (academics and journalists) and the data producers (official statistics and private survey organisations). The project is funded by the Telematics Applications Programme an activity under DGXIII of the European Commission's 4th Framework Programme. The main project has run for two years and will end January 2000. More information about the NESSTAR-project can be found at: http://www.nesstar.org.

31. NESSTAR is building on a distributed model where data are stored and maintained at separate units across the World. In this way the data providers and archives are guaranteed maximum control over their own holdings. However, for the end users the various collections will appear as elements in a totally integrated data archive. As a consequence NESSTAR might be described as a virtual data library offering global access to locally supported holdings.

32. The resource discovery system of NESSTAR is metadata-driven. Given the detailed structure of the DDI DTD, the user will be allowed to search for data with a very high precision, avoiding the overflow of hits that usually is the case when searching the Web. Researchers that are interested in particular subjects will be able to move beyond the keywords and abstracts that normally are included in on-line catalogues and search directly on variable descriptions, question texts etc. Likewise searches can be conducted on concepts such as method of data collection (e.g. telephone interviews, face-to-face interview or self-completion questionnaires) or sampling strategy (e.g. random, stratified, etc). Searching across archives is in addition fully seamless in the NESSTAR system.

33. The search-technology in NESSTAR is currently based on the Cheshire search engine that is developed by Professor Ray Larson at the University of Berkley. This is a very powerful engine originally developed for advanced searching in SGML repositories. More information about the Cheshire system can be found at: http://cheshire.lib.berkeley.edu/

34. Although many data providers and data archives are producing high quality metadata, the integration between metadata and the data as such are usually less than satisfactory. The bulk of metadata

are often stored in separate systems and must be viewed by other software devices than the data. This is partly due to the fact that the formats of most statistical packages do not support inclusion of metadata (beyond simple variable and value descriptions). NESSTAR is trying to bridge this gap between metadata viewing and data analysis. Metadata in NESSTAR are, therefore, not only used for resource discovery. As soon as a dataset is located, the accompanying XML-formatted metadata travel the net and can be viewed in a combined metadata/data browser. In this tool the user can easily jump from full text descriptions of variables to statistical analysis. When analysing data, the relevant parts of the metadata will never be more that one keystroke away.

35. The on-line data browser in NESSTAR includes basic statistical methods like n-way crosstabs, breakdown analysis, correlation and regression. For every statistical method, relevant graphical visualisation methods are available. The data browser is designed for two purposes:

- For the advanced user, the browser will mainly be used to decide whether or not a particular dataset fits a research purpose. If data can be used, the advanced user will probably download the data in an appropriate format and continue the analysis in his/her own favourite statistical package. In situations where the user decides to continue the analysis locally, the dataset can easily be sub-setted by variable and/or cases to create manageable selections or simply to save bandwidth.
- For the more inexperienced user, the methods that are available in the NESSTAR browser might be sufficient. In this case downloading of data for local analysis will not be needed.

36. The remote analysis system is designed to minimise net traffic without reducing client-side interactivity - a dilemma that is well known to all designers of distributed information systems for the Web. The basic idea is to divide the production of statistical output in two processes: the operations that are calculated on the raw data (and consequently need access to these data) and the post hoc manipulations on these calculations (like making percentages in a cross-tab, turning numerical information into graphical displays etc.). By dividing the processes between the client and the server in this way, we are able to create a highly interactive system with response times that are more or less equal to desktop statistical packages.

37. The NESSTAR hyperlink- and bookmark-facilities are integrating the system fully into the Web environment. NESSTAR will provide a framework for bringing live data into on-line texts, as well as a framework for linking on-line scientific texts into the metadata body of a data material. The former is achieved by a naming convention, which makes it possible to name, bookmark and hyperlink all relevant resources in the NESSTAR repositories. The latter is achieved by traditional hyperlinks in the XML-formatted metadata.

38. When a link to NESSTAR data is activated from an on-line text in an ordinary Web-browser, the client will be fired up to display the relevant information. Correspondingly, when a link to external documents (or any external web-object) is activated from the NESSTAR metadata browser, external viewers like Netscape or Acrobat will be fired up to display the material.

39. In theory every bit and piece in the data and metadata parts of the NESSTAR repositories can be named and hyperlinked, but for practical reasons we are limiting the implementing to the following:

Searches: The search strings that is the result of a search for data in one or several NESSTAR archives. When activated, the link will return a hit-list which will take the user directly to the data. Note that this is a “dynamic” link that will return different results whenever a new dataset that meets the requirements is added to the repositories.

Dataset: The location and name of a particular dataset. When activated, the metadata of the dataset will be loaded into the NESSTAR client and the data will be available for on-line analysis or download.

Analysis: The request for an analysis (like a crosstab) on a particular dataset. When activated the metadata of the dataset will be loaded and the requested analysis performed and displayed in the NESSTAR data explorer. Data will be available for further analysis or download.

Download: The specification of a download including sub-setting information and output-formats. When activated, the output file will be created and delivered to the specified destination.

40. As in ordinary Web-browser hyperlinks can also be bookmarked. The NESSTAR client includes standard bookmark functionality which can be used to create short-cuts to favourite datasets or tables and graphs that the user wants to be able to reproduce in an easy way.

41. A user of NESSTAR that is searching for data within a special research area, will have the opportunity to leave the search criteria in the system and be automatically informed by e-mail whenever new data within the special field of interest are made available somewhere around the world. This is a very simple implementation of the active agent or knowbot technology and similar to the type of services that are currently offered by electronic bookstores like Amazon (http://www.amazon.com).

42. In NESSTAR the active agent functionality can be regarded as an extension of the hyperlink/bookmark concept. An entry to the active agent is simply a "subscribed bookmark" that is automatically activated on behalf of the user on regular intervals. A consequence of this is that it is not only searching that can be performed by the knowbot. The user will also be able to subscribe to particular tables (analysis) or downloads and let the knowbot do the work. This feature will of course only be relevant for highly dynamic data that is updated frequently or on a regular basis.

43. The NESSTAR system has attracted a lot of interest from various groups of data distributors as well as data producers. A beta version of the system has been available since May 99 and test installations are currently being set up at various data archives and libraries in Europe, USA and Canada. From October 99 a public beta version of the client will be made available on the Web to allow extensive end-user testing. This is the final stage of a usability program that started with testing of throw-away prototypes almost two years ago and have involved a large number of users as well as a variety of methods.

V. INTERDEPENDENCIES

44. The NESSTAR system is building upon the DDI DTD. It might even be regarded as a software implementation of the standard. Given the ambitions of the venture it is difficult to see how the system could have been built without a widely accepted, highly structured and over-all Web-friendly metadata standard like the DDI.

45. The NESSTAR system is metadata-driven:

- The internal run-time files of the system (used for on-line analysis) are produced directly from the DDI-metadata.
- As much as possible of the DDI-metadata will also be included in any export-files produced by the system (depending on the metadata-friendliness of various output formats). NESSTAR can also produce HTML-versions of the complete metadata, which might accompany the export of data.
- The complete and indexed DDI-files are the basis for the resource discovery system of NESSTAR.
- The XML-trees created from the DDI-metadata provide the maps that are used to navigate datafiles
- The DDI-metadata provides the information used to guide the user when weighting the data

- The DDI-metadata will (hopefully) provide the information for the automated access control system.

46. Without a standard that is widely accepted among the data distributors, the integration across local metadata repositories will have to be taken care of by software. A promising example of this "software-approach" is the ComeIn project carried out by Space-Time Research (Australia) and run Software (Germany) in close co-operation with several statistical offices. The aim of ComeIn is to develop a general metadata interface that can serve as an integration layer between local metadata repositories and different user tools (resource discovery tools, data dissemination tools, tabulation tools etc.). In order to provide a maximum level of portability across systems and platforms, the ComeIn interface will according to the plans be accessible through CORBA, COM, as well as XML.

47. Another project that promises to access DDI-data as well as several other metadata standards or systems in an integrated manner is the Virtual Data Centre currently under development at Harvard-MIT (King 1998). More information about the Harvard-MIT approach can be found at: http://thedata.org/.

48. Under this heading we should also mention the work carried out by the Analytical Data Management DSIG (formerly Statistics Domain Special Interest Group) of the OMG. According to the mission statement the aim of this group is "to prepare to use the Object Management Group (OMG) technology adoption process to standardise the interfaces for software tools, services, frameworks, and components in statistical data collection and dissemination". For more information about the work of the OMG Interest Group, see http://www.omg.org/techprocess/sigs.html#statistics). We regard the development of a general object model for statistical metadata to be a long-term goal that if realised and universally accepted will make the life much easier for any developer of software aiming at integration across local systems. In the absence of such a model, standardisation efforts like the one accomplished by the DDI, is needed for systems like NESSTAR to be technically feasible.

49. However, not only is the NESSTAR system dependent on the DDI. The DDI standard is also dependent upon software systems like NESSTAR to prove its usefulness. No organisations are willing to invest the resources needed to change internal routines and to convert existing metadata to a new standard, without software systems that can demonstrate that the investments are worthwhile. Without productivity gains or improvements in the quality of products or services, acceptance of a new standard might be hard to justify.

50. The history of the Web itself might provide us with the most prominent illustration of the mutual dependency between the acceptance of a standard and software development. Without Tim Berners-Lee's first version of HTML, Marc Andersen and his group at NCSA would not have made a fortune on developing Mosaic and eventually Netscape. Likewise, without the development of the first Web-browsers (like Mosaic and Netscape), HTML would probably have remained a local hypertext dialect for technical documentation at CERN, Switzerland.

51. We would also like to stress that software implementation is an excellent way to detect shortcomings or ambiguities in a standard. As a consequence, the specification of a standard should never be carried out as a theoretical exercise only.

VI. SHORTCOMINGS AND CHALLENGES

52. In every project there is room for improvement. A major challenge facing the DDI-effort, is to improve the link to the data production. We agree totally with the statement expressed in Statistics Netherlands' paper to this conference that "the primary goal is to ensure that the reported metadata presented to the end-users match the metadata that drove the process and emerged during the process". (Bethlehem J, Kent J., Willeboordse A. and Ypma W., 1999) The current practice of the social science data archives is to compile the material that goes into the metadata documents from sources handed over by the data producers (either electronically or on paper) and to squeeze this information into the local schemas. The practice ensures a certain level of standardisation across data from different producers, but

is not able to guarantee that all relevant information are included, let alone that the end result matches the metadata that drove, and emerged from, the production process. What we would like to see is the end-user metadata to be extracted directly from the repositories of the data producers. The metadata could of course be refined and extended, especially in situations where the quality standard of the archive exceeds the quality of the formal metadata from the data producers (not an infrequent situation). What a direct link could prevent is end-user metadata of lower quality than the metadata of the producer. A standard like the DDI might provide a platform or an architecture for such a direct link. The remaining challenge is organisational.

53. To encourage a closer integration of the production and dissemination lines, the compatibility with other statistical metadata standards like ISO/IEC 11179 should be investigated. The fact that the current DDI-version includes a mapping to Dublin Core (more info at: http://purl.org/dc/index.htm) but lacks a corresponding interface to alternative statistical metadata standards might be indicative of a closer connection to the library community than the data producer community.

54. Another challenge facing the DDI as well as NESSTAR is the trade-off between completeness and coverage (the more elements and complexity you add to a standard, the lesser the chance that organisations will accept it). The 15 elements of the Dublin Core are one solution to this trade-off. The current DDI –solution is minimalistic to the extreme by only requiring the title-element to be included leaving the rest as optional. Although it is expected that most users of the standard will find its own balancing point somewhere between completeness and the minimum requirements, the solution is unsatisfactory for many reasons. Not knowing what elements are included makes the life difficult for any software that is going to process data directly from the information in the DDI.

55. A final challenge, which ought to be mentioned, is the extension of the DDI-standard and NESSTAR to support complex data. The current standard is a good tool for documenting independent survey files. Attempts have been made to make constructs which also can support aggregate data and hierarchical files, but these are still rudimentary and lack the level of generalisation that is needed.

56. Plans are being made both within the NESSTAR-project and the DDI-committee to approach these and other challenges as soon as the current efforts are completed by the end of 1999. The NESSTAR group has initiated a project, which among other tasks will make an attempt to extend and generalise the DDI-metadata model and to seek a closer integration between this model and other metadata initiatives, especially within the domain of official statistics. For this purpose two more statistical agencies have been invited to participate, Statistics Netherlands and Statistics Norway (in addition two CSO Ireland which is a current member of NESSTAR). ICPSR have taken similar initiatives to launch a project under the label DDI II.

VII. FINAL REMARKS -- METADATA AS COMMUNICATION

57. Metadata is all about communication. Metadata might be looked upon as a structured conversation between the different persons, offices or organisation working with a dataset all the way from the design process to the final users. The main purpose of this structured conversation is to make sure that all relevant information are passed on from one station to the next and that all participants have a chance to add their own relevant knowledge to this information exchange (parallel to the distinction between driving and reporting metadata made in the paper from Statistics Netherlands to this conference).

58. Most producers of statistical data will look upon the end users as legitimate receivers of relevant metadata. That end users also might contribute to the metadata conversation is more unfamiliar. What we should aim at is a feedback system where users of statistical data are allowed to share their experiences with other users as well as with people engaged in the creation of the data. This will include the ability to create links from the metadata to reports and other products of the research process, as well as systems where users are allowed to append comments, advises or warnings to the core body of the metadata. Metadata should consequently be looked upon as open and dynamic over the entire life-span of a datasource and the metadata conversation as multi-directional.

59. All communication needs a shared language and a voice (medium). The same is through for the metadata conversation. Standards, like the DDI DTD, are providing the metadata conversation with a language which can allow persons from different organisations and communities as well as computer processes to participate. But to stage the conversation we also need systems like NESSTAR to give the participants (people and software) a communication medium.

REFERENCES

Bethlehem J, Kent J., Willeboordse A. and Ypma W. (1999) “On the use of metadata in Statistical data processing”, Working Paper No. 23, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, 22-24 September 1999.

Karge R. (1999) “ComeIn – Common Metadata Interface”. Unpublished working document.

King, Gary et.al. “An Operational Social Science Digital Data Library”, Proposal responding to NSF 98-63 Digital Data Library Phase II Program, Harvard University, Cambridge 1998. (http://thedata.org/harum.pdf)

Nielsen, Jan (1997) “From OSIRIS to XML. Markup and Internet Presentation of Structured Data Documentation”. Unpublished thesis.

Musgrave, S. and Ryssevik, J. “The Social Science Dream Machine. Resource Discovery, Analsysis and Delivery on the Web”. Paper presented at IASSIST Conference "Building bridges, breaking barriers: the future of data in the global network", Toronto, May 99. Available at http://www.nesstar.org/M_Paper.shtml

The Dublin Core, Dublin Core Element set. Available at http://purl.org/dc/about/element_set.htm
Also see http://www.dlib.org/dlib/april99/04weibel.html for an assessment of the state of the Dublin Core as of April, 1999.