Please cite this paper as:

Green, T (2009), "We Need Publishing Standards for Datasets and Data Tables", *OECD Publishing White Paper*, OECD Publishing.
doi: 10.1787/603233448430
http://dx.doi.org/10.1787/603233448430

OECD Publishing White Paper

We Need Publishing Standards for Datasets and Data Tables

Toby Green, OECD Publishing, Paris
toby.green@oecd.org

REVISED VERSION (FEBRUARY 2010) AVAILABLE AT:
doi: 10.1787/787355886123
http://dx.doi.org/10.1787/787355886123

TABLE OF CONTENTS

We Need Publishing Standards for Datasets and Data Tables

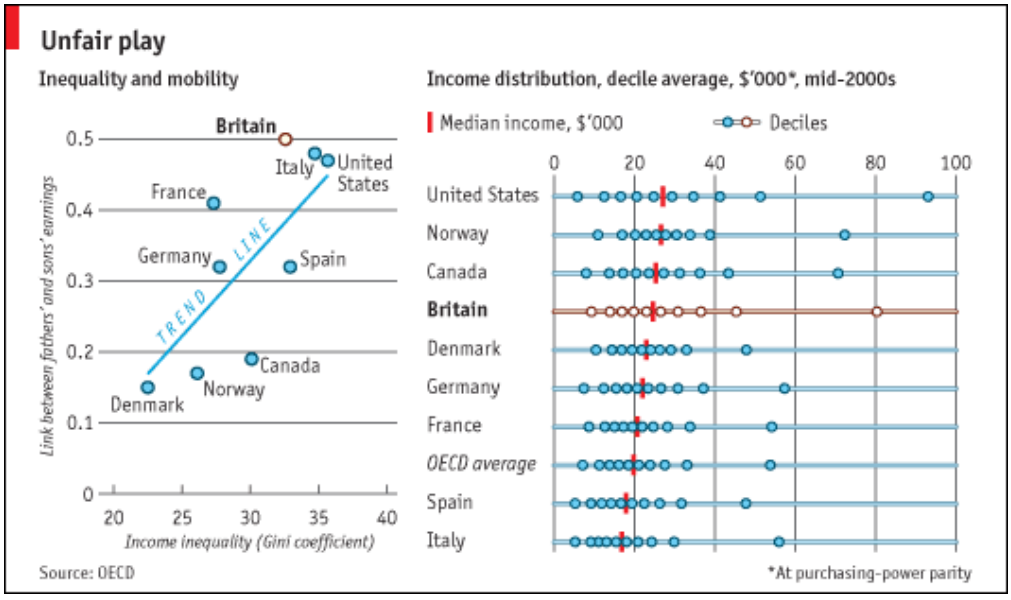
Go to Google.com and search for 'journals'. The first page of results is full of references to professional publishing websites such as OUP, PubMed, AMS et al. Just what one would expect.

Go to Google.com and search for 'datasets'. On the first page of results you'll find only two references to anything that might be professional – a link to the US Census website and a link to some data posted by researchers at the World Bank. (Interestingly, this World Bank site is not their famous World Development Indicators website). The other results are a mish-mash of poorly presented and poorly maintained pages from universities and other research bodies. None of these sites is presented in a professional manner and each has its share of broken links.

Research is all about gathering data. Academic papers, journal articles and monographs cannot be written without data. Before the Internet, data could not be made available easily, but now datasets are being posted on departmental websites in universities and research centres around the world. But can you find them? Will they be there tomorrow? Judging from the Google search, the answer is not positive.

Despite not appearing in the Google search results, publishers are getting involved. According to a recent ALPSP report on scholarly publishing practice, 45% of journal publishers said they provided access to data sets associated with the journal articles they publish1. Since 160 publishers replied to this question, this means that at least 72 journal publishers report they are handling data.

However, dig a little deeper and it's easy to see there are no rules about how to publish, present, cite or otherwise catalogue datasets. Consider this chart published by The Economist. Published as part of a piece called 'The importance of fairness in an economic downturn' it gives the reader no clue beyond "Source: OECD" as to which of the 298 datasets OECD publishes it came from.



[1] Scholarly Publishing Practice 3, ALPSP, London, 2008 ISBN: 978-0-907341-40-6

A search of ScienceDirect shows that many authors are using OECD data in their research. Yet when they need to cite OECD data, they cite OECD's data in wide variety of ways. They point to:

- OECD's print editions (which often have extensive statistical annexes) rather than the original dataset.
- OECD's main website (www.oecd.org)
- Nothing more than "OECD"
- Pages deep in OECD's website, none of which use persistent links

The example below is from a recent issue of Elsevier's journal *Economic Modeling*.

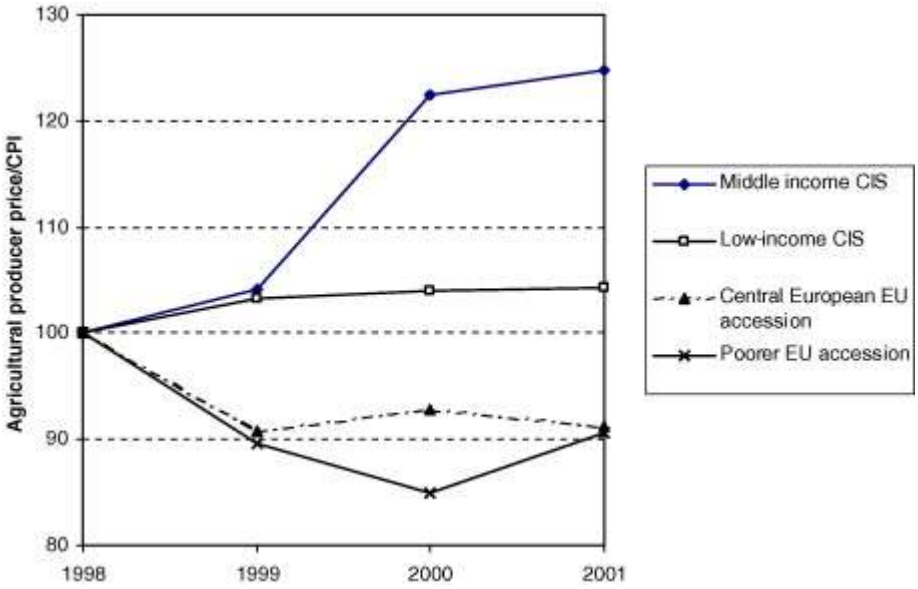


Figure 3. Change of relative agricultural producer prices since 1998. Middle-income CIS show average for Russia, Kazakhstan, and Ukraine. Low-income CIS show average for Azerbaijan, Kyrgyzstan, Moldova. Central European EU accession countries show average for Czech Republic, Estonia, Hungary, Poland, Slovakia, Slovenia. Poorer EU accession countries shows average for Bulgaria, Latvia, and Lithuania. Source: OECD, 2004 and CIS Statistics, 2003.

In another example, an author publishing in the journal *World Development* went to great trouble listing all the data sources in an extensive Appendix. However, some of the items listed would not be of much help to a reader:

> *Main mortality estimate: Estimated settler mortality. Settler mortality is calculated from the mortality rates of European-born soldiers, sailors, and bishops when stationed in colonies. It measures the effects of local diseases on people without inherited or acquired immunities.* **Source*: Acemoglu* et al. (2001), based on Curtin (1989) and other sources**.
>
> *Tertiary school enrollment: School enrollment, tertiary (% of gross).* **Source*: Barro and Lee (2000) and their databases.***

This is not to single out Elsevier. A similar result is found in Repec[2], the world's largest collection of papers in economics and on other publishers' e-journals sites. ESDS International[3], a UK data aggregation service, asks its users to cite OECD data as follows:

> *Citation information*
> *Publications based on ESDS data collections should always acknowledge the data source with a citation in the reference section. The bibliographic citation for this database is:*
> *Organisation for Economic Cooperation and Development, <Dataset name>, ESDS International, University of Manchester*

Not only is there no linking URL provided, the same data now risks being cited in two different ways depending on where the user sourced the data: one way via OECD's own services, another via ESDS'.

In view of the different advice received, it is no surprise authors and publishers are so unsure about how they should cite data sources. But what about librarians – are they doing a better job managing and curating data sets?

Librarians take the trouble to catalogue each and every book and journal they acquire – thousands of titles a year. Yet when it comes to datasets the situation looks very messy. Leaving aside a confusion about the meaning of the term 'dataset' (some librarians list IngentaJournals, Cambridge Scientific Abstracts, and Inspec as 'datasets'), a quick trawl through some library OPACs shows that none have managed to catalogue all 298 datasets published by OECD. In most cases, OECD's datasets are bundled together under the brand name we use for our online publishing platform, SourceOECD. In some cases the link is to the OECD's Statistics Department website, not the publishing platform where the actual datasets are housed. Again, this is not unique to OECD. Other datasets are either left uncatalogued (how many librarians have catalogued the 'Delve datasets' from University of Toronto[4]?) or are bundled together in unstructured lists in some corner of a library website. Since datasets are so difficult to catalogue, what chance is there for federated search tools to discover datasets alongside articles and book chapters?

None of this is to criticise authors, publishers, aggregators or librarians. They can hardly do better in the absence of an accepted system for how datasets should be cited and catalogued – but now that datasets are becoming widely available and so many publishers are beginning to get involved (if only to provide links from their journal articles) there is a need for a bibliographic system to help authors cite datasets and for librarians to catalogue them.

In today's world, does this matter? Surely everyone will find what they need from for a general search like Google?

---

[2] www.repec.org

[3] www.esds.ac.uk/international/introduction.asp

[4] http://www.cs.utoronto.ca/~delve/data/datasets.html

But do they? In Inger & Gardner's white paper on how readers navigate to scholarly content[5] they found that:

- When a reader already has a reference or citation and wishes to read the article online, a general web search engine ranks fourth behind specialist bibliographic search engines, library web pages and journal homepages as the starting point.
- When keeping up to date with the latest issues of journals, email alerts prove to be the most used, followed by visiting journal homepages and library websites ahead of general search engines.
- When searching for unknown articles on a specific subject, they prefer to search specialist bibliographic databases to general search engines (like Google) and rank library web pages as highly as journals gateways as starting points.

Clearly, if data providers rely on a post-it-and-Google-will-find-it approach, they will miss out on a great deal of traffic from readers who are using alternative routes to discover content.

Proof that readers want to access underlying data

OECD launched a service called StatLinks in 2004. The concept is simple. Under each table, chart and graph appearing in an article or book chapter, a DOI (Digital Object Identifier[1]) link is printed alongside the traditional 'Source' legend. By following the DOI link, readers are able to download a spreadsheet containing the data used to create the table, chart or graph. By 2008, OECD had put 20,000 StatLinks into its publications and in 2008 alone, 980,381 spreadsheet files were downloaded. Proof, if it were needed, that readers do take the chance to get hold of original data when it's offered.

As e-journal publishers know, and as e-book publishers are finding out, readers want everything to be connected. Readers don't want to find e-books in different e-silos from e-journals. They want all scholarly content to be interconnected via bibliographies and reference listings and that includes the underlying datasets.

OECD is not the first to recognise this. Altman and King[6] proposed a citation standard for scholarly data in 2007 but they did not include cataloguing metadata in their proposal.

In the UK, librarians are getting interested. UK Research Data Service is working on a feasibility study which will recommend a national data management service. Research Libraries UK and the IT directors from the Russell Group of Universities are bidding for funding from HEFCE and JISC to undertake the study. As Jean Sykes, librarian and director if IT services at London School of Economics noted, "The data underlying a research project can be extremely important, and can be regarded as part of the research output just as much

[5] Inger S., Gardner T, How Readers Navigate to Scholarly Content, 2008 http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf

[6] Altman M., King G., A Proposed Standard for the Scholarly Citation of Quantitative Data, D-lib Magazine, March/April 2007, Vol 13 No.3/4

as the finished experiment or publishing article. Research data . . is an invisible asset . .". In Australia, a similar study is already underway[7].

Researchers themselves understand the need to share their data. Aside from the World Bank researchers noted at the beginning of this paper, scholars at Oxford University were polled in the summer of 2008 for their views. The survey showed that informal mechanisms for sharing data were in place, but they were patchy and uneven across disciplines. A similar survey at Bristol University showed the same need for better management of data repositories and decisions on what needs to be stored and how it will be indexed[8].

In Germany, to enable citations and better retrievability of data, the German Research Foundation (DFG) created a project on Publication and Citation of Scientific Primary Data. Starting with the field of earth science, the German National Library of Science and Technology (TIB) has now established itself as a DOI-registration agency for scientific primary data and claims to have registered 500,000 datasets[9].

TIB has now gathered a group of Europe's leading research libraries and technical information providers to establish an as yet un-named partnership to improve access to research data on the internet. The German National Library of Science and Technology (TIB), the British Library, the Library of the ETH Zurich, the French Institute for Scientific and Technical Information (INIST), the Technical Information Center of Denmark and the Dutch TU Delft Library all signed a Memorandum of Understanding to this effect during the meeting of the International Council for Scientific and Technical Information (ICSTI) in Paris on 2nd March 2009. Their goal is to establish a not-for-profit agency that enables organisations to register research datasets and assign persistent identifiers so that research datasets can be handled as independent, citable, unique scientific objects[10].

So, it seems the time is right to propose a bibliographic standard for datasets.

OECD's vision is to make data outputs as accessible and as easy to find and use as written outputs like working papers, journal articles and book chapters. Moreover, the vision is to make all published outputs compatible with and discoverable from all scholarly publishing and discovery systems. This means it should be easy for an author, writing a paper for a journal, to cite a database accurately in such a way that the publisher can offer a reader a persistent link to the database without having to do anything more than they're doing today to link to articles. Equally, a librarian should be able to catalogue datasets with no more effort than it takes to catalogue and maintain records for e-serials. Readers who like to discover content via specialist bibliographic databases or other specialist websites should also expect to find references to datasets among references for articles, working papers, books and so on. In short, datasets must be as discoverable and as linkable as any scholarly publication.

Inger and Gardner[11] studied eleven different discovery channels for e-journal articles ranging from the informal (listing on author's departmental page) to the formal (specialist bibliographic databases). Since each

---

[7] Academia's Buried Treasure, Information World Review, October 2008

[8] Academia's Buried Treasure, Information World Review, October 2008

[9] http://www.tib-hannover.de/en/special-collections/research-data/

[10] http://www.tib-hannover.de/en/the-tib/news/news/id/114/ accessed on March 25th 2009

[11] Inger S., Gardner T, How Readers Navigate to Scholarly Content, 2008
http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf

of the eleven scored in terms of being helpful to readers in discovering journal articles, to achieve the vision outlined above, a parallel system for data needs to be considered which should be compatible with that used for journals and books.

Inger & Gardner's list of discovery channels for e-journals	**Proposed equivalents for datasets**
Specialist bibliographic database	The same – requires bibliographic metadata for datasets to be provided to these databases in a compatible format (e.g. Onix)
Library web pages	The same – requires bibliographic metadata for datasets to be provided to these databases in a compatible format (e.g. MARC 21 or Dublin Core)
Specialist, subject-specific portals	The same – requires bibliographic metadata for datasets to be provided in a compatible format
Key research group website	The same – unlikely to use a bibliographic standard, so simply requires persistent links to the dataset or a widget to be posted
Department webpage on institutional website	The same – unlikely to use a bibliographic standard, so simply requires persistent links to the dataset or a widget to be posted
Publisher's website	The same – if a formal catalogue of other publications is presented, requires bibliographic metadata for datasets in a compatible format; otherwise persistent links to the dataset or a widget to be posted
Email based alerts	The same – requires bibliographic metadata for datasets with a suitable date-driven trigger when updates are made available
Journal homepage	Database homepage would be the equivalent. This page needs to have its own persistent URL.
Journals gateway[12]	The same – requires bibliographic metadata for datasets to be provided to these gateways in a compatible format
General web search (e.g. Google, Yahoo!)	The same – higher rankings will result if bibliographic metadata for datasets is exposed to search engines
Scholarly Society website	The same (if appropriate) – if a formal catalogue of other publications is presented, requires bibliographic metadata for datasets in a compatible format; otherwise persistent links to the dataset or a widget to be posted

[12] Examples are IngentaConnect, SwetsWise, EBCSO Host et al

As the above table shows, if bibliographic metadata is prepared to the same standard, most of the discovery channels for e-journals can easily be exploited by datasets. In fact, OECD has been exploiting some of these channels for its datasets since 2001 – often without the channel owner realising what is going on. An example is journal gateways - users of IngentaConnect, and SwetsWise can discover OECD datasets among the journal articles. This was achieved by simply producing metadata for each dataset to the same style and standard as used by journals and sending it to these gateways along with OECD's journals metadata. All done without Ingenta or Swets realising because the metadata structure was the same as that used for journals.

Looking beyond discovery, what other requirements are there for a bibliographic standard for datasets?

Cross-referencing. Launched in 2000, CrossRef now has more than 35.6 million links to scholarly materials registered from 2,740 publishers[13]. The bulk of this content is scholarly articles from journals, but CrossRef's mission is to provide the citation linking backbone for all scholarly information in electronic form. This clearly encompasses datasets. The benefits are straightforward. By using CrossRef, dataset publishers can interweave datasets into the scholarly information network alongside journal articles and book chapters. Authors will be able to cite datasets they've used in their manuscripts, confident that their publisher will be able to not only find the link, but maintain the link into the future. Readers will be able to click on references to cited datasets, not just to cited articles, with confidence. Forward linking, now common in journal publishing, would be possible, so users could find published research articles and book chapters which draw on a particular dataset.

Library systems. Books and journals have been catalogued by librarians for generations and international standards have emerged to make improve the process. Library catalogues are now online, and as Inger and Gardner have shown, they are still used intensively as discovery tools by readers. The main bibliographic standard for library systems is MARC (MAchine Readable Cataloguing). This is a format standard for the storage and exchange of bibliographic records and related information in machine-readable form. In recent years, ONIX has emerged as the international standard for representing and communicating book industry product information in electronic form. ONIX for serials also exists. The key message here is that if standards can help streamline the cataloguing of books and journals, why can't they be extended to include datasets too?

Taking these requirements into account, OECD is proposing to implement a metadata standard for publishing datasets, collections of datasets and individual tables. The detailed proposal can be found in an annex to this paper, but a summary can be seen in the table below – and it also shows how OECD's proposal compares with Altman and King's 2007 proposal.

Proposed Dataset Publishing Metadata Fields	OECD Proposal	Included in Altman & King's proposal?
Unique, persistent, global identifier e.g. DOI *(digital object identifier of the content)*	**Mandatory**	Yes
Main Title *(in all available languages)*	**Mandatory**	Yes
Subtitle *(in all available languages)*	**Optional**	No

[13] Source: www.crossref.org on 25th March 2009

Author(s): **-** *first name* *- last name* *- affilitiation*	**Mandatory**	Yes
Publication date *DD-MM-YYYY*	**Mandatory**	Yes
Next publication date (3 fields) *- Year, Day (when available ,) Month (when available)* *to show when the dataset is next due to be updated*	**Mandatory**	No
Periodicity *(e.g. Annual, Monthly, Quarterly, etc.)* *Some datasets have a regular update frequency – shown here*	**Mandatory**	No
Languages of the content (ISO Codes): *Useful when publishing datasets in more than one language*	**Optional**	No
Size: number of cells (algorithm to calculate size must be defined) *Useful to show total file size*	**Optional**	No
Countries covered: *When the dataset is country specific, this relates to the Country ISO Code of the related country*	**Optional**	No
Period covered: Start Year	**Optional**	No
Period covered: End Year	**Optional**	No
Variable Index *- A classification with variable titles*	**Mandatory**	No
Short abstract	**Mandatory**	No
Long abstract	**Mandatory**	No
Keyword(s)	**Optional**	No
Classification *- for example JEL classifications in the case of economics*	**Optional**	No
Belongs to (parent-child relationship): *Links a dataset to any other parent dataset from which it is associated. This is useful when a dataset is released which is a sub-set of a larger dataset.*	**Optional**	No
Has Main Parent *For a dataset belonging to more than one larger dataset, links the dataset to its main parent dataset (as opposed to other "step" parent datasets.*	**Optional**	No
Has Physical Form Link to dataset in various output formats (XLS, PDF…) *This allows a dataset to be released in different file formats*	**Mandatory**	No
Is related to *Relates complementary products to facilitate linking.*	**Optional**	No

External links *Relates objects to external web pages, e.g. author's website.*	**Optional**	No
Supersedes *Indicates that a dataset has replaced an earlier version*	**Optional**	No
Is continued by *Indicates that a dataset has been discontinued and is replaced by a new one.*	**Optional**	No
Is Imprinted By *Indicates the legal / organisational body which owns the dataset at imprint level*	**Optional**	No
Is Copyrighted By *Indicates the legal / organisational body which owns the publication at copyright level*	**Mandatory**	No
Universal numeric fingerprint *Method for showing if a dataset has changed in any meaningful way since it was initially released* *In OECD's case this would be managed by the Supersedes/ Continued by fields*	**Not in OECD's proposal**	Yes

If these fields are stored in an XML or relational database, exporting them to discovery channels and library systems in standard file formats such as ONIX or MARC 21 is a simple procedure. Many of the fields are not needed for citation purposes or for cataloguing but they would be invaluable for users when they come to the dataset's homepage.

Similarly, it is a simple step to create downloadable citations for use in the common bibliographic management systems used by authors, e.g. RefWorks and EndNotes.

OECD is proposing to cite datasets in the following way:

· <Author> (<Year of Publication Date>), "<Dataset title>: <Dataset subtitle>", <*Parent dataset collection title*> (database)
<doi>
<doi link>
(Accessed on <date>)

Which would give:

· OECD (2008), "Social Expenditures aggregates", *OECD Social Expenditure Statistics* (database).
doi: 10.1787/000530172303
http://dx.doi.org/10.1787/000530172303
(Accessed on 21 December 2008))

In addition to datasets, OECD is proposing standards to cite tables. This is important since so many tables, rather than the original datasets, are used as source data by authors. OECD is proposing the following ways to cite tables:

When the Table comes from a publication (e.g. book):

- <Author>, <Year of Publication Date>, <Table number>.<Table title>:<Table subtitle>, in <Publication title>: <Publication subtitle>, <publisher>
  <doi>
  <doi link>

Which would give:

- Smith, J. (2008), Figure 1.2. Broadband Penetration in OECD Countries, in *OECD Communications Outlook 2008*, OECD Publishing
  doi: 10.1787/000530172303
  http://dx.doi.org/10.1787/000530172303

When the Table comes from a stand-alone series of tables rather than a publication:

- <Author>, (<Year of Publication Date>), "<Table title>:<Table subtitle>", <Table series>, <Table Order number>
  <doi>
  <doi link>

Which would give:

- OECD (2008), "Health Expenditure in OECD Countries", *Health Key Tables Series from OECD, No 5*
  doi: 10.1787/000530172303
  http://dx.doi.org/10.1787/000530172303

Dynamic Datasets

Many datasets are being updated on a rolling basis, adding new data as and when received. Occasionally, revisions are made to the entire dataset which changes the old data. All of these changes are noted and explained in the statistical metadata (i.e. the metadata which describes the data itself, rather than the publishing metadata which is used to describe the dataset). A citation, however, is supposed to link a reader back to the *same* publication which the citing author read. In the case of a dynamic dataset, linking back to the dataset as it was when an author used it to write a paper is clearly impossible. This poses a significant challenge.

The case of dynamic datasets is not the same as versioning. With versioning it is possible to track-back to earlier versions as is done with websites like Wikipedia. In the case of dynamic datasets the volume of changes can be so large or frequent to make tracking back impossible to manage.

OECD has discussed this issue with CrossRef and there is no immediate solution. Further discussions in the industry are needed. In the meantime, OECD will use a unique DOI to link to each dataset's homepage, dynamic or not, and use the publishing metadata to alert users to the dynamic nature of the dataset. Details of changes to the data will be found in the associated statistical metadata.

Static Tables

OECD already publishes a large number of tables as part of its publications. In 2009 OECD will launch a collection of key table series, for example a series of key tables on health. All these tables share a common characteristic: they are static objects. This makes them as easily citable as a journal article. However, many

tables are updated on a monthly, quarterly or annual basis. This means they are 'serials', just like an annual publication. OECD is therefore creating 'serial' metadata for these 'series' or 'collections' so they can be managed and cited just like an annual reference work. This means authors can cite a particular table 'edition' (e.g. the 2008 edition) and the DOI link will take readers to that 'edition'. A link on that edition's homepage will offer the reader the option of clicking 'forward' to the latest, 2009, edition. Backlinks to previous editions will also be provided.

Renditions

Data objects appear in a variety of versions, or renditions corresponding to the different languages of the same original data content and/or to the different types of proposed update methods (e.g. data objects can be updated on a continuous basis where previously released data may be over-written (dynamic object); or can be updated periodically, where each update is a static, unchanging object). For example, a table may be available in two languages (English and French) as two static objects, and for each language be available in HTML, Excel and PDF formats. OECD is proposing to use a single DOI link for each table, linking to the homepage for the data object rendition (e.g. the object in English), from which the reader can choose which electronic format to download. The same citation information will be embedded in each rendition. An example of different renditions for individual tables can be seen in OECD's beta Key Tables service[14].

Conclusion

Datasets are a significant part of the scholarly record and are being published more and more frequently, either formally or informally. Many publishers are beginning to link to them from their journals and authors are trying to cite them in their articles. Librarians would like a way to manage them alongside other publications. In short, they need to be integrated into the scholarly information system so that authors, readers and librarians can use, find and manage them as easily as they do working papers, journal articles and books.

In this paper, OECD is proposing some standards for citing and bibliographic management of datasets and data tables. OECD is currently building a new online publishing platform which will host working papers, journals, books, tables and datasets. Due to be launched in mid-2009, this platform will use the standards proposed above. Librarians will be offered MARC 21 records for datasets, alongside records for OECD books and periodicals. Users of the platform will be invited to download citations for datasets and tables in a form compatible with popular bibliographic management systems. All the DOIs for the datasets and tables will be deposited with CrossRef, ready for other publishers to use.

It is hoped other data publishers will join OECD in this initiative to establish an agreed bibliographic metadata standard for datasets and data tables and that the scholarly information industry accepts that datasets are a vital part of the scholarly record.

[14] http://www.oecd.org/statistics/keytables - accessed on March 25th 2009. On that date the citation information was not appearing in the data objects since it was not planned to be added until a later date in 2009.

ACKNOWLEDGEMENTS

The author would like to thank his colleagues at OECD who have worked hard to develop these proposed standards: Pascale Cissokho-Mutter, Terri Mitton, Eileen Capponi, Matt Brosius, Jerome Cukier Maria Bjurström and Deborah Scott-Douglas.

ANNEX 1 – DATASET METADATA

Definition of Dataset

A content type (consistent set of related data such as an OECD.Stat cube) published:
* as part of a collection or
* stand-alone (in this case it can be subject to subscription and managed as a serial)
* Has a DOI (DOI suffix = data**-<Number on 5 digits>** e.g. data-00002, data-00023)

A dataset can belong to more than one collection of datasets. One collection of datasets is necessarily the **default** parent while the eventual others should be step-parents.
The **default** parent is always attached to the dataset when the dataset is accessed straightforward online (search, DOI, etc.). The step-parent is displayed only when the dataset is accessed from the Table of Contents of the step-parent > see link **Is Default Parent Of** at level of database component.
Main Metadata exist at 3 levels: conceptual, rendition, item

- This object is cited.
- This object is searchable on the iLibrary

Datasets belonging to statistical collection are cited as follows:
<Copyrightowner(s) acronym(s)> (<year of publication date>), "<dataset main title>: <dataset subtitle>", *<default parent collection main title>* (database).
doi: <doiprefix>/<doisuffix>
<doi URL>
(Accessed on <date>)
where (database) is a label to be displayed
where date is formatted as follows: dd month label yyyy

Which would give:
OECD (2008), "Social expenditure aggregates", *OECD Social Expenditure Statistics* (database).
doi: 10.1787/data-00001
http://dx.doi.org/10.1787/data-00001
(Accessed on 21 December 2008)

Which would give (in this example there is no dataset subtitle and joint copyright OECD and FAO):
OECD/FAO (2008), "World prices", A*gricultural Outlook* (database).
doi: 10.1787/data-00002
http://dx.doi.org/10.1787/data-00002
(Accessed on 21 December 2008)

Stand-alone datasets are cited as follows:
<Copyrightowner(s) acronym(s)> (<year of publication date>), *<dataset main title>: <dataset subtitle>*(database).
doi: <doiprefix>/<doisuffix>.
<doi URL>
(Accessed on <date>)

Which would give (in this example there is no dataset subtitle and one unique copyrigh owner):
OECD (2008), *OECD Telecommunications Statistics* (database).
doi: 10.1787/data-00001.
http://dx.doi.org/10.1787/data-00001.
(Accessed on 21 December 2008)

Dataset metadata	Status
ISSN ***Relevant only for stand-alone dataset which are subject to subscription*** *An ISSN is assigned for a dataset in online medium/format in a given language*	**Optional**
Dataset Code *(in OECD.Stat)*	**Optional**
DOI *(digital object identifier of the content)* Each dataset in one language should have its own DOI. - The DOI of the English dataset should resolve to the library homepage of the dataset on the English interface - The DOI of the equivalent French dataset should resolve to the library homepage of the dataset on the French interface	**Mandatory**
DOI number (required for management of DOI suffix syntax) *A DOI is assigned for a dataset in a given language*	**Mandatory**
Main Title *(in all available languages)* *given by PAC/PUB and to displayed on the iLibrary in citation, search results, homepages*	**Mandatory**
Subtitle *(in all available languages)*	**Optional**
Author(s) * Institutional author will be OECD unless external * Physical author - *first name* - *last name* - *affilitiation* - *order (when several authors - managed by OECD)*	**Optional**
Author(s) - order (when several authors - managed by OECD)	**Optional**
Publication date *(quality insurance/validation, overwritten regularly and to be used for email alerts)* *DD-MM-YYYY*	**Mandatory**
Next publication date 3 fields: - *Year* ***YYYY*** - *Day (when available)* ***DD*** - *Month (when available)* ***MM*** ***Not relevant for static one-off dataset***	**Optional**
Periodicity Label "Data are" on the SV3 page*(e.g. Annual, Monthly, Quarterly, etc.)To be expressed in number of months associated to a label (e.g. 1month and label "Monthly", 12 months/ label ""annual"; 18 months and label "18 months", etc.)* ***Not relevant for static one-off and dynamic dataset***	**Optional**
Update method - Static one-off - Static versioned - Dynamic *A dataset may be published with different update methods.* *Examples:* *- dynamic, updated continuously: MEI, STAN* *- static versioned, updated as a whole on a regular basis: SOCX* *- static one-off, updated once only:* Economic outlook	**Mandatory**

Languages (ISO Codes) of the content - English - and/or French - and/or other language *Dataset are composed of both data and statistical metadata. Statistical Metadata exist in both English and French. The language of a dataset relates to the language in which titles, column and row headings, labels and notes are displayed.* *Currently, for most of the dataset contents there are:* *• One dataset in English, being the combination of data and statistical metadata in English* *• One dataset in French, being the combination of data and statistical metadata in French*	**Mandatory**
Size: number of cells (algorithm to calculate size must be defined)	**Optional**
Related Countries *When the dataset is country specific, this relates to the Country ISO Code of the related country*	**Optional**
Period covered: Start Year	**Optional**
Period covered: End Year	**Optional**
Time range to be always formatted YYYY-YYYY for datasets (where the first year is the start year and the last year is the end year)	**Optional**
Variable Index A classification with variable titles in both English and French, although the French titles link to English content.	**Optional**
Short abstracts *(in all available languages)*	**Mandatory**
Long abstracts *(in all available languages)*	**Mandatory**
Keyword(s) *(in all available languages) when available*	**Optional**
JEL classification *(in all available languages when available)* *- order (when several JEL classifications)* *- multiple JEL classifications are possible and they are ordered by relevance*	**Optional**
Theme(s) *(in all available languages when available)* *- classification into themes (stats portal, main web site and SourceOECD)*	**Mandatory**
Image file *Different types of images may be associated to a dataset (for OECD.Stat). Images can be attached at the level of the rendition of a given format e.g.* *- type: print screen of web page delivering dataset* *- type: top banner for stand-alone dataset* *Note: as for book covers different sizes of the same image may be required for specific channels*	**Optional**
Links	**Status**
Belongs To (parent-child relationship) Link to the statistical collection it belongs to ***Not relevant for stand-alone dataset*** **Attribute of Link: Order number** of the dataset within the statistical collection (managed by Editorial) --> used to display the dataset in the statistical collection TOC.	**Mandatory** for non stand-alone dataset
Has Main Parent link from dataset to its main parent collection (as opposed to other "step" parent collections) (relevant for datasets belonging to more than one collection)	**Optional**

Has Physical Form Link to content in various output formats/media: (XLS, PDF, OECD.Stat, external link…). Relates a rendition and its physical items *- PDF, XLS, CSV, IVT format relate to a dataset filename.* *- while OECD.Stat and External link online media relate to a URL* ***OECD.Stat URL is subject to access rights.***	**Mandatory**
Is related to *Relates complementary products (e.g. analytical material), sharing the same or similar topics, all of them being recorded into Kappa.* *Relates concepts together and products together.* **+ *It should be possible to order related links of a dataset (OECD.Stat need)*** *(SourceV3 Further reading link)*	**Optional**
Has external links *Relates Kappa objects with external links.* ***+ It should be possible to order external links of a dataset (OECD.Stat need)***	**Optional**
Has Renditions Relates a product and its content at rendition level. *One rendition corresponds to a specific language version of content. So this link allows relating the same content to several language versions.*	**Mandatory**
Supersedes *Indicates that a concept or product has been changed but the content remains the same. Can be used for table versions.*	**Optional**
Is continued by *Indicates that a product has been discontinued or is no longer available, and is replaced by a new one. Links products together.*	**Optional**
Is Imprinted By *Indicates the legal / organisational body who owns the publication at imprint level (e.g. Development Center, IEA)* - A logo in different sizes is associated to each imprint. Note:as for book covers different sizes of the same logo may be required for specific channels	**Optional**
Is Copyrighted By *Indicates legal / organisational body(ies) who owns the publication at copyright level .* By default OECD	**Mandatory**
Is Edited By *Indicates the legal / organisational body who owns the publication at editorial responsibility level.* e.g. PAC, STD, etc. For each potential owner a percentage of ownership will have to be assigned (e.g. PAC 25%, STD 30%, etc.).	**Mandatory**

ANNEX 2 – COLLECTION OF DATASETS

Definition of collection of dataset

3 types of collections

1) **Collection of datasets** which belongs to a parent collection **(=sub-collection)**
* Does not have ISSN - Is NOT subject to subscription
* Has a DOI (DOI suffix built as: <CollectionAcronym>-data)

2) **collection of datasets** which does not belong to a parent collection
*Has an ISSN
*Has a DOI (DOI suffix built as: <CollectionAcronym>-data)
***Is subject to subscription**

3) **collection of sub-collections** (and in some case also datasets)
*Has an ISSN
*Has a DOI (DOI suffix = <CollectionIAcronym>-data)
*Is subject to subscription

Main Metadata exist at 3 levels:
-conceptual
-rendition
-physical

- This object is searchable on the iLibrary
- This object is not cited: *Statistical Collection is not cited as such, but the collection main title is included in the dataset citation - see dataset.*

Publishing Metadata	Status
ISSN ***Not relevant for sub-collections of datasets (case 1 above)*** All OECD statistical collections are by definition all available in electronic format (vs. Print format). However, collections in electronic format do not necessarily contain data in one unique medium (e.g. OECD.Stat) but may contain datasets in different media (OECD.Stat, IVT, PDF, XLS, etc…). The ISSN is assigned for a collection in online medium in a given language and may contain datasets in different online media (OECD.Stat, IVT, PDF, XLS, etc…)	**Optional**
DOI *(digital object identifier of the content)* Each collection in one given language should have its own DOI. - The DOI of the English collection resolve to the library homepage of the collection on the English interface - The DOI of the French collection should resolve to the library homepage of the collection on the French interface	**Mandatory**
Main Title *(in all available languages)*	**Mandatory**
Subtitle *(in all available languages)*	**Optional**
Acronym *required for DOI syntax (see in above definition of collection)*	**Mandatory**

Update method - Static stand-alone - Static versioned - Dynamic *A database collection may be published with different update methods.*	**Mandatory**
Language(s) (ISO Codes) of the content - English - and/or French - and/or other language *Relates to the language in which titles, column and row headings, labels and notes are displayed.* *Similarly to datasets, 2 distinct collections in a given language can be defined for the same collection content depending on the languages of its datasets:* *• One collection in English, composed of datasets or sub-collections in English* *• One collection in French, composed of datasets or sub-collections in French*	**Mandatory**
Short abstracts *(in all available languages)*	**Mandatory**
Long abstracts *(in all available languages)*	**Mandatory**
Keyword(s) *(in all available languages when available)*	**Optional**
Theme(s) *(in all available languages) when available* *- classification into themes (stats portal, main web site and sourceOECD)*	**Mandatory**
JEL classification *(in all available languages when available)* *multiple JEL classifications are possible and they are ordered by relevance*	**Optional**
Image file *Different types of images may be associated to statistical product:* *Images can be attached at the level of the rendition of a given format e.g.* *- type: print screen of web page delivering collection* *- type: top banner for the statistical database Collection (on OECD.Stat)*	**Optional**
Links	**Status**
Belongs To (parent-child relationship): Link to the statistical collection it belongs to ***relevant for sub-collection without ISSN belonging to a top collection*** **Attribute of Link: Order number** of the sub-collection within the parent collection (managed by Editorial) --> used to display the sub-collections the top collection TOC.	**Optional**
Has Main Parent link from sub-collection to its main parent top collection (as opposed to other "step" parent collections) **relevant only for sub-collection** belonging to more than one collection	**Optional**
Has Default Child Link to the dataset or sub-collection for which it is the default child A dataset/collection may be the "default" dataset/sub-collection of one or many collection (for OECD.Stat) This link allow identifying the FTI related to the collection and sub-collection (sourcev3 export)	**Optional**
Has Renditions *Relates a product and its content at rendition level.* *One rendition corresponds to a specific language version of acontent. So this link allows relating the same content to several language versions.*	**Optional**

Is related to *Relates complementary products (e.g. analytical material), sharing the same or similar topics, all of them being recorded into Kappa.* *Relates concepts together and products together.* ***+ It should be possible to order related links of a collection (OECD.Stat need)***	**Optional**
Has external links *Relates Kappa objects with external links.* ***+ It should be possible to order external links of a collection (OECD.Stat need)***	**Optional**
Supersedes *Indicates that a concept or product has been changed but the content remains the same. Can be used for table versions.*	**Optional**
Is continued by *Indicates that a product has been discontinued or is no longer available, and is replaced by a new one. Links products together.*	**Optional**
Is Imprinted By *Indicates the legal / organisational body who owns the publication at imprint level (e.g. Development Center, IEA)* A logo in different sizes is associated to each imprint. Note: as for book covers different sizes of the same logo may be required for specific channels.	**Optional**
Is Edited By *Indicates the legal / organisational body who owns the publication at editorial responsibility level.* e.g. PAC, STD, etc. For each potential owner a percentage of ownership will have to be assigned (e.g. PAC 25%, STD 30%, etc.).	**Derived from datasets for display**

ANNEX 3 – KEY TABLE COLLECTION

Definition of key table collection

Statistical collection related to a theme or to countries used for key tables publishing (e.g. OECD Key table on Taxation ; OECD Country Statistical Profiles).

Main Metadata exist at 1 level: conceptual.

This object is not cited.

This object is searchable on the iLibrary

DOI suffix= ISSN

Publishing Metadata	Status
ISSN *Each language version of a Key table collection edition has a distinct ISSN*	**Mandatory**
Collection type (code of key table country collection or key table thematic collection) *used to distinguish type of collection for online publishing and citation rules*	**Mandatory**
DOI (digital object identifier of the content) *Each language version of a Key table collection has a distinct DOI*	**Mansatory**
Collection Main Title *(in all available languages)*	**Mandatory**
Collection Subtitle *(in all available languages)*	**Optional**
Update method - Static stand-alone - Static versioned - Dynamic *A key table collection may be published with different update methods.*	**Mandatory**
Language(s) (ISO Codes) of the content: EN and maybe another language - English - and/or French - and/or other language *Relates to the language in which titles, column and row headings, labels and notes are displayed*	**Mandatory**
Short abstracts *(in all available languages)*	**Mandatory**
Long abstracts *(in all available languages)*	**Mandatory**
Keyword(s) *(in all available languages when available)*	**Mandatory**
Theme(s) *(in all available languages when available)*	**Mandatory**
Links	

Has Renditions *Links to the same content in another language,* *Relates a product and its content at rendition level*	**Mandatory**
Is related to *Relates complementary products (e.g. analytical material), sharing the same or similar topics, all of them being recorded into Kappa.* *Relates concepts together and products together.*	**Optional**
Has external links *Relates Kappa objects with external links.*	**Optional**
Supersedes *Indicates that a concept or product has been changed but the content remains the same. Can be used for table versions.*	**Optional**
Is continued by *Indicates that a product has been discontinued or is no longer available, and is replaced by a new one. Links products together.*	**Optional**
Is Edited By *Indicates the legal / organisational body who owns the publication at editorial responsibility level.* **e.g. PAC, STD, etc.** For each potential owner a percentage of ownership will have to be assigned (e.g. PAC 25%, STD 30%, etc.).	**Optional**

ANNEX 4 – KEY TABLES

Definition of key table

Defines table content singled out as very key interest, at content and rendition levels.
This object is not cited. This object is searchable on the iLibrary.

DOI suffix syntax:
<Key table collection ISSN>-table<Key table OrderNumber>
e.g. 16097319-table1

Note that Country Statistical Profile tables do not have a KeyTableOrderNumber, but the ISO3 country acronym is used to identify each table:

DOI suffix syntax:
<Key table collection ISSN>-table-<ISO3>
e.g. 16097319-table-aus
(OECD Country Statistical Profile of Australia)

Publishing Metadata	Status
DOI (digital object identifier of the content) *Each language version of a Key table has a distinct DOI*	**Mandatory**
Main Title *(in all available languages)*	**Mandatory**
Subtitle *(in all available languages)*	**Optional**
Periodicity May be daily weekly, monthly, etc. To be expressed in number of days associated to a label	**Mandatory**
Update method - Static stand-alone - Static versioned - Dynamic *A key table series may be published with different update methods.*	**Mandatory**
Language(s) (ISO Codes) of the content - English - and/or French - and/or other language *Relates to the language in which titles, column and row headings, labels and notes are displayed*	**Mandatory**
Source Note Free area of text when there is no 'source' expressed as "Is Sourced From" metadata (e.g. when the data is calculated by author based on an unpublished database). *Source Note is language qualified (i.e. one note in English and one note in French can be supplied)* *This metadata is mutually exclusive with the metadata link "IsSourceFrom".*	**Optional**
Short abstracts *(in all available languages)*	**Mandatory**
Long abstracts *(in all available languages)*	**Mandatory**

Keyword(s) *(in all available languages when available)*	**Optional**
JEL classification *(in all available languages when available)* *- multiple JEL classifications are possible and they are ordered by relevance*	**Optional**
Variable Index A classification with variable titles in both English and French, although the French titles link to English content.	**Optional**
Related Countries *When the key table is country specific, this relates to the Country ISO Code of the related country*	**Optional**
Theme(s): 1 unique theme *(in all available languages when available)*	**Mandatory**
Links	**Status**
Belongs To (parent-child relationship) Link to the key table collection it belongs to *Relates component concept to parent concept, or component product to parent product (at ISBN and ISSN level)* **Attribute of Link**: Order number of the key table within the parent collection (managed by Editorial) --> used to display the sub-collections the top collection TOC. ***not relevant for country profiles key tables***	**Mandatory**
Is related to *Relates complementary products (e.g. analytical material), sharing the same or similar topics, all of them being recorded into Kappa.* *Relates concepts together and products together.*	**Optional**
Has external links *Relates Kappa objects with external links.*	**Optional**
Has Renditions *Links to the same content in another language,* *Relates a product and its content at rendition level*	**Optional**
Is Sourced From A database (e.g. can be the total database OECD.STAT for the Country Profile) or a publication. If both exist as references, the database is the source while the publication is related to the table via the "Is Part Of" link. *Relates a product to the source(s) of the product* *This metadata link is mutually exclusive with the metadata "SourceNote"*	**Optional**
Supersedes *Indicates that a concept or product has been changed but the content remains the same. Can be used for table versions.*	**Optional**
Is continued by *Indicates that a product has been discontinued or is no longer available, and is replaced by a new one. Links products together.*	**Optional**
Is Imprinted By *Indicates the legal / organisational body who owns the publication at imprint level (e.g. Development Center, IEA)* A logo in different sizes is associated to each imprint. Note: as for book covers different sizes of the same logo may be required for specific channels	**Optional**
Is Copyrighted By *Indicates legal / organisational body(ies) who owns the publication at copyright level.* By default OECD	**Mandatory**
Is Edited By *Indicates the legal / organisational body who owns the publication at editorial responsibility level.*	**Mandatory**

ANNEX 5 – KEY TABLE EDITION

Defines an edited table edition (physical representation of a table as two dimensional object) singled out as very key interest
Main Metadata exist at 3 levels: conceptual, rendition, physical

This object is not searchable on the iLibrary.

Citation rule:
<author physical/institutional> (<year of publication date>), "<Key tableTitle >", <*Key table collection title*>, No. <Key table Order number>.
doi: <doiprefix>/<Key table edition doisuffix>
<doi URL>
(Accessed on <date>)

OECD (2009), "Income tax plus employee social security contributions", *OECD Key Tables on Taxation,* No.1.
doi: 10.1787/16097319-2009-1-table1
http://dx.doi.org/10.1787/16097319-2009-1-table1
(Accessed on 21 January 2009)

Example Country Statistical Profiles citation:
OECD (2008), OECD Country Statistical Profile of Australia.
doi: 10.1787/16097319-2008-1--table-aus
http://dx.doi.org/10.1787/16097319-2008-1-table-aus
(Accessed on 21 December 2008)

DOI suffix syntax:
<Key table collection ISSN>-<EditionYear>-<YearNumber>-table<Key table OrderNumber>
e.g. 16097319-2009-1-table1

Note that Country Statistical Profile tables do not have a KeyTableOrderNumber, but the ISO3 country acronym is used to identify each table:

DOI suffix syntax for Country Statistical Profile:
<Key table collection ISSN>-<EditionYear>-<YearNumber>-table-<ISO3>
e.g. 16097319-2009-1-table-aus

Publishing metadata	Status
DOI (digital object identifier of the content) *Each language version of a Key table edition has a distinct DOI*	**Mandatory**
Main Title *(in all available languages)*	**Mandatory**
Subtitle *(in all available languages)*	**Optional**
Author(s) * Institutional author will be OECD unless external * Physical author - first name - last name - affilitiation - order (when several authors - managed by OECD)	**Optional**

Publication date (quality insurance/validation, overwritten regularly and to be used for email alerts) DD-MM-YYYY	**Mandatory**
Update method - Dynamic - Static The Full-text items of a current *key table edition are continuously updated so "dynamic" and overwritten online until a new key table edition is created. At this time the FTI of the previous edition become "static-one off"*	**optional**
Language(s) (ISO Codes) of the content: EN and maybe another language - English - and/or French - and/or other language *Relates to the language in which titles, column and row headings, labels and notes are displayed*	**Mandatory**
Related Countries *When the key table is country specific, this relates to the Country ISO Code of the related country*	**Optional**
Edition year Each Key Table Edition has a year edition	Mandatory
Year number To distinguish Key table editions released during the same year.	Mandatory
Period covered : Start Year To be used for years structured as follows: yyyy : a single year yyyy-yyyy: a range of years (EVERY year between the two specified values) > it is mutual exclusive with time range	**optional**
Period covered : End Year To be used for years structured as follows: yyyy : a single year yyyy-yyyy: a range of years (EVERY year between the two specified values) > it is mutual exclusive with time range	**optional**
Time range > it is mutual exclusive with period covered start and end year To be used for years structured as follows: […], yyyy: the previous range AND the specified values. Examples: **1980, 1995-2003** means: 1980, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003. **1970, 1980, 1990, 2000-2005** means: 1970, 1980, 1990, 2000, 2001, 2002, 2003, 2004, 2005.	**optional**
Links	**Status**
Is Direct Member Of (parent-child relationship) Link to the key table it belongs to	**Mandatory**
Has Physical Form : Link to content in various output formats/media: (XLS, PDF, OECD.Stat, external link…). Relates a rendition and its physical items *- PDF, XLS, CSV, IVT format relate to a filename.* ***Is not subject to access rights.***	**Mandatory**
Has Renditions Relates a product and its content at rendition level. One rendition corresponds to a specific language version of a content. So this link allows torelate the same content to several language versions. (	**Optional**
Supersedes Indicates that a concept or product has been changed but the content remains the same. Can be used for table versions.	**Optional**

Is Edited By *Indicates the legal / organisational body who owns the publication at editorial responsibility level.* e.g. PAC, STD, etc. For each potential owner a percentage of ownership will have to be assigned (e.g. PAC 25%, STD 30%, etc.).	**Optional**

BIBLIOGRAPHY

ALTMAN, M., KING, G. (2007), "A Proposed Standard for the Scholarly Citation of Quantitative Data", *D-lib Magazine*, March/April 2007, Vol 13 No.3/4

Academia's Buried Treasure, *Information World Review*, October 2008

INGER, S., GARDNER, T. (2008), "How Readers Navigate to Scholarly Content", http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf

NOTES

1. Scholarly Publishing Practice 3, ALPSP, London, 2008 ISBN: 978-0-907341-40-6
2. www.repec.org
3. www.esds.ac.uk/international/introduction.asp
4. http://www.cs.utoronto.ca/~delve/data/datasets.html
5. Inger S., Gardner T, How Readers Navigate to Scholarly Content, 2008 http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf
6. Altman M., King G., A Proposed Standard for the Scholarly Citation of Quantitative Data, D-lib Magazine, March/April 2007, Vol 13 No.3/4
7. Academia's Buried Treasure, Information World Review, October 2008
8. Academia's Buried Treasure, Information World Review, October 2008
9. http://www.tib-hannover.de/en/special-collections/research-data/
10. http://www.tib-hannover.de/en/the-tib/news/news/id/114/ accessed on March 25th 2009
11. Inger S., Gardner T, How Readers Navigate to Scholarly Content, 2008 http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf
12. Examples are IngentaConnect, SwetsWise, EBCSO Host et al
13. Source: www.crossref.org on 25th March 2009
14. http://www.oecd.org/statistics/keytables - accessed on March 25th 2009. On that date the citation information was not appearing in the data objects since it was not planned to be added until a later date in 2009.