CONFERENCE OF EUROPEAN STATISTICIANS WP.5-2

Workshop on Statistical Data Dissemination and Communication 12 June 2017

28-30 June 2017, Geneva, Switzerland

Integrating Data Collections with Open Data Resources: a communication-centred approach

Thomas Bourke (European University Institute, Florence)

EconLibrary@EUI.eu

This presentation describes recent developments in the provision of data services from the perspective of a research library. Some of the issues concerning re-dissemination and communication in a community of advanced users may shed light on the overall relationship between data providers and users. For the purposes of this presentation, the term 'data' includes statistical and other forms of data.

The European University Institute, Florence, is a postgraduate institution established by the European Union member states, dedicated to research in economics, law, sociology, political science and history.

There are multiple 'players' in the international data community: official statistics producers; data publishers (commercial); data users; researchers as data-creators; data repositories; the ICT community, and the library community.

For research libraries there are two important priorities (i) the centrality of the data *user* and (ii) the necessity to index, disseminate, communicate and support data literacy for a *heterogeneity* of resources.

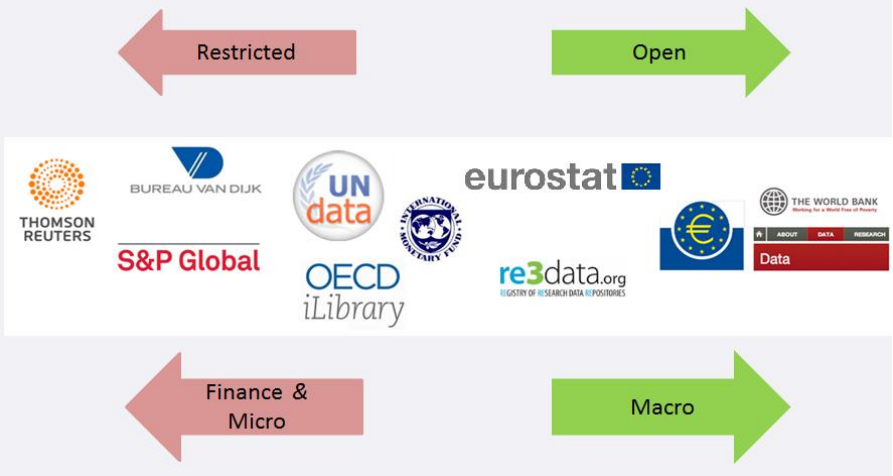
From a data user's perspective, there are two paramount considerations: (i) relevance and (ii) quality. Scholars are rarely interested in the dissemination model used by the data producer – eg. whether the data requires payment, or is distributed on an open basis. Researchers are interested in finding relevant and quality data for their research project.

Libraries are therefore in constant contact with multiple stakeholders: (i) official statistics providers – including providers of micro-socioeconomic data, (ii) commercial publishers of databases and (iii) the research data repository community (sometimes referred to as the 'open data community').

Research libraries have to manage different data access models: (i) open and free – eg. official statistics (ii) open, with 'fremium' option for added services (iii) restricted access and free – eg. micro-socioeconomic data and (iv) restricted access and paid – eg. commercially published financial databases.

Restricted – free eg. micro data	**Open – free** eg. official statistics
Restricted – paid eg. financial data	**Open – paid** eg. fremium

It is possible to place major suppliers on a spectrum broadly indicating access from 'restricted' to 'open' (below). What can be observed is that macroeconomic data is largely available on an open, non-subscription basis, while financial data is largely available under subscription model.



Two of the most important topics of the past decade – the Financial Crisis and socio-economic inequality – lie at the more 'restricted' end of the spectrum. For research libraries, access to quality financial data is expensive. Access to micro data is usually free of payment – but is labour intensive in terms of contract management, data protection and infrastructure for internal re-dissemination.

There are two main categories of open data, and both are growing steadily: (i) open public data and (ii) open research data. Research libraries are actively involved in indexing, re-disseminating, providing data literacy training and preserving both open public data and open research data. At the same time, research libraries continue to invest financial and human resources in the acquisition, indexing, re-dissemination and infrastructure for commercially-sourced financial databases and publicly-sourced micro-socioeconomic data resources. No significant change in the latter two categories is envisaged. This is due to (i) publishers' database copyrights; and (ii) data protection provisions pertaining to access to, and use of, micro data observations about persons, families and households.

Research libraries have a traditional role with regard to the acquisition, indexing and re-dissemination of data. More recently – universities have become active in the world of research data management and open research data. These activities are mostly – but not always – undertaken by the university library. However, the 'traditional' data collection role is not always undertaken in tandem with the 'new' research data management role.

The European University Institute merges the two roles – and advocates this as an appropriate way to create synergies for data users. This is because – from a data user's perspective – relevance and quality are more important than the dissemination model of the data producer.

The 'traditional' library data collection role is linear:

Acquisition/ deposit	Serials maintenance	Paper / CDs DVD / Databases	Cataloguing/ classification	Terms & conditions of access and use	Reference support	Preservation

Newer university research data roles (usually undertaken by the research library) have a more cyclical dynamic:

Important features of the new data ecosystem include: 'mixed' data portals which combine both paid and free resources; the recognition that researchers are data-creators; the provision of research data management services including data management plans; metadata assistance; data literacy training and the maintenance of data repositories for preservation and sharing of data outputs.

EUI Library data strategy has three pillars: (i) The Data Portal (ii) Data Services and (iii) the EUI's new ResData repository. There are seven thematic components:

- Data discovery
- Terms and conditions of use

- Support, software and infrastructure
- Research data management and data management plans
- Data management in EU Horizon 2020
- EUI ResData repository, preservation and open data
- Qualitative data in humanities and social sciences

Universities retain their roles as data collectors and re-disseminators – but have new roles in support of data elaboration, management, repositing and sharing.

Some of the issues concerning re-dissemination and communication in the research community may shed light on the overall evolution of the relationship between data providers and users. There are three aspects: (i) communication, (ii) dissemination and (iii) infrastructure.

Based on the evolution of the research library experience of creating synergies in the provision of collections and services; it is recommended that there be a greater dialogue between these three international communities:

- Official statistics producers (including micro data providers)
- Commercial data publishers
- The research data / open data community

This dialogue could produce new modes of dissemination and re-dissemination of data.

For example, would it be possible to create an international registry – a union catalogue – of quality data resources, comprising (i) official statistics providers (ii) abstracts of commercial data resources and (iii) research datasets?

Such a union catalogue could use advanced metadata and semantic web indexing to create synergies between data resources – for example linking primary databases and research-generated derivative datasets and citations.

On European Statistics Day, 20 October 2016, the then Director-General of Eurostat, Dr. Walter Radermacher noted in Budapest, that Eurostat and the broader European Statistical System "see quality as the basis of our competitive advantage in a world experiencing a growing trend of instant information which often lacks the necessary proof or quality."

This concern for quality – supported by strong branding – is shared by all three international communities: (i) official statistics providers (ii) commercial data publishers and (iii) the new research repository community.

Could the recent experience of research libraries be a model for greater collaboration between these three communities?