Наводя мосты между словами и числами: внедряя данные в книги и статьи

Building Bridges between Words and Numbers: Integrating Data into Books and Papers

Зводячи мости між словами і числами: інтегруючи дані в книги та статті

Ален Деймон
Организация по экономическому сотрудничеству и развитию (OECD), Берлин, Германия

Damon Allen
OECD – Organisation for Economic Co-operation and Development, Berlin, Germany

Деймон Аллен
Організація з економічної співпраці та розвитку (OECD), Берлін, Німеччина

Доступ к необработанным данным становится все важнее для исследователей. Но как лучше всего включить необработанные данные (статистику) в анализ этих данных (книги, статьи и т.п.)? Как организация, занимающаяся изданием и данных, и их аналитической обработки на протяжение более 50 лет, Организация по экономическому сотрудничеству и развитию (OECD) имеет уникальную возможность построить мосты между словами и числами. В данном докладе представляется решение, созданное OECD – библиотека iLibrary, в которой слова и числа взаимосвязаны, что позволяет читателю легко переходить с одного уровня контента на другой, загружать его, работать с ним и цитировать его.

Access to raw data is becoming more and more important for researchers. But how can the raw data (statistics) best be integrated into the analysis of that data (books, papers etc)? As a publisher of both data and analysis for over 50 years, OECD is in a unique position to build bridges between words and numbers. This paper will present OECD's solution - the OECD iLibrary - in which words and numbers are interlinked, allowing readers to easily navigate from one layer of content to another and to download, work with and cite that content.

Доступ до необроблених даних є надзвичайно важливим для дослідників. Але як найкраще включити необроблені дані (статистику) в аналіз цих даних (книги, статті тощо)? Як організація, що займається виданням даних і результатів їх аналітичної обробки понад 50 років, Організація з економічного співробітництва та розвитку (OECD) має унікальну можливість побудувати мости між словами і числами. У цій доповіді пропонується рішення, створене OECD – бібліотека iLibrary, в якій слова і числа взаємопов'язані, що дозволяє читачеві легко переходити з одного рівня контенту на інший, завантажувати його, працювати з ним і цитувати його.

Context

Jim Gray from Microsoft Research talks about the "Fourth Science Paradigm": A thousand years ago science was *empirical*, describing natural phenomena. In the last few hundred years it became *theoretical*, using models, generalizations etc. In the last few decades it has become *computational*, simulating complex phenomena. And today, it is characterized by *data exploration*. The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Only then does the scientist analyse the database or file using data management and statistics.

Science has become extremely data intensive, with the size of many datasets now measured in terabytes or petabytes rather than megabytes or gigabytes. The challenge for scientists and publishers alike is – how to best publish the raw research data? And how can publishers help users by making research data discoverable, comprehendible and usable?

Data publishing challenges at OECD

OECD has been publishing social and economic data since the Organisation began in 1961, initially in the form of statistical books and periodicals. With the arrival of electronic publishing in the late 1990s, OECD was one of the first to make all its content accessible online. When it came to rebuilding its online platform in 2009/2010, perhaps the single most important goal was to increase the discoverability and usability of its data.

To address the challenge of better integrating data and analysis, we identified the following must-do tasks:

- Create publishing metadata at a more granular level so that smaller and more relevant pieces of information can easily be searched and found, whether data or analysis.
- Create statistical metadata so that the numbers are never delivered without the textual environment making them understandable and usable.
- Link, as much as possible, analytical content to the statistical evidence supporting it.
- Adapt content to the different audiences and levels of specialisation.
- Deliver the content in the most user-friendly environment.

OECD's solution

Launched in 2010, OECD iLibrary is the global knowledge base for OECD's statistics and publications. It fully integrates OECD data and analysis thus facilitating the discovery process. OECD iLibrary contains all books and papers published since 1998 (over 10,000 full text publications) – as well as a vast collection of statistics, with data going back to the early 1960s (300 datasets with 4 billion datapoints).

Users can search and click directly to chapters, articles, books, tables, datasets and databases. These content items are available in different formats and languages. Supported by significantly enriched metadata, and the systematic use of Digital Object Identifiers (DOIs), users can find and cite individual tables, graphs and datasets as easily as articles and chapters. The search function allows quick access to all content types (books, chapters, journals, articles, tables, datasets, etc) in any available content format: PDF, WEB, XLS, DATA, ePUB.

Different formats for different users

While access to raw data is important for high-level researchers, other types of user (undergraduates, non-specialists) find this data too complex and are looking for simplified data on key topics of interest. OECD iLibrary therefore provides entry to three types of statistical content according to the profile of the user and the format of information required.

- **Statistical databases:** For experienced users wanting to access the latest available raw data. Users can build tables and extract data from across databases as well as work within individual databases. Extractions of up to a million datapoints can be made in a range of formats. Each database includes detailed metadata to help people understand the numbers.
- **Key tables:** The most popular and frequently requested indicators have been formatted for users to copy and paste straight into their reports. Accessible in formats to suit the needs of the busy user, more than 100 socio-economic and environmental indicators are available. These tables are easy to use and are available in three different formats: WEB, PDF and XLS.
- **Statistical books:** Provides the archive version, essential for users wishing to check older numbers, with a back catalogue to 1998. Each table is available in three different formats: WEB, PDF and XLS to facilitate discoverability and usability of the data.

Creating a user friendly environment

Content in Context: On OECD iLibrary every single item, from a complete database down to a table in a book chapter, has its own home page, DOI and bibliographical metadata. The contextual navigation allows users to easily locate the original source data for a publication, related publications or data, or even tables within a title.

Digital Object Identifiers (DOIs): The use of DOIs for all published items in OECD iLibrary allows the use of persistent and independent links to the described content anywhere on the internet. Information about a digital object may change over time, including where to find it, but its DOI name will not change and thus always points to its original online presence.

Citation tools: All citation tools in OECD iLibrary feature the DOI name as the reference link. A new "Cite as" feature allows consistent referencing of any full-text item on the site. As an innovative tool, this feature also includes the citation of dynamic databases with a link to the original dataset and the exact time at which it has been accessed. The citation tool is compatible with popular bibliographical reference software such as EndNote, RefManager, ProCite and RefWorks.

Exporting embedded data: Most OECD books now include "StatLinks" - a unique service which allow readers to download Excel versions of charts, tables and statistical annexes in OECD publications. A dynamic link (StatLink) is provided for each table, directing the user to a web page where the corresponding data are available in Excel format.

Conclusion: Data and publications belong together

OECD has unequivocally established that data makes publications discoverable and vice-versa. Publications also gain depth by supplying data. Our experience is that the addition of research data to a publication can add great value, but that a publisher can add even greater value to the process by ensuring the following:

- Careful treatment and storage of research data
- Creation of appropriate metadata
- Adaptation of data for different audiences
- Usage of bi-directional links and persistent identifiers (DOIs)
- Addition of additional features (e.g. OECD StatLinks)
- Establishment of citation practice