

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

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Work Session on Statistical Metadata (METIS)
(Geneva, Switzerland, 10-12 March 2010)

METADATA FOR GOVERNANCE: QUANTIFYING LICENSING AND REDISTRIBUTION ISSUES

Submitted by The Board of Governors of the Federal Reserve System, United States¹

I. INTRODUCTION

1. The importance of data in everyday business life cannot be understated, nor can the amount of information available to users who must choose what is relevant for the important decisions to be made. Hence metadata is of critical value to aid in finding, interpreting, and processing data. As data increase in value, both monetary and strategic, data providers are increasingly asserting rights over their intellectual property. In some cases, this assertion manifests as an enforcement of copyright after a perceived infringement; in others, it involves contracts and licensing, sometimes with fees and monetary costs related to the use of data. For many data providers, selling data is big business, and they are careful to make sure that purchasers follow the appropriate terms and conditions.

2. In many institutions, however, the purchaser or contract reviewer will not be the user of the data. In these instances, it is critical to convey the information about the terms of use to those who will actually be using the data. The need to develop a system to convey this information without having every user review lengthy legal documents was the impetus behind a project being undertaken by data-management staff at the Federal Reserve Board to translate the legal terms and conditions on data use in a manner that the data users would find easy to understand.

3. In 2005, the Division of Research and Statistics at the Board launched a metadata catalogue to capture some basic information about the data that were being purchased from private data providers in support of the Board's research, monetary policy, and regulatory functions. The fields in this catalogue were fairly straightforward: data set name, data provider, purchasing unit, category, key words, and a free-text field to capture a more detailed description of the data product, usually from marketing materials. The metadata types were not very complicated and were initially created as an original specification but then were subsequently adapted to align with the Dublin Core Metadata Initiative.² This catalogue succeeded in meeting its primary

¹ Prepared by San Cannon (sandra.a.cannon@frb.gov). The views expressed are those of the author and do not indicate concurrence by the Board of Governors of the Federal Reserve System.

² The Dublin Core Metadata Initiative is a metadata standard developed in 1995 to "provide simple standards to facilitate the finding, sharing and management of information." See Dublin Core Metadata Initiative, "Mission and Principles," webpage, <http://dublincore.org> (retrieved January 21, 2010).

goal: letting users know what data have already been purchased to prevent duplication of spending. The catalogue also provided information regarding who to contact for questions on the license and storage location.

4. As is quite common with successful applications, users began to demand more from the catalogue than was originally envisioned. In addition, the data managers wanted to add functionality to allow for more types of information to be readily available, especially the contracting information. A redesign was started in 2008 to increase not only the amount of metadata stored for each data set, but also the range of data sets covered to include aggregate government data and data obtained either without charge or without negotiated contracts. Many of the changes were straightforward: adding metadata fields for geographical coverage, dates of availability, and links to other data products, including data collected by the Board. For many of these fields, the database architects could again draw on established metadata standards. Unfortunately, trying to draw from those same metadata standards for the licensing information was less successful.

II. PROBLEM

5. A variety of metadata standards cover metadata of different types and for different purposes.³ They cover concepts important to the arena for which they are intended: data warehousing, document management, survey documentation, statistical time series, and so on. Some even have fields to handle the notion of “rights” or “terms of use” for the content they describe. Unfortunately, most of these fields or attributes contain either lengthy string descriptions of the associated rights or a uniform resource locator, or URL, pointing to a webpage that contains the information. Neither format of information is particularly useful for automated processes or parsimonious storage in a metadata repository. Machine-actionable metadata on rights and usage information are available in various “rights expression languages,” some of which are quite detailed and allow for complicated expressions of digital rights management.⁴ Regrettably, these languages and expression mechanisms do not have other metadata expressions to describe other useful information needed when managing data.

6. Several related concepts are easily conflated when trying to discern how to specify usage terms or rights in many existing metadata specifications. The following is a list of common metadata terms and the questions they really answer:

- **Availability:** Do the data exist? Are they published for use? These concepts really deal with a publication issue rather than a redistribution issue.
- **Access(ibility):** If data are available, how does one get to them? This issue is really a technical one; the Statistical Data and Metadata eXchange (SDMX)⁵ defines access as “[t]he ease and the conditions under which statistical information can be obtained.” The Dublin Core definition deals more with rights pertaining to access (accessRights): “Information about who can access the resource or an indication of its security status.”⁶ But this definition does not explain how the data can be used once access is obtained.

³ For a brief list of such standards, see Metadata Advisory Group, MIT Libraries, “Selected Metadata Standards,” *Metadata Reference Guide*, <http://libraries.mit.edu/guides/subjects/metadata/standards.html> (referenced December 20, 2009). A more detailed treatment, albeit in a different context, can be found in Jesse M. Blum and Kenneth J. Turner (2008), “The DAMES Metadata Approach,” DAMES Node, Technical Paper 2008-2 (Stirling, Scotland: Universities of Stirling and Glasgow, December), www.dames.org.uk/docs/tech_papers/DAMES_tp2008-2.pdf (referenced December 20, 2009).

⁴ For an overview, see Karen Coyle (2004), *Rights Expression Languages: A Report for the Library of Congress* (February), www.loc.gov/standards/relreport.pdf (referenced December 20, 2009).

⁵ The SDMX standard is an ISO technical specification (ISO/TS 17369) developed “to foster standards for the exchange of statistical information.” See the SDMX definition of “Accessibility” in *Cross-Domain Concepts* (2009) http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf (retrieved January 15, 2010).

⁶ See the Dublin Core definition of “accessRights” at Dublin Core Metadata Initiative, *DCMI Metadata Terms*, “Section 2: Properties in the /terms/ namespace,” <http://dublincore.org/documents/dcmi-terms/#H2> (retrieved December 20, 2009).

- **Confidentiality:** If the data are available and I can get them, what are the security considerations for their use? Confidentiality is really a privacy issue. Again, the SDMX definition: “A property of data indicating the extent to which their unauthorised disclosure could be prejudicial or harmful to the interest of the source or other relevant parties.”⁷
- **Rights:** Do I have permission to do anything with these data? If so, what? This issue is really the heart of the licensing and redistribution conundrum; it is a permission issue. The Dublin Core definition: “Information about rights held in and over the resource. Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.”⁸

7. Of the metadata specifications reviewed for this work, that of Dublin Core defines the most attributes for various aspects of rights (e.g., rights, rightsHolder, license), but none of them are in a format that would allow an automated response to the question “What can I do with these data?” In fact, most could not allow an automated process to formulate a response to a yes or no question such as “Can I draw a chart of these data?” These types of questions are the kinds for which data users need ready access to answers.

III. SOLUTION

8. Given the lack of a standard solution, data managers at the Board set about to define the questions to which their users needed answers and to see how those answers could be best fit into a metadata specification. The current prototype is very specific to the data-management needs currently faced by the Board, but work is continuing to try to broaden the scope to be more widely applicable.

9. One major concern with data redistribution for Board staff is the presentation of materials on the public website. Many data contracts, licenses, or terms of use restrict the user’s right to “redistribute” the data or “publish derivative works,” which in many cases translates into tabling or charting the data. Data managers wanted to be able to clearly communicate when these activities were allowed. Some data providers also distinguish between print and electronic media, as technology allows for the ready harvesting of data from webpages; print is not immune to data piracy, but the costs are much higher. Additional clarifications were also needed for outright data sharing; in many instances, Board staff members wish to share data with colleagues at Federal Reserve Banks, coauthors at academic institutions, and even the public.

10. To that end, data managers outlined the binary fields necessary for a simple graphical presentation of what the permitted uses are:

- Chart the data for a printed publication?
- Table the data for a printed publication?
- Chart the data for online publication?
- Table the data for online publication?
- Share the data with Federal Reserve Banks?
- Share the data with others?

11. One additional twist to this simple delineation is that microlevel data often have different permissions for the raw data than for aggregates derived from the raw data. For providers of microlevel data, this list is then replicated and answered for both the raw data and aggregated data derived from it. For data that are published at an aggregated level, there is no microlevel counterpart.

⁷ See the SDMX definition of “Confidentiality” in *Cross-Domain Concepts* (2009) http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf (retrieved January 15, 2010).

⁸ See the Dublin Core definition of “rights” at Dublin Core Metadata Initiative, *DCMI Metadata Terms*, “Section 2: Properties in the /terms/ namespace,” <http://dublincore.org/documents/dcmi-terms/#H2> (retrieved December 20, 2009).

12. For the actual metadata fields, we have chosen to use a few categories with multiple values:
- Chart: paper, web, both, none
 - Table: paper, web, both, none
 - Share: all, banks, none
13. These tags are stored in the metadata warehouse and rendered as yes/no representations that work well in a graphical display for the users:

Can data be published?		
	Raw data	
	Chart	Table
Publication	No	No
Website	No	No

Can data be shared with?
Reserve Banks: Yes
Public: No

Can data be published?		
	Aggregate data	
	Chart	Table
Publication	Yes	Yes
Website	Yes	Yes

Can data be shared with?
Reserve Banks: Yes
Public: Yes

14. By doing so, an author writing a paper using a particular data source can easily see whether there are any restrictions on charts or tables appearing in that paper once it has been published, regardless of the media. Should colleagues or journal referees want a copy of the data set for that paper, the guidelines are also clearly

IV. SOLUTION

15. The metadata fields themselves are not particularly complicated; the challenge has been to translate licensing agreements and terms-of-use statements from typical legal jargon into the simple binary statements displayed in the graphic. Legal counsel at the Board has worked closely with data managers and library staff to clarify and delineate terms for existing agreements. The Research and Statistics Division has also revamped its data acquisition procedure to allow an electronic resources librarian to shepherd negotiations through the process and get these questions answered before contracts are signed.

16. For data that we receive either without payment or without a signed contract, more work needs to be done to ensure that Board data users know the terms with which they need to comply. The ease of downloading data from a website makes it critical that there be a clear understanding of how rights and restrictions are communicated. While U.S. courts have not ruled definitively on the validity of “browse-wrap” agreements,⁹ it is not in the Board’s best interest to wait for an ex post judgment; data managers prefer either to be proactive and work with research staff members to ensure understanding and compliance with stated terms or to work with legal staff to negotiate new terms. In several cases, data managers or other research staff at the Board have requested the specific permissions listed earlier from data providers for data that are generally considered to be “freely available” but that may have restrictions in the terms of use that prohibit the type of use that would be desired by Board staff.

17. For example, researchers at the Board were interested in using the S&P/Case-Shiller Home Price Indices data published by Standard & Poor’s. No cost was associated with downloading these data, but the website had fairly restrictive terms of use:

⁹ “Browse-wrap” or “browse through” agreements are agreements in which the terms-of-use pages are surreptitiously linked to a webpage and state that usage of the site constitutes agreement with said terms. See Christina L. Kunz et al. (2003), “Browse-Wrap Agreements: Validity of Implied Assent in Electronic Form Agreements,” *Business Lawyer*, vol. 59 (November), pp. 279-312.

The contents of the Web Site made accessible by Standard & Poor's on the Web Site including, but not limited to, the Standard & Poor's ratings and other opinions, text, data, reports, images, photos, graphics, graphs, charts, animations and video (the "Content"), may be used only for your personal individual use. Except as expressly permitted under these Terms of Use, you agree not to copy, reproduce, modify, create derivative works from, or store any Content, in whole or in part, from the Web Site or to display, perform, publish, distribute, transmit, broadcast or circulate any Content to anyone, or for any commercial purpose, without the express prior written consent of Standard & Poor's.¹⁰

18. It would be very limiting if data managers could not store the data locally and if researchers could not "create derivative works" such as research papers, policy documents, or charts for presentations. An e-mail to Standard & Poor's requesting permission to store the data, chart and table the data, and present such derivative works on the public website was promptly answered in the affirmative and the appropriate metadata recorded.¹¹

19. Therefore, the existence of a contract or other signed document is not a prerequisite for the recording of licensing information in the metadata warehouse. For official statistics or other aggregate time-series data that are retrieved from the websites of U.S. government agencies, few restrictions apply, as U.S. government agencies cannot assert copyright for the products of their employees. Many data series are provided to the Board by statistical agencies through other channels, however, and it is important to make appropriate notations about the restrictions on use for such "suppressed" or "unpublished" data.

V. CONCLUSION

20. Work is in process to streamline the collection of metadata for governance even further: A new data contract addendum, developed by Board legal staff for use in the data acquisition process, makes plain what the restrictions on usage are for contracted data, and further research is under way to try to map more easily between common copyright and contract terminology and the licensing metadata stored in the warehouse. Data managers and electronic resource librarians are working to fill all the licensing metadata fields for existing data contracts as well as including those data sources for which there are no signed agreements. As data types and usage expand, it is conceivable that more metadata fields on permitted usage will be necessary to cover procedures not yet envisioned. As the appetite for data continues to grow, the importance of metadata, including metadata for governance, will continue to expand.

¹⁰ See the Internet Archive, Standard & Poor's website from June 29, 2008, with terms of use dated from November 16, 2007, <http://web.archive.org/web/20080629231315/http://www2.standardandpoors.com> (accessed February 12, 2010).

¹¹ Terms of service for Standard & Poor's website are available from www.standardandpoors.com/terms-of-use/en/us/; the current version was posted October 1, 2009 and permits the usage for which we previously had to request permission.