IJPIS

METADATA FOR GOVERNANCE:
QUANTIFYING LICENSING AND REDISTRIBUTION ISSUES

SAN CANNON
Division of Research and Statistics
Federal Reserve Board
Washington D.C. 20551
sandra.a.cannon@frb.gov

Abstract

Like all metadata, the information about data governance, specifically here the rules and rights for usage, is important for users to have readily available in an understandable format. This paper will discuss some issues involved with licensing and redistribution of data, summarize how those issues are handled by some widely-used metadata standards, and outline how data managers at the Federal Reserve Board are building a metadata structure to capture the redistribution information they need to properly support the work done by economists at the US central bank.

Keywords: Data, metadata, standards, Dublin Core, rights, licensing

1. Introduction

The importance of data and news available to users in everyday business life cannot be understated. From the vast array of information they face daily, they must choose what is relevant for making decisions. Hence the critical value of metadata to aid in finding, interpreting, and processing data. Metadata are crucial to making valuable data assets more visible and easier to use, thereby maximizing efficient use of the data. As data increase in value, both monetary and strategic, data providers are increasingly asserting rights over their intellectual property. In some cases this manifests as enforcement of copyright after a perceived infringement; in others, it involves contracts and licensing, sometimes with fees and other 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 of contracts.

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 that will actually be doing the using. Developing a system to convey this information without having every user review lengthy legal documents was the impetus behind a project undertaken by data management staff at the U.S. Federal Reserve Board (hereinafter, the Board) to quantify the legal terms and conditions on data use in a manner that the data users would find easy to understand.

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: dataset 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 fields weren't very complicated and were drawn mainly from the Dublin Core Metadata Initiative. [DCMI, 2008] This catalogue succeeded in meeting its primary goal: let users know what data

have already been purchased to prevent duplication of spending. The catalogue also provided information on who to contact for questions on the license agreement and storage location. 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 dataset but also the range of datasets covered. This included aggregate government data and data obtained without charge or 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.

2. Problem to be Solved

There are a variety of metadata standards to cover metadata of different types and for different purposes. [MIT Libraries 2009; Blum and Turner 2008]. They cover concepts important to the arena for which they are intended: data warehousing, document management, survey documentation, statistical time series, etc. 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 URL pointing to a web page that contains the information. Neither format of information is particularly useful for automated processes or parsimonious storage in a metadata repository. In addition, there are several related concepts that 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 is really a technical issue. For instance, the Statistical Data and Metadata eXchange (SDMX)[1] definition of access is "The ease and the conditions under which statistical information can be obtained." On the other hand, the Dublin Core definition deals more with rights pertaining to access (accessRights). Its definition is "Information about who can access the resource or an indication of its security status." But neither explains how the data can be used once access is obtained.
- **Confidentiality**: If the data are available and one can get them, what are the security and privacy considerations for their use? Again, the SDMX definition is "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." This is a privacy or disclosure concern separate from intellectual property rights. Confidentiality may be a reason for strict terms of use but it is not a necessary condition for all cases.
- **Rights:** Is it permissible to do anything with these data? If so, what? This is really the heart of the licensing and redistribution conundrum: a permission issue. The Dublin Core definition is "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."

[1] The Statistical Data and Metadata Exchange (SDMX) standard is an ISO Technical specification (ISO/TS 17369) developed "to foster standards for the exchange of statistical information. (Retrieved on January 15, 2009 from http://sdmx.org/)

Of the metadata specifications reviewed for this work, the Dublin Core defines the most attributes for various aspects of rights (e.g. rights, rightsHolder, license) but they are designed for human readability, not automated processing. An automated response to questions, such as ”What can I do with these data?” or ”Can I draw a chart of these data?”, is not easily attained. Machine-actionable expressions of rights and usage are available, however, in various “Rights Expression Languages” (RELs). Some of these specifications are quite detailed and allow for complicated expressions of digital rights management (Coyle [2004]). Several expression mechanisms reviewed for this work, which are not actually metadata standards, could possibly be used to capture some of the basic concepts. Full adoption of these underlying rights models, however, would not work so well. The level of abstraction for these languages makes implementation too complicated [ODRL 2009, XrML 2002]. Some rights expression languages assume a default class of “all users” which is too broad for our purposes and made them unsuitable for further investigation [ccRel 2008].

3. Solution: Take One

Given the lack of an obvious standard solution, data managers at the Board set about to define the questions to which their users needed answers and see how to best fit those answers into some kind of metadata specification. This prototype was very specific to the data management needs faced by the Board and while this initial approach was not successful, it was a good preliminary foundation on which subsequent work is based.

One major concern with redistribution is presentation of materials on a public-facing 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 to creating tables or charts of the data. Data managers should be able to clearly communicate when these activities were allowed. It initially appeared that some data providers distinguish between print and electronic media, also. Technology readily allows for harvesting of data from web pages, but print is not immune to data piracy problems, albeit with higher costs. Additional clarifications were also needed for outright data sharing; in many instances researchers may wish to share data with colleagues at other agencies, co-authors at academic institutions, and even with the public.

To that end, data managers outlined the binary (yes/no) fields necessary for a simple graphical presentation of what the permitted uses are:

- Create charts for the data for a printed publication?
- Create tables for the data for a printed publication?
- Create charts for the data for online publication?
- Create tables for the data for online publication?
- Share the data with others?

One additional twist to this simple delineation is that micro-level data often have different permissions for the raw data than for aggregates derived from the raw data. For those providers, this list was then replicated and answered for both the raw data and the aggregate data.

For the actual metadata fields in the implementation, we chose to simplify the system by using a few categories with multiple values:

- Chart: paper, web, both, none
- Table: paper, web, both, none
- Share: all, banks, none

These elements were stored in the metadata warehouse and rendered as a yes/no representation that works well for a graphical display of permitted uses for the users:

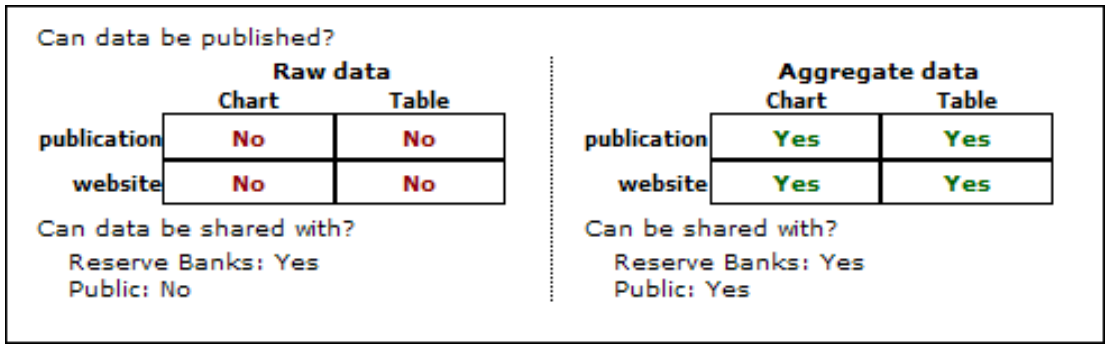

Figure 1. Permitted use matrix, version one.

So an author writing a research paper using a particular data source could 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 dataset for that paper, the guidelines were also clearly outlined.[2]

4. Legality Meets Technology

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

For data that the Board receives either without payment or without a signed contract, more work needs to be done to ensure that data users know the terms with which they need to comply. The ease of downloading data from a website makes it critical for there to be a clear understanding of how rights and restrictions are communicated. Many data users are not aware that nearly all websites dictate some terms for the use of their websites. Terms-of-use pages are linked, sometimes in an inconspicuous manner, to a web page and may state that the mere usage of the site constitutes agreement with said terms. While US courts have not ruled definitively on the validity of such agreements [Kunz et al. 2003], it is often not in an institution's best interest to wait for an ex post legal judgment; data managers should be proactive and work with data users to ensure understanding and compliance with stated terms, or work with legal staff to negotiate new terms. In some cases, data managers may have to request specific permissions from data providers for data that are generally considered to be "freely available" but may have restrictions in the terms of use that prohibit the type of use that would be desired.

[2] The Federal Reserve Board has an additional challenge in that from a legal organization perspective, the 12 Reserve Banks in the Federal Reserve System are separate and private legal entities. The Board of Governors is also part of the system but it is considered an independent government agency. It is therefore quite common for data contracts to be written allowing access to the Board but not the Reserve Banks and vice versa. For more information, see http://www.federalreserve.gov/pubs/frseries/frseri.htm (Accessed May 10, 2010)

For example, in 2006 researchers at the Board became interested in making use of the Case/Sheller House Price Index data published by Standard & Poor's. There was no cost associated with downloading these data but at that time the website had fairly restrictive terms of use:

> "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." [Standard & Poor's 2007]

It would be very limiting if data managers couldn't store the data locally and if researchers couldn't "create derivative works" such as research papers, policy documents, or charts for presentations. An email to Standard & Poor's requesting permission to store the data, create charts and tables from the data, and present such derivative works on the public website was promptly answered in the affirmative and the appropriate metadata recorded[3].

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 US government agencies, there are few restrictions as US government agencies cannot assert copyright for the products of their employees. This restriction is not universally true, however. Agencies of other governments can, and often do, assert copyright or dictate licensing terms [Cannon and Rodriguez 2010].

5. Solution: Take Two

Unfortunately, the straightforward representation originally designed did not fit well with how contracts are actually written. Upon further investigation, it was discovered that most data contracts do not distinguish between print and on-line publication and those that do are a small enough subset to be dealt with as a special case instead of creating a major distinction to be applied to all contract information. The bigger question which was not adequately addressed originally involved the disposition of data as a standalone entity versus data as incorporated in a "derivate work" or "work product." Here, "Work Product" is defined as output that incorporates portions of the information supplied by a data provider into other output such as research papers, speeches, testimony, reports, and official publication and policy documents. These are the common types of output created using data that most people would be concerned with disseminating. Therefore, specific treatment of the rules involving the inclusion of data within a work product in a way that easily maps from what the contract says to what users want to do is necessary. The work on this mapping did not address the other contractual issues (such as storage and retention, for example) as they are outside the scope of what the matrix is trying to do. There are several major changes to the permitted use matrix described earlier:

1. Addition of "Other agencies" as potential counterparties distinct from the public - A recent working paper highlighted the importance of a combined data set using data from two different government agencies [Eichner et al. 2010]. Such data sharing could potentially improve national statistics and related policy work and would be necessary, at least between financial regulators, in any centralized systemic risk monitoring system.

[3] The current terms of use posted on the Standard & Poor's site, dated October 1, 2009, (www.standardandpoors.com/terms-of-use/en/us/, accessed April 25, 2010) do not present such strict usage rules and likely would not have required the specific permission that was requested.

2. Separation of disposition of data from work product - Again, given the greater potential for data sharing discussions, it was necessary to make clear where the restrictions on the disposition of the work product as a separate entity from the data are.
3. Change in the wording for "aggregate" data - Much discussion revealed that there were two different definitions of the word "aggregate" being employed and the subtle difference was causing problems when trying to be consistent across the different usages. The real distinction from a contractual perspective is not between "raw" and "aggregate", which mean different things in different contexts, but between "unmodified" and "derived." "Summary" would also be applicable to the latter category as well so we could say "Derived or summary." If we are providing "unmodified data" or "identifiable raw data" either as a standalone entity or incorporated in a Work Product, it is likely that we'll need to follow some guidelines to ensure compliance with any legal language preventing the Board from publishing information that may constitute a substitute for the purchase of the data from the original provider. For "derived or summary data," there really should be no notion of substitution.
4. Clarification of "derived" or "aggregated" data as a work product - A data series created from raw source data in a manner that incorporates our intellectual property and from which the source data cannot be reverse engineered constitutes a work product itself.

It is important to note that data tables are often presented as an alternative to graphical depictions to fulfill the Board's accessibility requirements under Section 508 of the Rehabilitation Act of 1973, as amended [29 U.S.C. §794d]. Sharing data in this context is considered to be part of the work product which the Board defines as a creative work that involves the application of intellectual effort. Work products can include tables, charts, spreadsheets, and databases so long as they meet the test of being a "creative work" and adhere to any additional permitted use guidelines in the contract. There are quantity limits as well so authors need to be mindful that they cannot incorporate information in a quantity or manner that would be viewed as a substitute for the data purchased from the vendor.

Incorporating the results of the discussions described above, we developed the permitted use matrix depicted in Figure 2.

Table 1. Data from this resource may be shared in the following formats with the audiences indicated:

		Audience			
		Internal (Board Only)	System (Reserve Banks)	Government Agencies	Public
Format	Work Product containing ***no identifiable raw*** or source data	Y/N/Q	Y/N/Q	Y/N/Q	Y/N/Q
	Work Product containing a ***limited quantity of identifiable raw*** or source data	Y/N/Q	Y/N/Q	Y/N/Q	Y/N/Q
	Identifiable raw or source data not part of a Work Product	Y/N/Q	Y/N/Q	Y/N/Q	Y/N/Q

Figure 2: New Permitted Use Matrix

The allowed values in each cell have the following meanings:

- **Y –** Unqualified yes. No restrictions on usage
- **Q –** Qualified yes. There are conditions on coverage, quantity or representation. These will be spelled out in a text box. Some instances, such as restriction on the number of concurrent users, may be common enough to be represented in a standard, possibly machine-actionable format.
- **N –** Unqualified no. No data (even a single observation) are to be included in a work product.

As used in the row headings in the matrix, "**Identifiable raw**" describes data as we receive it from the provider or source including modifications that allow reverse engineering to retrieve source data.

Finally, the 4 following categories define the columns of the matrix:

- **Internal:** For use by Board staff only.
- **System:** Accessible to Board and System staff.
- **Agencies:** Other regulators or statistical agencies, particularly of concern for policy and regulatory work.
- **Public:** General access including co-authors outside the System and other agencies.

The information will be lifted directly from the contract language by either data managers or librarians and the user interface will display the relevant information, including the text qualifiers[4]. The stored representation will be more complicated than that for the first solution because of the additional audiences (other regulatory or statistical agencies) as well changing the choice from binary to multiple options. Practically, the audience combinations break down to the following:

- Board only
- Reserve Banks only
- Board and Reserve Banks
- Board and Agencies
- Board, Reserve Banks, and Agencies
- Public

Public is an encompassing state. If the license allows for public access, then the Board, the Reserve Banks, and other agencies can be included as part of the public. In addition, it is unlikely that the Reserve Banks would enter a contract that allowed access to other Agencies while excluding the Board, so that combination can be eliminated. Each of the three categories (work product with no data, work product with data, data by itself) would then need to be combined with the possible rights holders to effectively store the information needed to depict the licensing for the users.

Where possible, we plan to use terminology consistent with the concepts in many rights expression languages even if applying the information model for that language is problematic. These languages generally make clear distinctions between "Modify" (encompassed in the top row of the matrix) and "Excerpt" (describing the bottom row). The concept of 'work product with data' (the center row) may be especially tricky to model in these languages as it is a combination of the two. Therefore, further

[4] For example, if there was a restriction of the number of observations that could be incorporated into a work product, that cell would contain the identifier "Q" and a text note below would outline that particular restriction.

research is underway to formally model these expressions and relationships employing standard rights expression terms.

6. Conclusion

The representation of the rights for data users to easily consult was originally the primary goal of this work. Further investigation into rights management expression has shown the need to separate the abstract notion of the right, the database representation of the right, and the display of the rights for user consultation. While the original goal has been met, further work is ongoing to develop a more robust representation that could simplify the workflow and allow for more automated processing of information either on data entry or when disseminating the information. Work being done around attaching Creative Commons licenses to previously copyrighted government data may help to offer insight into how data producers think about providing licensing information to a broader audience even in the absence of a contract.

Operationally, there is still work to be done to streamline the collection of metadata for governance at the Federal Reserve Board even further. 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 kinds of usage expand, it is conceivable that more metadata fields on permitted usage will be necessary to cover procedures not yet envisioned. For example, a governance concern that will need to be addressed in the future deals with the "mosaic effect"[5]: data that are just pieces of a puzzle alone do not have deidentification issues that can arise when those pieces are combined with other pieces and a different picture emerges. It is possible that combining different bank-level data sets, for example, would allow a researcher to obtain a different picture than using either data set individually, which is desirable. It may also be possible that the joined data set provides confidential information about, or identification of, one of the institutions in the data set. Just the possibility of such a discovery may violate terms of a contract or agreement even if the unearthing was inadvertent and the exposed information is never used. These issues need to be dealt with in a systematic way. As the appetite for data continues to grow, the importance of metadata, including metadata for governance, will grow as well.

Acknowledgement

The views expressed are those of the author and do not indicate concurrence by the Board of Governors of the Federal Reserve System or its staff. I am grateful to two anonymous referees and to Alistair Hamilton for his extensive comments and suggestions. I am indebted to Linda Powell, Chris Black, and Joanne Kee for the collaborative work done in this area but all errors or omissions are my own.

References

Abelson, H., Adida, B., Linksvayer, M., and Yergler, N. 2008 *ccREL: The Creative Commons Rights Expression Language*.

Blum, J. and Turner, K. 2008. *The DAMES Metadata Approach. Technical Report CSM-177,* Department of Computing Science and Mathematics, University of Stirling.

Boettcher, A. 2007. Data and Knowledge Management at the Federal Reserve Board. *IASSIST Quarterly 31(2),* pp. 12-19.

Cannon, S., Rodriguez, M. 2010 Survey of On-line Data Dissemination Practices for Government and International Statistics. *The OECD Statistics Newsletter.* Issue 49, July.Coyle, K. 2004. Rights Expression Languages: A Report for the Library of Congress. Library of Congress.

[5] For a discussion of the Mosaic Effect in the context of personal identification see Vijayan, Jaikumar. Sidebar: The Mosaic Effect. *Computerworld.* March 15, 2004. (Accessed at http://www.computerworld.com/s/article/91109/Sidebar_The_Mosaic_Effect on April 24, 2010)

Dublin Core Metadata Initiative 2008. DCMI Metadata Terms. Retrieved January 10, 2010 from http://dublincore.org/documents/dcmi-terms.

Eichner, M. J., Kohn, D. L., & Palumbo, M. G. 2010. Financial Statistics for the United States and the Crisis: What Did They Get Right, What Did They Miss and How Should They Change? *Financial and Economics Discussion Series,* 2010-20, Board of Governors of the Federal Reserve Board, Washington D.C.

Kunz, C., Ottaviani, J., Ziff, E., Moringiello, J., 2003. Browse-Wrap Agreements: Validity of Implied Assent in Electronic Form Agreements. 59 *BUS. LAW*, pp. 279-331.

MIT Libraries. 2009. Selected Metadata Standards. Retrieved April 19, 2010 from http://libraries.mit.edu/guides/subjects/metadata/standards.html

Open Digital Rights Language (ODRL V2.0). 2009. Retrieved August 30, 2010 from http://odrl.net/2.0/DS-ODRL-Model-20090306.html

Standard & Poor's. 2007. Terms of Use. Retrieved February 12, 2010 from http://web.archive.org referencing Standard & Poor's site from June 29, 2008 with terms of use dated November 16, 2007.

XrML 2.0 Technical Overview, 2002. Retrieved August 30, 2010 from http://www.xrml.org/reference.asp