STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 4
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington D.C., United States, 28-30 November 2000)

Topic (i): Statistical metadata for dissemination

BAZAAR STYLE METADATA IN THE AGE OF THE WEB - AN 'OPEN SOURCE' APPROACH TO METADATA DEVELOPMENT

Submitted by The Norwegian Social Science Data Services [1]

INVITED PAPER

I. INTRODUCTION

1. Metadata is all about communication. Metadata might be looked upon as a structured conversation between the different persons, offices, organisations and even software processes 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 is 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.

2. Most producers of statistical data will look upon the end users as legitimate *receivers* of relevant metadata. The idea that end users might also *contribute* to the metadata conversation is more unfamiliar. What we might envisage are feedback systems 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, advice 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 data source and the metadata conversation as multi-directional.

3. The aim of this paper is to discuss metadata standards and metadata development in the light of this communication perspective. We will also explore the consequences of the move to Internet and the Web as the dominant communication and dissemination medium for statistical information. Our assumption is (following the ideas of Marshall McLuhan) that the introduction of a new communication medium like the Web has an impact far beyond the structuring and packaging of content. New technology changes the very models of communication and creates new methods and patterns of collaboration. In the final section of the paper we are using one of the most interesting models of co-operation rooted in the Web-revolution - the open-source software development movement - to challenge our traditional monolithic view on metadata development as well as metadata standards.

II. METADATA AS COMMUNICATION

4. Many discussions on the nature of metadata are set in the conceptual triangle: "*finding*", "*understanding*" and "*assessing*".

- *Finding*: Metadata facilitates high precision resource discovery. A user never searches for numbers, but for concepts represented by numbers. Trough catalogue information, study descriptions, question texts, definition of concepts or descriptions of sampling procedures, etc.; users are able to locate the collection of numbers that might fulfil their data needs.

[1] Prepared by Jostein Ryssevik.

- *Understanding*: Metadata is giving meaning to numbers. Without human language descriptions of their various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end users. The metadata provides the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

- *Assessing*: Metadata is giving end-user a chance to assess the quality and relevance of a collection of numbers. By describing methodologies and procedures, as well as features related to the context of a particular study, end users are permitted to decide whether or not a data collection meets their professional or scientific standards.

5. There is a fourth concept that is highly relevant to this discussion, and that is "*sharing*". The final product of the statistical production process is not the dataset (the collection of numbers) or even the table report. The dataset and the table report are both artifacts of a more general knowledge production process where the ultimate goal is to improve the understanding of our physical and social environment. The evidence based knowledge production process is an activity with many groups of participants, each bringing different skills and resources to the table. It is also an activity that normally will be distributed in space as well as time. The majority of actors that are involved in this process have not been engaged in the creation of the data and have therefore no access to the "undocumented" and informal knowledge that follows from direct participation in data production. They might also be using the data for other research purposes than those intended by the creators (secondary analysis) and will frequently do their analysis many years after the data were collected.

6. The metadata might be seen as a facilitator for the interchange of information and insight that is the driving force of this process. Metadata makes it possible to extract knowledge from numbers and to share this knowledge with others. At the same time the conversation around the data and the various layers of knowledge products that derive from this conversation should become part of the metadata. For a secondary user of a data source it is of course important to get access to all relevant information about the data production process as provided by the data producer. However, it is also of immense value to get access to the knowledge of previous users, not only to avoid walking down analytical paths that are already fully explored, but also to learn from past experiences and to make it possible to add new approaches and new insight to the layers of already accumulated knowledge. Empirical research is one of the few arenas where it makes sense and indeed also is legitimate to stand on others shoulders.

7. The knowledge and insight acquired through the use of data is not only of value to other users, but can also be exploited by the data producers. Information about how data are used and evaluated by secondary analysts might provide important input to the data production process that on a longer term might improve data collection instruments and methods.

8. The communication perspective on metadata is summarised in the following diagram that uses feedback loops to illustrate how important information derived from the process is fed back into the system.

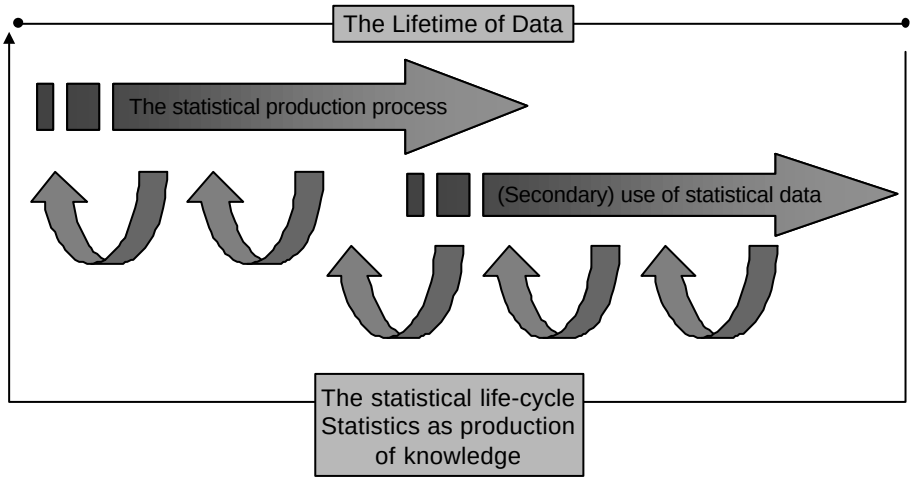

Figure 1: Metadata in the statistical life cycle

9. The perspective leads to an *extended* metadata concept where not only descriptions of the data are relevant information, but also various types of knowledge products (formal as well as informal) deriving from their use. It also implies a *dynamic* concept where metadata is seen as a collection of information that is developed and enriched all the way through the life cycle of the dataset and not something that can be created and published once and for all. Finally, the perspective leads to a concept where a broad spectre of actors are seen as legitimate contributors to the metadata holdings. Whereas the core metadata are still developed by the data producers as part of the data production and publishing process, further layers of metadata will be provided by others as an ongoing activity lasting for many years after the data themselves have left the production line.

III. THE METADATA MEDIUM IS THE METADATA MESSAGE

10. The Web is about to become the dominant media for publication and dissemination of statistical information products, gradually replacing paper-based publications as well as other more static digital products (like CD-roms, etc.). The move has initiated a discussion (within this group and elsewhere) on how statistical metadata should be designed, packaged and optimised to fit the new format. This discussion is important and should continue. The Web is radically different from the communication technologies that we know and are leaving behind and therefore places new and unfamiliar requirements and constraints on the content providers. An illustration: an important metadata concept like footnote which makes perfect sense in the paper-based world from which it originates, might at the best serve as a vague metaphor in a Web-based publication environment.

11. However, there are much deeper and more profound changes following in the wake of the Web-revolution than new requirements on metadata design. With the aphorism "the media is the message" the Canadian communication theorist Marshall McLuhan wanted to put a stronger focus on the medium as such, and not only on the messages or content that are delivered through the various communication channels (see McLuhan, Marshall 1964 and Levinson, Paul, 1999). According to McLuhan the very fact that we are using one media as opposed to another (TV instead of radio, or the Web instead of the printed publication), has a more significant impact on the way we think, work and collaborate, than the given content of any communication. Or applied to the topic of this paper: moving the dissemination of statistical information to the Web not only affects the structuring and design of the accompanying metadata. The move to the Web changes the way in which statistical information is perceived and used in society and thus alters the fundamental concept and function of metadata. In other words, the predominant communication technology has a deep impact on the structuring of the knowledge production process and consequently affects the way numbers and statistical evidence are linked into this process.

12. So, what are the significant features of the Web, which hold the potential of changing the very concept and function of metadata? At least the following should be mentioned:

- *From "one to many" to "many to many"*: The Internet and the Web is the first mass media that really challenges the traditional "from one sender to many receivers" model. The costs and skill requirements needed to provide content are reduced to a minimum, allowing more or less everybody to become a "publisher". The Web also has several layers of formality allowing "quick-publishing" and informal exchange of ideas to exist side by side with more formal contributions. Also part of this picture is the interactivity of the Web, which encourages the user to participate, not only to consume.

- *From publishing to collaboration*: The Web is gradually changing from a publishing media to an arena for collaboration totally independent of time and space. Tim Berners-Lee (the only person that legitimately can name himself the initiator of the Web-revolution) has always seen the Web as an environment for collaboration, but admits that it has taken longer time than expected to reach this aim. However, the direction is unmistakably correct (see Berners-Lee, Tim, 1999).

- *From several local to a global hypertext space*: The Web is *hyperlinked.* It makes it possible to link one piece of information to another, more or less in the same way as the human brain snaps from one idea to the next by means of associations. It is also constituting *a global hypertext space* (breaking the confines of prior hypertext technologies), allowing information objects, totally independent of location or content, to be inter-linked. Moreover, there is no such ting as a linking authority weaving

the Web on our behalf. Everybody can link other resources to their own or create resource-pages or portals that bring together information that according to the creator are related. In this way the Web should be seen as a collective effort growing like a "global brain". (for an interesting view on this aspect, see http://pespmc1.vub.ac.be/.)

- *The Web is genuinely multi-media*: The Web has taken as its content more or less all existing media. Text, pictures, animations, sounds, moving pictures, games or software tools - you name it - they are all available and inter-linked through a single interface, the ever-present Web-browser (note, that it is the browser and not the PC that is the integrating element - through WAP and other emerging technologies the Web functionality is about to break the confines of the computer screen).

- *The Web has memory*: The Web has the ability to remember artefacts as well as interactions and activities. Part of this memory is private (like local mailboxes, bookmarks- and history-lists), but the major part is public, constituting a huge public archive of "historic" documents as well as information exchange (like organised contributions to public news-groups and mail-lists etc.).

- *The Web has the" right" amount of standardisation*: W3C (The World Wide Web Consortium, which is the closest that we come to a Web authority) has taken care not to over-standardise the Web. W3C is developing and recommending the basic protocols (like HTTP), and the general and "all purpose" languages (like HTML, XML, RDF etc.), but these are only providing the necessary level of stability and interoperability to allow the industry as well as the domains (like the statistical community) to develop their own higher-level domain-specific standards. By using languages like XML or RDF to represent standardisation efforts, there is at least a chance that standards and schemas developed within different domains are able to "talk to each other".

13. The list of features could have been made longer. However, the intention of this paper is not to describe the Web in all its exciting details, but to discuss the relevance of some of the more basic traits for our understanding of statistical metadata.

IV. METADATA IN THE AGE OF THE WEB

14. It should not come as a surprise that the communication perspective on metadata outlined above is pretty well served by Web technology. The Web is providing a communication platform that will allow us to establish a metadata-based conversation and to feed the important knowledge products deriving from this conversation back into the system. It is also a system that allows us to link various types of information objects like:

- data descriptions (traditional metadata),
- producer provided knowledge products like reports and fact sheets,
- user provided knowledge products like research notes, discussions, papers and articles,
- people (data collectors, domain experts, external researchers that have used a particular data source etc.)

15. The various elements will create a metadata-space, with layers of hyperlinked information. Closest to the data-core we find the traditional metadata elements developed by the data producers to allow the users to *find*, *understand* and *assess* the described data. Further from the core are objects that provide important contextual knowledge for secondary analysts, but which have been developed for other purposes than to serve as metadata. In the outer circles we will also find information provided by users, as opposed to producers and information that are of a more informal nature than the formal and structured information closer to the core (see Figure 2).

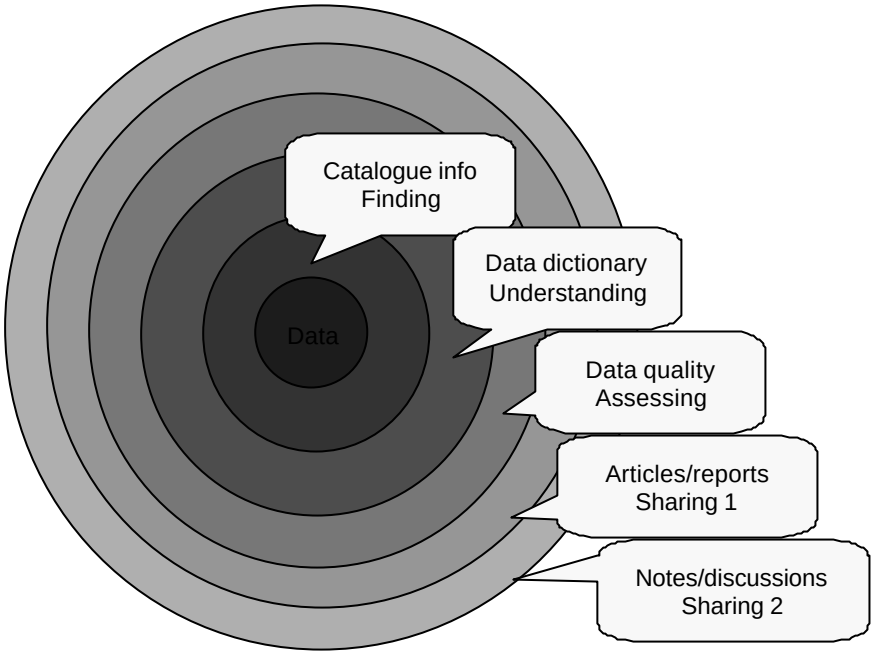

Figure 2: Elements of a hyperlinked metadata-space

16. A network of hyperlinks, which makes it easy to jump from one piece of information to the next, interweaves the entire metadata space. The links are not only providing an efficient navigation network. By associating conceptually related but physically dispersed information they are also adding knowledge to the system. (Part of) the knowledge are in the links.

17. *Scenario 1*: A user that is looking at a data source to test a theory might be alerted to other studies based on the same data or to comments on data quality provided by previous users. She might also be allowed to inspect the relevant contribution to a mail-list, which contains a discussion on the validity of the data for the type or research that she is doing and even to give her own contribution to this discussion. Further down the line she will be allowed to enter a link to her completed research paper to allow future users to take her work into account before they proceed on their own.

18. *Scenario 2*: A user that is reading an article in an on-line journal finds a link that connects him to the data that was used by the author to underpin the arguments. The link allows the reader to rerun the analysis, and also to dig deeper into the same data source. Through the metadata he is even made aware of several other sources that might be used to shed light on the phenomena and which might give another answer than the source used by the author. References or links to papers and articles based on these additional sources are of course also available.

19. Both scenarios are blurring the traditional division of roles between providers and users, authors and readers. In the first scenario a data user becomes a partner in the creation of metadata. In the second scenario a reader becomes a participant in the analysis of the data source. Both scenarios are facilitating cumulative research and both are made possible through hyperlinked metadata.

20. It might be argued that the above description is leading to a vague metadata concept that encompasses more or less all information that can be of any relevance to a user of a dataset. This is absolutely true, and we should even add "over the entire life-span of the data." For the users of a dataset, the "borders" of the metadata concept are totally irrelevant. We know that as we move away from the core metadata as described in Figure 2, we will sooner or later reach a territory, which hardly have anything to do with metadata. But to draw that border between metadata and non-metadata is of no more value than to

draw the border between hot and cold water in the Atlantic Ocean. It is not the object as such that defines it as metadata or not, but the role it can play for potential users.

21. The Web is providing us with the technology and models, which might allow such metadata-spaces to grow. The role of the data producer is not to develop the entire space, but:

- to initiate the process and to publish the core metadata that provides the basic foundation for the metadata conversation
- to use metadata standards that are able to interoperate on the Web and are open-ended enough to include hooks to external information,
- to provide feedback systems that allow users to append or link information objects to the core

The rest of the process should be left to the users and to the collective energy that characterises the Web. It is indeed exactly in this way the entire "Web project" has grown and prospered over the last decade.

22. There is obviously not much left of the traditional monolithic view on metadata in this vision. Metadata is not seen as a coherent and centralised collection of information with clearly defined boundaries, provided by a single authority for a defined community of users. Rather we are looking at a multifaceted collection of information, distributed in space and constantly growing over time, created by a loosely connected network of contributors who are doing it for themselves as much as for any other potential participant.

V. BAZAARS VERSUS CATHEDRALS

23. The approach to metadata development described above has several parallels to a phenomenon that has existed for quite a while, but which first caught attention outside its own narrow tribe of hackers with the release of Linux - the first operating system to challenge Microsoft on it's home ground. Linux is not developed according to the standard models and procedures of the software industry, but is the joint product of thousands of volunteer programmers collaborating over the Internet. By building a software system as complex and vulnerable as an operating system, the loosely connected Linux confederation managed to overturn many of the old-standing truths of the software industry. Overnight "the open source movement", of which the Linux tribe was a prominent member, reached the headlines of the computer press and convinced giants like Netscape, Oracle and IBM to join the party in order to keep up the innovation.

24. Eric S. Raymond, the self appointed participant anthropologist of the open source movement has described the new style of development as a "great babbling bazaar of differing agendas and approaches". In his now famous essay, published on the Web in 1998 (see Raymond E.S., 1998), Raymond compared this Bazaar style of development with the centralised and carefully planned Cathedral-building models of the software industry and he pointed to several distinguishing factors that may explain it's success.

25. One obvious factor is the total *openness* of the Bazaar model. The initiator of an open-source project makes all the source-code and documentation publicly available and invites others to criticise the approach and to come up with improvements. All the way trough the lifetime of the project all developed artefacts are open for inspection and discussion, and superior approaches are gradually replacing those that fail to convince. Open source projects thus resemble a *conversation* gradually moving foreword through exchange and evaluation of arguments.

26. Another important factor is the recognition of the value of different types of knowledge and the simple fact that many heads inevitably are better than a few. By excluding no one from the party, intelligently managed open source projects can master a pool of talent and profit from a constant flow of new perspectives and ideas that can boost innovation as well as quality.

27. According to Raymond, a third and very decisive factor is the belief in the value of the users. Users are treated as co-developers and not as receivers of a finalised commodity. By allowing users to make proposals or even to take part in the discussions among the developers, users can add value to the product. Or to site Raymond: “The next good thing to having good ideas is recognizing good ideas from your users”. Feedback is a key to success. As in our metadata-scenarios (see paragraph 17 and 18 above)

the division of roles between providers and users, authors and reader is blurred – and again to the profit of the knowledge production process.

28. The real challenge of the open source model is to understand what drives thousands of smart programmers to spend parts of their highly valued time to develop code and (even to document the code so that everybody else can understand it) for free. What is the motivation? Again according to Raymond, open source hackers are not genuine altruists. The driving force behind their voluntary efforts is maximisation of reputation and status within the group. To invent the killer algorithm or to locate the faulty line of code that have caused the system to crash for no obvious reason allows the contributor to climb in the culture's status hierarchy (see Raymond E.S., 1999 and Kuwabara, K., 2000).

29. Primitive as it may sound, the motivating force is not very different from the one we find in any non-profit academic research environment, including empirical social science that we are addressing in this paper. Merits are measured according to intellectual contributions and status allocation resides on a system of peer reviews.

30. Developing software and metadata are obviously two different things, so the parallel should not be over-stretched. However, the important thing to notice is that there are alternatives to cathedral building in software as well as metadata and that this alternative resides on a system of interaction, feedback and involvement of a variety of actors, many of whom we traditionally have classified as users.

VI. TOWARDS THE DATA-WEB

31. Many of the ideas presented in this paper are implemented or about to be implemented in a Web-based data access and dissemination system called NESSTAR (Networked Social Science Tools and Resources), developed by the UK Data Archive, the Danish Data Archive (DDA) and the Norwegian Social Science Data Services (NSD). The basic philosophy of NESSTAR is to provide a software system that will allow producers and disseminators of statistical data to publish their resources on the Internet, either as standalone service to a limited group of identified users, or as an open offering connected to a distributed virtual data library. For the end users of statistical data, the system will provide a flexible interface that will allow them to search for data across the holdings of a broad range of data publishers, to brows detailed descriptions of the data (metadata), to visualise and analyse data on-line and to download data in a variety of formats ready for further local processing. (Musgrave, S. and Ryssevik, J. 1999 and 2000, more information are available at www.nesstar.org and www.faster-data.org)

32. NESSTAR is building upon an XML-based metadata standard developed by an international committee of data archives and data producers called the Data Documentation Initiative (DDI). This is a very end-user oriented metadata standard, allowing a rich amount of semi-structured information to travel along with the data on their way from the production line to the secondary analysts. One of the most important features of the DDI-standard is the ability to embed Web hyperlinks (URIs) in every metadata-element allowing external resources to be referenced or linked. This might include references to external knowledge products provided by the data publisher, as well as products and information provided by others (Ryssevik, J., 1999).

33. NESSTAR is supporting this feature of the DDI-standard. Moreover, the entire communication protocol of the NESSTAR system is based on messages composed as URIs. This is allowing every information object on a NESSTAR server to be hyperlinked or bookmarked, from within the NESSTAR client as well as from external Web objects. Any search for data, any dataset or any table or analysis derived from stored data can thus be described as a URI and activated from any other resource that are stored on the Web. NESSTAR is therefor providing 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.

34. Both of the metadata scenarios described above (see paragraph 17 and 18) are consequently feasible in a NESSTAR environment. A research paper or report published on the Web as an HTML or PDF document can perfectly well include a hyperlink (URI) that gives the reader direct and live access to the underlying data stored on a NESSTAR server. The reader will be allowed to rerun the analysis, bring in new variables or even use the active data as a springboard to find similar data sources that can throw

light on the research topic in question. Starting from the other end - the data source - users will also be allowed to find relevant reports, papers or other documents that are based on the particular data.

35. What is lacking in the current NESSTAR system is the feedback loop that gives external users the chance to link their contributions directly into the metadata. In the current scenario links to external resources must be created by the data publisher - a procedure far to rigid to support the dynamic and open-ended metadata conversation that we have argued for in this paper. However, this technology is underway. What is aimed for is a system that allows the user of a dataset to create links from the metadata to reports and other products of the research process, as well as a system where users can append comments, advises, warnings or proposals to the core body of the metadata.

36. NESSTAR is one of several projects engaged in the development of what gradually has become known as the "data Web". Other relevant projects on this arena are the Virtual Data Centre currently under development at Harvard-MIT (King 1998) and the Ferret system developed by the U.S. Census Bureau (http://dataferrett.census.gov/). The common goal of all of these projects is to use open standards to build a true "data Web" where the models, technologies and collective energy of the Web is brought to the world of statistics.

37. Metadata specification languages like XML and RDF, developed by W3C, are providing important building blocks for this endeavour. However, to succeed we might be forced to rethink what we really mean by metadata standards and how we organise their development. In our current statistical information systems even standards are Cathedrals - they take ages to build and if ever completed they are literally "cut in stone". In the environment of the Web, development of new standards is normally measured in months, not in years. Any standard that takes more than a couple of years to develop bears the risk of becoming obsolete before it is even published.

38. Current metadata standards, including the DDI, are also static - they need to be revised in order to support a new requirement. What we will need in order to build the "data Web" are metadata standards that are flexible enough to evolve without initiating a costly and time-consuming revision process. In addition to an agreements on the key concepts and their relationships, the new generation of standards should include extensibility mechanisms that make it possible to add new concepts and relationships by building on what is already known.

39. Cathedral standards are based on an assumption that it is possible to reach global agreement on every little detail of a complex construction. There are few real-life examples to support this assumption. Following Tim Berners-Lee's vision of the Semantic Web we might be satisfied with "partial understandings", that is agreements on the key concepts and a common logical framework to express the local variations (Berners-Lee, 1999, pp. 201 -).

REFERENCES

Berners-Lee, Tim (1999), "Weaving the Web - The Past, Present and Future of the World Wide Web by its Inventor", Orion Business Books, Great Britain.

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

Kuwabara, Ko (2000), "Linux: A Bazaar at the Edge of Chaos", published in the Web-Journal First Monday Volume 5, Number 3 - March 6th 2000, available at http://www.firstmonday.dk/issues/issue5_3/kuwabara/index.html

Levinson, Paul (1999), "Digital McLuhan - a Guide to the Information Millennium", Routledge, London

McLuhan, Marshall (1964), "Understanding Media - The Extensions of Man", MIT Press Edition, 1994.

Musgrave, S. and Ryssevik, J. (1999) “The Social Science Dream Machine. Resource Discovery, Analysis and Delivery on the Web”. Paper presented at IASSIST Conference "Building bridges, breaking barriers: the future of data in the global network", Toronto, May 1999. Available at http://www.nesstar.org/M_Paper.shtml

Musgrave, S. and Ryssevik, J. (2000) "Beyond NESSTAR: FASTER Access to Data" Paper presented at IASSIST Conference , Chicago, June 2000. Available at: http://www.faster-data.org/FASTER.doc.

Raymond, E.S. (1998), "The Cathedral and the Bazaar", published in the Web-Journal First Monday Vol.3 No.3 - March 2nd. 1998, available at http://www.firstmonday.dk/issues/issue3_3/raymond/index.html

Raymond, E.S. (1999), "Homesteading the Noosphere" published in the Web-Journal First Monday Vol.3 No. 10- October 5th. 1998, available at http://www.firstmonday.dk/issues/issue3_10/raymond/index.html

Ryssevik, J. (1999), "Providing Global Access to Distributed Data through Metadata Standardisation – The Parallel Stories of NESSTAR and the DDI", Working paper no. 10 from the UN/ECE Work Session on Statistical Metadata, Geneva, September 1999.