Building a Strong Foundation with Dublin Core

Linda Powell
Board of Governors of the
Federal Reserve System,
U.S.A.
Linda.Powell@frb.gov

Abstract

When the Board of Governors of the Federal Reserve System needed to create a new metadata system, staff decided to start with a time-tested standard, Dublin Core. The U.S. central bank consumes a variety of metadata, including metadata that define collections of data and metadata that describe variable-level data. This paper follows the process used to create collection-level metadata and variable-level metadata, then to retrofit existing metadata to make them all usable by economists, financial analysts, and information technology professionals.

Keywords: metadata; financial data; banking.

1. Introduction

The Board of Governors of the Federal Reserve System (the Board) collects, gathers, and purchases data from a variety of sources. Advances in technology and industry standards over the past 10 years have enabled firms, including the Board, to improve data repositories. For a data repository to be complete, it needs good metadata, which are generally defined as "data about data," or the electronic data dictionary for the data of interest. Although there are many types of metadata, for the purposes of this paper, we will focus on collection-level metadata (sometimes called meta-metadata), which describe groups of data or data sets, and variable-level metadata, which define each item (or element) in the data set. For example, the financial data of companies can be broken into collection-level metadata and variable-level metadata. The collection- level data could include (1) a description indicating that the data represent individual companies' financial statements, (2) how the data are compiled, and (3) the date range available. The variable-level metadata would include information about the data within the financial statements such as (1) a description for each variable (for example, total assets, total liabilities, or capital ratio), (2) the data type (for example, monetary, percent, text), and (3) formulas used to calculate the numbers.

2. The Business Problem

Over time, as technology has grown, the availability of data from the private sector has expanded and the financial markets have increased in complexity, resulting in changes to the types of data the Board uses. Historically, when the Board purchased data, they were used for specific research projects rather than market or policy analysis. Today, significant volumes of financial market data are purchased. Most of these data sets are time series cross-sectional data, meaning that they cover data for a large number of institutions over many periods. The periods can range from daily (for example, the value of derivative contracts) to quarterly (for example, the financial statements of all U.S. banks) or periodic observations. In 2006, the change in business practices at the Board became apparent, and the need for better information about purchased data was addressed by creating a new collection-level metadata repository.

As of 2009, the Board has three main metadata repositories for firm-level data: one for collection-level data called the Data and News CatalogueE (DANCE), one for data collected by the Federal Reserve System called the MicroData Reference Manual (MDRM), and one for data purchased from vendors (Vendor Metadata). Each of these repositories was developed to meet a

https://doi.org/10.23106/dcmi.952109594

specific need. At the time the repositories were developed, the metadata requirements were based on a subset of users' needs and requirements.

Over the past several years, these repositories have been evaluated to determine if they meet the needs of the user community and if and how they can be used collectively. In summary, we found that the repositories should not be combined but should be revamped to provide better information and enable working collaboratively. To revise these repositories, we looked to the body of international standards for metadata and chose Dublin Core as the primary standard for revamping our metadata repositories.

3. Existing Metadata Repositories for Non–Time Series Data

Metadata have been used at the Board since before the World Wide Web came into existence. The metadata for data collected by the Federal Reserve System have been maintained and well organized for more than 30 years in the MDRM. The MDRM provides both collection- level and variable-level metadata. The presentation has improved and the scope of explanatory data in the MDRM has increased over the years. Although the MDRM is an established and useful tool, it could benefit from adherence to industry standards and consistency with other metadata tools.

The DANCE repository was first developed in 2004 to collect and disseminate basic information about data purchased by the Board. The original purpose was to facilitate finding various types of data. The need for additional information became apparent as soon as the application reached production. In addition to the need for more information about the data sets, it was discovered that users of the data sets needed a centralized location to record information about their experiences using the data. The implementation of an international standard would provide a strong foundation for determining the scope of the expanded variables. To meet the need for experiential information, it was determined that a wiki page for each data collection would provide users with the opportunity to share information.

The Vendor Metadata repository development began in 2007 to facilitate the programmatic creation of internal SAS data sets from purchased financial data. It was quickly expanded to document the data structure and variables for use by end users. In addition to variable-level metadata, this repository contains system or technical metadata detailing file layouts and formats. Because this repository was originally designed to store system-level metadata, using an industry standard to expand on the metadata provided guidance on the scope of the variables and consistency among the repositories.

4. Review of Metadata Standards

Although the DANCE application met the requirements of its initial purpose, as soon as the application hit production, it became obvious that this application could be expanded, and additional demands surfaced. Once the decision was made to expand the application, a review of metadata standards was performed. While examining the various industry standards, it was determined that all three of the Board's metadata repositories could benefit from review and the application of industry standards. In addition, there was an effort to achieve consistency among the repositories.

Although additional standards were reviewed, five candidates for the Board's data warranted in-depth evaluation: SDMX (Statistical Data and Metadata Exchange), XBRL (Extensible Business Reporting Language), MARC, DDI (Data Documentation Initiative), and Dublin Core. The review began with the standards that Board staff were familiar with: SDMX, XBRL, and MARC. SDMX appeared to be geared toward aggregate time series data rather than cross-sectional data (SDMX Standards). XBRL provided guidance for the variable-level metadata but did not provide the scope of collection-level metadata (which was the primary focus of the review) that other standards provided (XBRL Specifications). And MARC was familiar to the Board's library staff but was more bibliographic than necessary for this purpose (MARC 21 Introduction).

https://doi.org/10.23106/dcmi.952109594

DDI provided good collection-level metadata as well as extensive variable-level metadata. However, at the time DDI was reviewed, it did not allow for data collections that persisted over time. In addition, the variable-level metadata were geared toward social science survey data, whereas the majority of data stored in these repositories are financial in nature. (DDI 2.1 Specifications.)

Although none of the standards were a perfect fit for the Board's needs, we found that Dublin Core had an excellent breadth of metadata elements without being overwhelming. Dublin Core was a particularly good match for the collection-level metadata. Dublin Core also had the advantage of being publicly well documented and relatively uncomplicated, thus enabling immediate application of the standard. (DCMI Element Set.)

Although Dublin Core was found to be a well-fitting standard for collection-level metadata, we found that we needed additional variable-level metadata; as such, we turned to XBRL. XBRL is a good fit for the variable-level data because it focuses on financial statement data, which comprise the majority of data that are stored in the metadata repository and the MDRM.

5. APPLICATION PROFILE

The DANCE application profile consists of the variables listed in Appendix 1. Because some variables can change over time (such as update accrual method) and other variables (such as relation) can have multiple entries, the data were divided into separate tables in the database. The application profile was designed to support the discovery of metadata and provide information to end users. Most of the variables in the application profile are from Dublin Core; however, some variables were added to supplement the Dublin Core elements and provide agency-specific information such as which department within the organization purchased the data. The metadata variables were chosen by a group of end users consisting of data managers, programmers, librarians, and financial analysts. The user group reviewed each of the Dublin Core elements and modeled example data sets to determine the applicability of each element. The Vendor Metadata and MDRM application profiles expand on the DANCE profile and incorporate XBRL elements specific to financial data.

6. Open Source Software

Once the decision to use the Dublin Core standard for collection-level metadata was made, we explored the possibility of capitalizing on one of the open source repositories. We preferred to use a PostgreSQL relational database but were open to MySQL as a backend storage facility. The systems chosen for evaluation were EPrints, FEDORA, and DSpace. After our initial review, DSpace appeared to be the best candidate for the new DANCE. However, in-depth review and initial testing showed that to meet the requirements for the new DANCE application, the open source system would need significant customization. In general, we wanted to be able to integrate our data security information and Vendor Metadata repository with DANCE to allow end users to drill down through the application for related information. In the end, we decided to build our own collection-level and variable-level repositories in-house using a PostgreSQL server on Linux.

7. Conclusions and Future Work

The greatest advantage of using the Dublin Core standard was that the collective experience and knowledge of hundreds of professionals went into creating the standard. Some administrative data that were specific to the Board's needs were added to enhance the standard's variables. A quick review of the original variables versus the complete list of variables for the DANCE repository in Appendix 1 demonstrates the greater breadth of information suggested by the Dublin Core standard than was originally identified. The Vendor Metadata repository, Appendix 2, and the variable-level metadata repository in the MDRM, Appendix 3, both benefited from the inclusion of Dublin Core and XBRL variable-level metadata.

https://doi.org/10.23106/dcmi.952109594

A second advantage of using the Dublin Core standard is the interoperability among systems that Dublin Core facilitates (that is, using a common standard for all of the Board's metadata repositories will assist in automating data retrieval from all three repositories). These repositories were designed for use within the Board, but the Federal Reserve System has 12 regional banks that have expressed interest in the data stored in the repositories. Using an industry standard should facilitate easier interoperability among the various locations. Finally, the Board currently publishes aggregate economic macrodata using SDMX. Because SDMX supports Dublin Core as a metadata format, there may be opportunities to capitalize on the interoperability of SDMX and Dublin Core when bridging firm-level microdata to aggregate-level macrodata.

Acknowledgments

The author wishes to thank her colleagues Wei-na Chow and Andrew Boettcher for their research contributions to this paper. The views expressed in this paper are those of the author and do not necessarily represent those of the Board of Governors of the Federal Reserve System.

References

CPIT. (2007). Technical Evaluation of Selected Open Source Repository Solutions, Version 1.3. Retrieved February 27, 2007 from https://www.eduforge.org/docman/view.php/131/1062/Repository%20Evaluation%20Document.pdf.

Data Documentation Initiative. (2006-2009). Retrieved August 3, 2009 from http://www.ddialliance.org.

DCMI. (2006-2009). DCMI Element Set. Retrieved February 27, 2007 from http://www.dublincore.org/index.shtml.

DCMI. (2006-2009). DCMI Metadata Terms. Retrieved February 27, 2007 from http://www.dublincore.org/index.shtml.

DCMI. (2006-2009). User guide. Retrieved February 27, 2007 from http://www.dublincore.org/index.shtml.

INNODATA ISOGEN. Content Metadata Standards: Libraries, Publishers, and More. White paper. Retrieved October 30, 2008 from http://www.innodata-isogen.com/knowledge_center/white_papers/content_metadata_standards_wp.

MARC. (April 2008). MARC 21 Introduction. Retrieved January 4, 2009 from http://www.loc.gov/marc/bibliographic/lite/genintro.html.

SDMX. (2006-2009). Retrieved February 27, 2007 from http://www.sdmx.org.

SDMX. (2006-2009). SDMX Standards. Retrieved January 4, 2009 from http://sdmx.org/?page_id=10.

Snijder, Ronald. (2001). Metadata standards and information analysis: A survey of current metadata standards and the underlying models. Retrieved September 10, 2008, from http://www.geocities.com/ronaldsnijder.

XBRL. (2005-2009). Retrieved January 4, 2009 from http://www.xbrl.org/Home/.

XBRL. (2005-2009). XBRL Specifications. Retrieved August 3, 2009 from http://www.xbrl.org/SpecRequirements/.

Appendix A: DANCE Variables

Category of information	Original variables	Internal name of retained/ new variables	Dublin Core element	Dublin Core definition/guideline or comment
Descriptive	Category			This item was not found useful and was removed in the redesign.
Descriptive	Database	Title	Title	The name given to the resource.
Descriptive		Alternative	Alternative	An alternative name for the resource.
Descriptive	Database URL	Product URL		

https://doi.org/10.23106/dcmi.952109594

Descriptive	Description	Description	Description	Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content, or a free-text account of the content.
Descriptive	Keyword(s)	Keyword	Subject	The topic of the content of the resource. Typically, a subject will be expressed as keywords or key phrases or classification codes that describe the topic of the resource.
Descriptive	Vendor	Publisher	Publisher	Recommends against using same name as in "creator." Suggests "publisher" for organizations and "creator" for individuals. If ambiguous, use "contributor."
Descriptive	Vendor URL	Vendor URL		
Descriptive		Creator	Creator	Entity responsible for the content.
Descriptive		Record ID	Identifier	An unambiguous reference to the resource within a given context.
Coverage		Date created	Date	
Coverage		Date range available	Available	Available should be used in the case of a resource for which the date of availability may be distinct from the date of creation, and the date of availability is relevant to the use of the resource.
Coverage		Geographical coverage	Coverage	The spatial or temporal topic of the resource.
Coverage	Status	Data set status	Accrual policy	The policy governing the addition of items to a collection.
Storage		Data origination	Source	In general, include in this area information about a resource that is related intellectually but does not fit easily into a relation element.
Storage		Related resources	Relation	Reference to a related resource.
Storage	Form of access	Data location		
Storage	Form of access	Format	Format	Physical or digital manifestation of the resource.
Storage		Type	Type	Nature or genre of the content.
Storage		Physical medium	Medium	The material or physical carrier of the resource.
Storage		Update method	Accrual method	The method by which items are added to a collection.
Storage		Software name	Software	A computer program in the source or compiled form.

https://doi.org/10.23106/dcmi.952109594

Storage		Update schedule	Accrual periodicity	The frequency with which items are added to a collection.
Contact	Data contact	Data contact		
Contact	Division	Purchasing division		
Contact		Data requestor		
Contact		Contributing section	Contributor	Use for entities with lesser responsibility or more ambiguous contributions than the entity in "creator."
Contact		Purchasing section		
Contact		Vendor contacts		
License	License contact	License owner	Rights holder	Defined as person or organization holding rights over the resource.
License	License information	Data confidentiality	Rights	Information about rights held in and over the resource.
License		Bibliographic notation	Bibliographic citation	A bibliographic reference for the resource.
License		License agreement	License	A legal document giving official permission to do something with the resource.
Other		Help files	Instructional method	A process used to engender knowledge, attitudes, and skills that the resource is designed to support.
Other		Data documentation		
Other		Document path or URL	Text	A resource consisting primarily of words for reading.
Other		Additional information		This item links to a wiki that is updatable by all users.
Admin		Criticality		
Admin		Cost		Restricted access.
Admin		Payment schedule		Restricted access.
Admin		Contract renewal date		Restricted access.
Admin		Purchase justification		Restricted access.
Admin		Purchase order		Restricted access.
Admin		Reason needed		Restricted access.
Admin		Contract length		Restricted access.
Admin		Record last updated		Not displayed.
Admin		Record updated by		Not displayed.

https://doi.org/10.23106/dcmi.952109594

Appendix B: Vendor Metadata Repository Variables

Internal name of retained/new variables	Dublin Core element	XBRL element	Dublin Core definition/guideline or comment
Variable storage name	Title		The name given to the resource. Typically, a Title will be a name by which the resource is formally known.
Variable name short	Alternative		An alternative name for the resource.
Mnemonic			To link to the MDRM database.
Description	Description		Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
MDRM number			To link to MDRM database.
ID variable	Identifier		An unambiguous reference to the resource within a given context.
Date start			
Date end			
Ordinal position			Display order.
Flow-stock		Flow-stock	
Flow periodicity		Flow periodicity	
Debit/credit indicator		Debit/ credit indicator	
Data type		Data type	Monetary, boolean, text, rate, percent.
Multiplier			This was an element in the initial XBRL requirements but was removed in subsequent versions (for example, thousands of dollars).
Numeric flag			
Derived flag			
Formula		Formula	Formula used if the variable is derived.
Variable length			
Variable precision			
Confidentiality			
Format	Hasformat		A related resource that is substantially the same as the pre-existing described resource, but in another format.
Record last updated			Not displayed.
Record updated by			Not displayed.

Appendix C: MDRM Variables

Internal name	Dublin Core element	XBRL element	Dublin Core definition/guideline or comment
Current variables			
Item name	Title		The name given to the resource. Typically, a Title will be a name by which the resource is

https://doi.org/10.23106/dcmi.952109594

			formally known.
Item short caption	Alternative		An alternative name for the resource.
Mnemonic			A four character mnemonic that identifies the data set or survey. When combined with the MDRM number, it creates a unique identifier.
Description	Description		Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content, or a free-text account of the content.
MDRM number			A suffix that identifies the accounting concept. When combined with the mnemonic, it creates a unique identifier.
ID variable	Identifier		The combination of the mnemonic and MDRM number.
Date start			
Date end			
Confidential			
Item type			This variable will be broken into data type and derived flag.
Record last updated			Not displayed.
Record updated by			Not displayed.
Future variables			
Ordinal position			A combination of the reporting form schedule and line number.
Flow-stock		Flow-stock	
Flow periodicity		Flow periodicity	
Debit/credit indicator		Debit/ credit indicator	
Data type		Data type	Monetary, boolean, text, rate, percent.
Multiplier			This was an element in the initial XBRL requirements but was removed in subsequent versions.
Derived flag			
Formula		Formula	Formula used if the variable is derived.
Collection-level metadata			
Title	Title		The name given to the resource.
Mnemonic	Identifier		
Sub series			
Segmented mnemonics			
Reporting form number	Alternative		Reference to the survey form number.
Other database ID			
Access method			
Status	Accrual policy		
Date start			

https://doi.org/10.23106/dcmi.952109594

Date end			
Record last updated			Not displayed.
Record updated by			Not displayed.
Future collection-level metadata—convert from text to a database			
Description	Description		
Frequency	Accrual periodicity		
Reporting panel	Source		
Data availability	Accrual method		
Confidentiality			
Major series changes	Event		A nonpersistent, time-based occurrence.
Additional information			Link to additional information for complex series.
Public release	Text		
Creator	Creator		
Data location			
Format	Format		
Data contact			

https://doi.org/10.23106/dcmi.952109594