Can the Library and the Data Archive Meet in Active Support of Research in the Social Sciences? *The Case of ILSES*

by Repke de Vries *

ILSES: development of tools for an Integrated Library and (Survey) Data Extraction Service. http:// www.gamma.rug.nl/ilses

Project under the European (EC) Telematics for Libraries Program.

Partners: ProGamma (Netherlands), ZA (Germany), NIWI (Netherlands), University of Amsterdam (Netherlands) and associate partners BSDP (France) and Trinity College (Ireland).

Introduction

Data material collected for empirical research has traditionally been computer stored and electronically distributed by data archives and data libraries. Whereas publications from the same research were kept, referenced and given access to by libraries.

As content providers data archives could not extend their services with relevant book and journal collections, cross referencing and lending of printed material. Libraries could not give access to data related to published research or had the means to expand bibliographic references to also point at data as machine readable outcome of the research process.

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. Neither can data 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 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.

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.

A recent publication[1] is referred to for a more complete treatment, including a few Internet related projects that already demonstrate first attempts in this direction. One of these is ICPSR's "Publication Related Archive"[2], another the European NESSTAR project[3]. In the same publication ILSES as Integrated Library and Survey-Data Extraction Service, a system of tools and (Internet) facilities, is expanded upon as a current project funded within the Library Programme of the European Commission. ILSES addresses the same goal of integrating publication and data. To achieve this, it accommodates both content providers (libraries and data archives) and end-users.[4]

Other approaches and further developments

In the UK e(lectronic)Lib(raries) program, the Open Journal project has been working on mechanisms and demonstrators for "citation linking" in the broadest sense.

"Using citations - the links made by authors themselves - users can navigate between their current work and a priori work in the archives of the research literature or take a recent paper and move forward, tracking the citations dynamically"[5]

In particular did the Open Journal develop Internet solutions to create, maintain and give access to "distributed links" between primary and subsequent secondary sources, when electronic information is available that never received embedded links or that simply does not have the internal format to adopt such links.[6] These could easily include formats like data material distributed over the Web, collections following from digitization or MARC type of bibliographic references, which by themselves do not have entries for cross linking. Another advantage of the Open Journal approach could be the fact that information to be

linked in “citation style” fashion, can be arbitrarily distributed over the Net and that establishing the links can be a separate, dedicated activity at any point in time.

Central thesaurus facilities over the Internet are another key issue in tying together distributed information. When metadata can be attached to both data and related publications that take indexing terms from a central (domain specific) thesaurus, future searches will bring up related material because of such common terms. The UK Data Archive has established such an Internet based thesaurus facility, that also takes into account web based forms to submit new possible additions to the thesaurus.[7]

Metadata

The Dublin Core metadata definition is both finalizing in details but also has a core set that already finds application in sometimes large scale projects. Last years DC 5 conference in Helsinki reflect this. Both a series of current DC projects were discussed[8] and several DC elements received further clarification and more precise definition.[9] Two projects in particular address the same issue of integrating diverse but related information types - one by the Australian Geodynamics Cooperative Research Centre and one by the UC Berkeley Digital Library Catalog.[10] It is interesting to see the integration approached in a distributed DC metadata fashion instead of by a central database model.

At and following the conference the DC.Relation element (among others) received further definition. The now proposed six sub-elements have enormous potential for cross-referencing and thus having build-in links to go from a DC data material description to following DC descriptions of article and book publications - especially where these are electronically available as well.[11]

Conclusion

The technology, connectivity and development of standards is available enough to start bridging the two infrastructures of access to data and access to publications following analysis of those data or touching upon the same theme.

ILSES addresses that goal with very concrete tools and solutions. Hopefully it can both be useful for end-users and at the same serve as an evaluation of the type of approach chosen, i.e. a central database keeping all the metadata and the linking information.

It seems that the library world with its digital library projects and its developing metadata standards and distributed models have an advantage for future solutions and more momentum to realize these. The data archiving world should be quick to take up the challenge and start working with library people on the common cause of giving researchers complete access to all the related information of any type, following from their research activities. From the beginning should this access be complemented by facilities for these same researchers (data collectors, authors, publishers, depositors) to create metadata and linking (citation-) information equally well themselves.

References

1. Vries, R.E.de: ILSES: how library and data archive meet in active support of research in the social sciences. INSPEL Vol. 31 No. 4, 1997. The article is also electronically available in PDF format: <http://www.fh-potsdam.de/~IFLA/INSPEL>

2. ICPSR “Publication Related Archive” <http://www.icpsr.umich.edu/ICPSR/Other_Resources/pra.html>

3. NESSTAR (Networking European Social Science Tools and Resources) <http://dawww.essex.ac.uk/projects/nesstar/> Though NESSTAR’s first goal is integrated access across holdings at different data archives, the choice for Z39.50 at least opens their catalog searching to libraries and thus bridges the separate infrastructures

4. ILSES <http://www.gamma.rug.nl/ilses>

5. From the IRISS’98 conference: Hitchcock, Steve et al.: Webs of research: putting the user in control. <http://sosig.ac.uk/iriss/papers/paper42.htm>

6. Carr, Leslie et al.: The distributed link service: a tool for publishers, authors and readers. <http://www.w3.org/Conferences/WWW4/Papers/178/> Please note the IRISS’98 paper in note 5 has references to further work on these tools by the Open Journal Project

7. The HASSET thesaurus project at the UK Data Archive, jointly with other parties: <http://biron.essex.ac.uk/cgi-bin/zhasset

8. DC conference October 1997: Project Presentations <http://linnea.helsinki.fi/meta/projects.html>

9. A very thorough account is given by Diann Rusch-Feja: Entwicklungen der Dublin Core Metadaten: Bericht ueber den 5. Dublin Core Metadata Workshop .... (full title abbreviated) <http://www.mpib-berlin.mpg.de/DOK/dc5ber.htm>

10. Project Presentations 19 and 25 from the document in note 8

11. See Rusch-Feja’s article in note 9 ; the relevant section is almost self explanatory and a series of further URL’s of mostly English information sources is given.

* Paper presented at the 1998 IASSIST/CSS Conference, Yale University, New Haven, USA, May 19-22, 1998,
Repke de Vries, ILSES - NIWI: repke.de.vries@niwi.knaw.nl