

# **ILSES: HOW LIBRARY AND DATA ARCHIVE MEET IN ACTIVE SUPPORT OF RESEARCH IN THE SOCIAL SCIENCES**

**By Repke Eduard de Vries**

## **Introduction and background**

Empirical social scientific research follows a cyclic path in which collecting of data is followed by data analysis, which is followed by publication of the analysis results. The publication in turn can lead to new questions that need new data, which are again analyzed and published upon. Or data analysis and book or article have not exhausted all the explorations and answers possible with a particular data collecting effort: secondary analysis then follows a primary one. Especially national central statistical bureaus gather survey data with high relevance for subsequent social scientific research, be it academic in nature, in a supportive role for government policy making, or to inform the general public. Supra national does Europe know such survey data collections as “Eurobarometers”, funded and initiated by the European Commission<sup>1</sup>. By consequence over the years a large body of social scientific literature has accumulated on the one hand and equally well collections of computer (survey) data files have been build on the other. In terms of responsibilities for preserving and giving access to what has accumulated and been collected however, book and data file - though coming from the same research process - have always been subject to different facilities and infrastructure.

Libraries with their long-standing experience with printed works, took care of publications but data archives as a relatively very young facility<sup>2</sup> started doing the same for what is called the “original” or “raw” data. The difference in physical appearance and technical format of printed books and computer data files and the different needed technology for long term storage and access, made suchdata archival specialization necessary at the time. As a consequence though

---

<sup>1</sup> Eurobarometers: <http://europa.eu.int/en/comm/dg10/infcom/epo/polls.html>  
For data distribution:  
<http://www.za.uni-koeln.de/data/en/eurobarometer/index.htm>

<sup>2</sup> From the early sixties onwards ; further information for Europe at the CESSDA Web site, that also has pointers to data archives world wide: <http://www.nsd.uib.no/cessda>

of the two different infrastructures and the different responsibility taking, bibliographic references for books and summarizing metadata for data files have been created separately and without keeping together the interrelated nature of publications and original data. On the library side does it take proper citation by authors of books and journal articles to data used, which linking information cannot be repeated and expressed in bibliographic records. Searching library catalogues therefore reveals the books but not the data. Standards have been developed for citation of “machine readable data files”<sup>3</sup> as has the library world developed a standard for cataloguing computer files as such.<sup>4</sup> But data citation is still not consistently practiced by authors and cataloguing rules have less of an impact if social science data are not part of library collections but remain a data archival responsibility.

At data archives technical documentation and publications are part of the collection but as an accompaniment to the data it self and not for lending. The study descriptions as equivalents to the bibliographic records in libraries, also focus on the research that produced the data and though they do reference to publications it is most of the time limited to the publication directly linked to data collection and first analysis results. Only some data archives and for some particular data series, maintain bibliographies of all publications that analyze particular data sets.<sup>5</sup> Searching the holdings of data archives therefore gives access to data and with the help of bibliographies to just a selection of books.

A situation where data and books are separately referenced without consistent cross linking, have to be searched for in separate catalogues and are given access to by different authorities and with different facilities, has consequences for any one embarking upon new research or in general needing social scientific information. It is not possible to start with general literature searches in libraries and easily trace back publications to the empirical research and collected data that is at the heart of it. If one succeeds at all, subsequent data access would take separate catalogue searching at data archives followed by ordering (electronic)

---

<sup>3</sup> Most notably the work by Sue Dodd but also see: <http://dpls.dacc.wisc.edu/cite.html>

<sup>4</sup> The ISBD - CF (computer files) ; information about this standard can be found at the IFLA site: <http://www.nlc-bnc.ca/ifla/>

<sup>5</sup> An arbitrary example at: <http://dpls.dacc.wisc.edu/pubs/publications.html#biblio> Another approach is having linking information to publications in the searchable data catalogue records: at the Dutch NIWI (<http://www.niwi.knaw.nl>) choose information type “data” and discipline “social sciences” and search for example for “election”

copies of the original data. Neither can archive catalogues (even when expanded with bibliographies) help with book and article searches starting from particular data collecting efforts. Properly linking data and publications would need metadata standards<sup>6</sup> that take such relationships into account and coordinated efforts between authors (proper citation of data sources or writing such metadata directly themselves), the library world (referencing with cross linking in new metadata formats) and the data archives (likewise referencing with cross linking). Part of those efforts would also have to be a common catalogue search facility or some form of easy access from one catalogue to information in the other.

Some coordination has been achieved already, especially now that the last few years the Internet and related technologies make it possible for every one to at least search separate catalogues with standard Web browsers and in some cases to access data and electronic publications directly. World Wide Web techniques for linking electronic resources on the Internet but also new metadata initiatives that explicitly hold linking information to related (electronic) resources, have the potential to finally bring data and book together again for searching and retrieval. Employing Internet techniques thereby overcomes the presently separate infrastructure, management of holdings and reference services.

A few Internet related projects can be mentioned that demonstrate first attempts in this direction.

ICPSR, as one of the largest data archives and having a global coverage, together with the academic journal publishing world for the social sciences broadly, started an initiative where (printed) journal articles reporting on empirical research not only have proper citations to data used, but actually have the data itself directly available over the Internet. The latter is a service by ICPSR called the "Publication-Related Archive" where authors deposit their data electronically and colleagues and interested readers of the publication, have subsequent access for downloading.<sup>7</sup> In operation now for a few years, this initiative shows and encourages data availability from the perspective of publishing and at the level of publications. It has several more potentials: with electronic journals, access to the data could right away be in the text with help of WWW "clickable links". And with consistent referencing by authors to the data as electronic resource, metadata

---

<sup>6</sup> Dublin Core, for which the Nordic Metadata Project is a good starting point for orientation: <http://renki.helsinki.fi/meta/> And the W3C RDF (Resource Description Framework) proposals in particular: <http://www.w3.org/Metadata/>

<sup>7</sup> At: [http://www.icpsr.umich.edu/ICPSR/Other\\_Resources/pr.html](http://www.icpsr.umich.edu/ICPSR/Other_Resources/pr.html)

for these published articles can take this linking into account too. Which in turn would in the long run weave a web of relations between books, articles and data, at the level of catalogue searching.

SOSIG, the UK Internet “Social Science Information Gateway” , is a longer existing subject information based gateway to help researchers locate relevant material. The type of information can be diverse: from data to ongoing research projects with a particular focus. All resources have standardized descriptions and are indexed with a subject classification. In this approach the classification helps bring together social scientific information types, that otherwise would have had to be searched for in very different catalogues and by themselves miss all the linkages that SOSIG now establishes after the fact.<sup>8</sup>

NESSTAR, Networked European Social Science Tools and Resources, is a European Internet project under way that explicitly seeks to open up data archival holdings information, the data related documentation in electronic fashion and the data it self, to existing electronic search facilities in the library world and to new metadata initiatives.<sup>9</sup> Over the next few years NESSTAR will pilot closing the infrastructural gap separating data and related publications. Initiatives like the “Publication-Related Archive” by the publishing world and ICPSR, can greatly benefit from closing these technical gaps.

Finally for Europe the initiatives by the European Commission should be mentioned that are “creating a platform for the library in the Information Society”.<sup>10</sup> The present Libraries work programme (1995-1998, and implemented under the Fourth Framework Programmes) knows three Action Lines, one of which is based upon the recognition that the “traditional role of the library has been to provide access to resources they themselves collect and store”, but that “increasingly libraries are required to offer access to networked information resources. These resources include file archives and data sets (documents, software, images, statistical surveys, etc.), as well as interactive services” This Action Line formulates a role for libraries “in the organization and distribution of networked information and to act as the intermediary between the end-user and the resource ..”<sup>11</sup> These assessments not only recognize how data (data archives)

---

<sup>8</sup> At: <http://sosig.esrc.bris.ac.uk/>

<sup>9</sup> At: <http://dawwww.essex.ac.uk/projects/nesstar/>

<sup>10</sup> Iljon, A. Objectives and strategies - creating a platform for the library in the Information Society. *IFLA conference, Copenhagen, 1997*

<sup>11</sup> Idem

and publications (libraries) for the social sciences have been resources with too much separate access for end-users, but also formulate new policy and how “networked information” has to play a key role in that policy.

ILSES, or “Integrated Library and Survey-data Extraction Service”, is funded under this European Libraries Programme<sup>12</sup> and is developing a service that enables individual users or librarians in an intermediary role, to access and retrieve in an integrated manner data and publications coming from large-scale surveys. It is an Internet service, with the Internet also being used for content providing. Experts at (specialized) libraries bring bibliographic reference information over the Net into the ILSSES system and access ILSSES over the Net to link publications with data and to index books and articles with thesaurus based keywords. Experts at data archives in a similar fashion bring data and data documentation (technical and about the research that produced it) into ILSSES. They too can apply keywords from the same thesaurus but classified now are the topics addressed with questions in surveys. Like SOSIG does keywording bring the two resources “data” and “publications” together in ILSSES. The linkage between book and data moreover is directly available when searching ILSSES as a library catalogue system: this is an improvement over the “Publication-Related Archive” initiative, where references to data only appear at the level of the journal article it self and cannot be searched for. Also is ILSSES more symmetrical: users not only have access to data starting their quest from publications, but ILSSES also makes survey questionnaires and data documentation searchable, including links from data to those publications where the data is analyzed or reported upon. The fact that ILSSES is both an Internet service but also software that content providers can apply themselves to provide this service, fits very well with the roles anticipated for libraries in the European Libraries Programme, i.e. a role: “in the organization and distribution of networked information and to act as the intermediary between the end-user and the resource .”<sup>13</sup>

NESSTAR creates unique data oriented services of its own (like exploratory data analysis and data visualization over the Internet) but by opening data archival infrastructure and resources in electronic and standardized ways, also creates the kind of “networked information” that enables services like ILSSES to be accomplished much more easily. Uniformly standardized data documentation and uniform data access together with longer established library standards for

---

<sup>12</sup> The Telematics for Libraries program is at: <http://www2.echo.lu/libraries/en/libraries.html> ILSSES is at: <http://www.gamma.rug.nl/ilses/>

<sup>13</sup> A. Iljon ; see note 10

bibliographic information like UNIMARC, will bring very easy building blocks for new types of user services like ILSES. Developing metadata standards like Dublin Core can further catalyze this because of their orientation on electronic resources and their taking into account of linkages between resources.

### **ILSES - an overview.**

ILSES is a project of a number of Dutch, German, French, and Irish institutes: together representing the library world, the data archiving world and the world of academic social scientific research. It is funded by the European Commission under the Fourth Framework, Telematics for Libraries and has been running since September 1996. With several modules finished and some to follow, the project will end September 1999. The ILSES Web pages and those for the EC Telematics for Libraries Program, give further general information.<sup>14</sup>

Following is a brief overview of the different ILSES modules, the standards that ILSES adheres to and how ILSES relates to existing models of Internet user services and Internet information-networking. The overview makes a distinction between modules for end-users and for content providers. End-users can be individual researchers directly or intermediaries at (data) libraries. Content providers are roughly libraries for bibliographic information and expertise, and data archives for data, technical data documentation, survey questionnaires and background information on the research that produced the data. Technically these responsibilities could be in one hand however but also could researchers or in general anyone with Internet access, be asked to collaborate over the Net and use ILSES to build up information on data and related publications.

### **ILSES modules**

For content providers ILSES has three modules: Administrator-ILSES, LIB-ILSES and DAT-ILSES. The first one manages general matters like access authorization to ILSES and importing the thesaurus to be used. At present this is a subset of HASSET, which in turn is based upon the 1977 Unesco thesaurus.<sup>15</sup> Administrator-ILSES also keeps the database that holds all data oriented information, all publication related information, the applied keywords to both and all further linking information between the two.

---

<sup>14</sup> See note 12

<sup>15</sup> HASSET (and its roots in the Unesco thesaurus) : <http://155.245.254.46/services/nhasset.html>

LIBrary - ILSES is a content provider module for importing bibliographic references into ILSES, for indexing the referenced publications with thesaurus terms or alternatively for linking books or articles directly to a particular data-collecting-effort (also known as “the study”) or even directly to particular questions in a questionnaire (also known as the “variables in the data”). Bibliographic records in ILSES can also be entered and corrected or enhanced manually. The standard for importing is UNIMARC of which a subset is implemented. Though UNIMARC is reported<sup>16</sup> to adopt in its definition a field for Electronic Location and Access (developed in USMARC as the “856 field”), ILSES for the present moment does not import this field but lets it create manually. Handling Electronic Location and Access information anticipates the growing number of electronic publications that end-users in this way will be able to retrieve directly from the Internet by means of the ILSES end-user module E-ILSES. Printed publications referenced will have to be lended in traditional ways.

LIB-ILSES recognizes monographs (books), “chapters” (from compilations, readers and so on), serials and articles appearing in a particular volume of a serial. It also handles the relevant UNIMARC fields for linking journal article descriptions to the higher level of the serial and likewise for linking “chapters” to the higher level of “compilations”. Subject classifications are accepted from UNIMARC records when they match with terms from the ILSES thesaurus.

DAT-ILSES is a content provider module for importing technical data documentation (or “codebooks”), for manually entering background information on research studies, for relating scanned images of the questionnaires to those parts of the data that hold answers to questions asked on a particular questionnaire page, and for indexing variables (or “questions”) in the data with thesaurus terms. Standards for data documentation have always been less established but DAT-ILSES recognizes one structured format and anticipates the emergence of a new standard based upon SGML.<sup>17</sup> End-users of ILSES interested in the data itself for further analysis, can make selections and do automatic ordering with the E-ILSES module. The data is not kept in the ILSES database, nor are the scanned questionnaires. Instead the database has pointers to data archives holding the data and electronic references to where the images reside.

---

<sup>16</sup> Fernanda Maria Campos Unimarc: the virtual format in the virtual age. *IFLA conference, Copenhagen, 1997*

<sup>17</sup> The codebook format implemented is “OSIRIS”; the new SGML initiative is the DDI: <http://www.icpsr.umich.edu/DDI/>

All these content provider modules have been finished. They all are software residing at a content providers' own PC and communicating over the Internet in standardized ways with the central ILSES system and database. Content material chosen for the ILSES demonstrator are a number of Eurobarometer large scale surveys (held in a series of European countries)<sup>18</sup> the publications based upon these surveys.

For end-users ILSES will have two modules: E-ILSES and NET-ILSES. The first module is a dedicated program that installs on a users' own PC and communicates (over the Internet) with the ILSES system under a client-server model and with SQL queries for information search and retrieval. The second is an approach where ILSES produces a number of Web pages from its information about data and publications and the relations between them. Such a Web site can be browsed, searched and retrieved from by end-users with standard Web browsers. Both modules are planned for 1998. Their goal is it to give researchers, intermediaries and anyone interested in social scientific information, an easy and integrated view on the continuous process of empirical research and its two outcomes: data and publications. Therefore a user can start from both perspectives: browsing data documentation and expand to related literature or search for publications and go to the data sources. ILSES also provides for both: data can be subsetted, combined and retrieved or bibliographies can be extracted and where possible electronic publications retrieved.

Repke Eduard de Vries  
NIWI: Nederlands Instituut voor Wetenschappelijke  
Informatiediensten  
(Netherlands Institute for Scientific Information  
Services)  
P.O. Box 95110  
1090 HC Amsterdam  
The Netherlands  
e-mail: repke.de.vries@niwi.knaw.nl  
<http://www.niwi.knaw.nl>  
Tel: +31 20 4628670  
Fax. +31 20 668 5079

---

<sup>18</sup> see note 1