WP 8
16 June 2010

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

UNECE Work Session on the Communication of Statistics
(30 June – 2 July 2010, Paris, France)

(i) Getting the numbers across in ways that external audiences understand

WEAVING DATA INTO INFORMATION NETWORKS USED BY SCHOLARS, RESEARCHERS AND STUDENTS

Submitted by the Organisation for Economic Co-operation and Development (OECD)[1]

I. INTRODUCTION

1. The online publishing revolution has drawn working papers, journals and books into a seamless information resource network, regardless of publisher. At the heart of this network are 'nodes': cross-linking systems like CrossRef[2], federated search systems like Serials Solutions, aggregation platforms like IngentaConnect, community sites like REPeC[3] and, of course, discovery engines like Google Scholar, Scopus and Web of Science.

2. The quality and range of the content available in this network is a boon to time-pressed researchers and students who no longer have to wade through the trivia that clutters general Internet search engines such as Google or Bing - they can trust what they find via these specialist information networks.

3. Currently, datasets are not included in these information networks – making it hard for researchers and students to find statistical data they can trust. The information resource network also provides tools to help researchers and students manage what they have read, prepare citations and build bibliographies - these tools are not provided by data providers, again, making it harder for researchers and students to get their work done quickly and efficiently.

4. This paper describes OECD's citation tool for datasets. Described by John Wilbanks, vice president for Science of Creative Commons and the executive director of Science Commons, as 'frighteningly sane'[45], this new tool is a centrepiece for a new publishing platform that will weave OECD's datasets (together with its working papers, journals, and books) into the global information resource network used by researchers and students around the world.

II. THE SCHOLARLY INFORMATION RESOURCE NETWORK

A. Nothing's changed

5. A visit to any library or bookshop shows that content has always been channelled into various silos: scholarly, general trade, literature, children's books and so on. Behind the scenes, each particular silo has evolved a supply chain to take an author's work to the intended audience. In the case of scholarly content, for example, much is written in article form. The articles are gathered into specialist journals that are mainly

[1] Prepared by Toby Green, Head of Publishing (Toby.GREEN@oecd.org).
[2] www.crossref.org
[3] www.repec.org
[4] wilbanks"cite this database" button from the OECD. make it easy to cite, not illegal to copy. frighteningly sane. #NASEJ, Twitter March 24th 2010

collected by libraries, via subscription agents, at major research institutions. Before the arrival of the Internet, researchers and students would visit the library to find, retrieve and read the articles they needed. Over the years, the supply chain evolved. First, the big journals 'twigged' into many niche titles, each serving ever-smaller specialties. In response to the explosion in the number of journals, secondary services emerged: alerting and discovery systems in the form of cataloguing and abstracting and indexing services helped readers find articles more efficiently. If your library didn't subscribe, it wasn't a problem, Inter-library loan and document delivery systems would get the articles you needed. In short, an entire ecology of services evolved to help scholars publish, discover and read what they needed.

6. To work more effectively, standards evolved in the system. For example: to help researchers 'link' content they had read to new content they were generating, citation standards emerged. Citations themselves were grouped into bibliographies and reference listings at the end of each article and book and everyone used them. Over time, stand-alone bibliographies appeared to help guide researchers new to a field to the core work that had been published.

7. The system worked because it was efficient and well-understood by all stakeholders. A researcher would no more waste their time looking for a scholarly article in a bookshop as they would try and buy meat from a baker. The combination of full-text publications (journals, books) with the various discovery systems (indexes, bibliographies, catalogues, reference listings etc) created a 'scholarly information resource system'. They know it not only contained all the trusted and reliable scholarly content needed for their work and studies but it also contained all the tools and services needed for them to publish, discover and obtain the scholarly information they needed.

8. The arrival of the Internet changed nothing and everything. It changed nothing because researchers still need a 'scholarly information resource system'. It changed everything because all parts of the system moved online. Journals still exist, as do cataloguing and abstracting and indexing services – they are simply delivered online as well as in print. Libraries still build collections and make content available to the communities they serve – but increasingly via online services rather than print. If your library doesn't subscribe, document delivery systems fix the problem – again, online. Citations in reference listings are now live links with the cited content just a mouse-click away.

9. It has taken a little over a decade for the scholarly information resource system to shift from an almost entirely off-line model to an almost entirely online model. But the key point is this: the audience has moved with the system - online. So, if your content isn't in the system the chances are it won't reach a maximum possible audience.

B. Everything's changed

10. ICT upgraded the whole supply chain. Every aspect from the way authors submit their manuscripts right through to the way readers access articles online improved. Along the way, new problems and new opportunities emerged, spawning new services and solutions that enriched the online version of the scholarly information resource network.

11. Examples include:

a. Digital Object Identifiers – the 'ISBN' for objects on the Internet: a unique identifier for objects published online (www.doi.org).

b. CrossRef – a 'telephone exchange' for citation-links, it enables any citation to be turned into a link regardless of where the publication is hosted (www.crossref.org).

c. Federated search tools that provide a single interface for the many scholarly catalogues and publisher sites, e.g. Serials Solutions.

d. Aggregation platforms where small publishers can pool their full-text content online at little cost to themselves e.g. IngentaConnect.

e. Specialist search engines that focus only on scholarly resources, e.g. Google Scholar, Scopus and Web of Science.

12. The emergence of Web 2.0 meant that users could self-publish. In economics, a community of researchers created RePEC (Research Papers in Economics) to share their working papers. This site has since grown to become the single biggest online collection of working papers, journal articles and book chapters in economics – with an audience to match.

III. DATA AND THE SCHOLARLY INFORMATION RESOURCE SYSTEM

13. Before the Internet, data was published in printed form, usually as books. These were made available via the scholarly publications supply chain and ended up being catalogued, cited and so on by librarians and researchers along with the other scholarly content. In short, they were integrated into the pre-Internet era scholarly information resource system alongside analytical content.

14. However, when the Internet opened up the possibility of publishing data online, data producers simply posted datasets online and expected that search engines and their own reputation would bring in the users. To a large extent this did happen: usage of data online has been very significant. However, in separating data from the online scholarly information system, data producers are not helping scholars, high level students, librarians and other stakeholders because by being apart, it makes it harder for end-users to find, access and incorporate data in their workflows.

15. If data producers remain outside the scholarly information system they will lose out because scholarly audiences may make do with alternative data sources that can be found inside the system. To maximise accessibility to data, producers should consider how to integrate their data services back into the system.

16. The benefits of integrating data outputs in the scholarly information system are significant for all stakeholders: users will find it easier to find and cite datasets; librarians (and other intermediaries) will find it easier to catalogue and index datasets; publishers will find it easier to link to datasets from reference listings in analytical publications. With the result that data producers will maximise the impact and usage of their data among the scholarly community and among high-level students.

17. The scholarly information resource system relies on a variety of publishing metadata standards to function. Metadata standards exist for journals and books and without them the system would be impossible to build on any scale. Today, data is mainly published online and the old common denominator, print books, has largely gone. With the demise of printed books have gone the accompanying metadata standards inherent with that publication type. This means that there is a lack of publishing metadata for the new online format and without metadata; data can't be incorporated in the scholarly information system.

18. To solve this problem, OECD has proposed a metadata standard for online data resources that is compatible with the scholarly information system[6]. OECD has since built a new publishing platform, OECD iLibrary, which puts the proposed standard into practice[7].

19. Users of OECD iLibrary will find every data object (dataset or table) has a unique homepage that has a persistent DOI identifier. From this homepage, and from inside each dataset, users can find a 'cite this dataset' button. When clicked, the user sees a pop-up box from where they are invited to download a fully-formatted citation that is available to be downloaded in one of common formats used by scholars' citation management tools. Users can then integrate this citation into any reference list they are building when they write a paper or chapter in a book. Because the citation is structured to the same standards as journal articles and books, publishers will be able to generate links in the online versions of the articles and books they publish, via the CrossRef system.

[6] Green, T. (2009), "We Need Publishing Standards for Datasets and Data Tables", *OECD Publishing White Papers*, OECD Publishing. doi:10.1787/787355886123

[7] http://www.oecd-ilibrary.org

20. By adapting the existing publishing metadata standards used in the scholarly information system, the OECD's proposed citation standard for data objects can be used by all stakeholders in the scholarly information system without any re-engineering or adaptation of their existing workflows or processes.

21. Once OECD iLibrary is fully launched (later in 2010), OECD's datasets will be integrated into the scholarly information system once more, alongside data and analytical publications. To speed the process, OECD will provide intermediaries with structured publishing metadata to populate their part in the system, starting with the provision of MARC[8] records to scholarly librarians.

22. Special search engines, Google Scholar, Scopus, Web of Science, will be invited to work with OECD iLibrary so that datasets will start to appear in scholarly search results, alongside journal articles and books. The popular search engines, Google, Yahoo! And Bing, will also find the structured homepages for each dataset in OECD iLibrary very easy to crawl and index.

23. The dataset format is not relevant: any OECD dataset regardless of format will be 'wrapped' with appropriate publishing metadata when loaded into OECD iLibrary and thereby be included in the scholarly system.

IV. CONCLUSION

24. Whereas in the pre-Internet era, data publications were included in the scholarly information system thereby benefiting scholars, students and information professionals such as librarians, the online datasets have not, as yet, been included in the online version of the scholarly information system. This is making it hard for scholars, students and information professionals to locate, cite, manage and link to online datasets.

25. OECD believes that to be fully used and reach a maximum audience in research and academic settings, online datasets must be wrapped with publishing metadata to similar standards found in electronic versions of journals and books.

26. OECD has elaborated a proposed standard for publishing metadata for datasets and has built a new publishing platform, OECD iLibrary, which puts the proposed standard into practice. As this new platform is rolled out during the second half of 2010, OECD will start the process that will integrate its online datasets into the scholarly information system alongside its books, working papers and journals.

27. OECD believes that in doing so, its datasets will be easier to find, easier to cite, easier to link to and easier to manage than they are today. In the long run, it believes that usage and impact will be greater too.

[8] MARC: Machine Readable Cataloguing records that are used by library systems worldwide.