This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

Maximizing the Potential of Data – Modern IT Tools, Best Practices, and Metadata Standards for SBE Sciences

Pascal Heus (pascal.heus@metadatatechnology.com), Metadata Technology North America Inc.
Arofan Gregory (agregory@odaf.org), Open Data Foundation
Version 1.0, October 15th 2010

Abstract

An unprecedented amount of data are produced by agencies and researchers in support of SBE research. However, the existence of data alone is not sufficient to maximize their utility and the need for more accessible, open, and usable data is growing. Meeting such demand is a major challenge for data managing institutions that are, not only facing a number of archetypal issues - such as quality, timeliness, adequate documentation, privacy, security, comparability, and harmonization - but also significant resource constraints. Furthermore, the globalization and virtualization of data and systems calls for a unification of practices to facilitate the process by which information may be delivered over the Internet or exchanged between organizations. New information technologies, tools, best practices, and metadata standards have emerged to meet these needs in a way that is both efficient and responsible, the adoption and diffusion of which can provide producers and researchers with a cohesive approach to data management and knowledge sharing. This in turn will open greater access to data, maximizing their usefulness for policymakers, students, researchers, and journalists. This paper examines how advances in technology and open standards may be leveraged upon to enhance manageability, discovery, accessibility, and usability of SBE data.

Copyright © 2010
Metadata Technology North America Inc., http://www.metadatatechnology.com
Open Data Foundation, http://www.odaf.org

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

Introduction

For centuries, statistical offices, academic and research institutions, governmental agencies, international organizations, and commercial entities have invested significant efforts and resources in the production of data for the advancement of social, behavioral and economic sciences (SBE). These data and derived research outputs, papers, reports and publications, are essential to measure the health of our society, support evidence-based policies, and measure results. They take various shapes, from low level microdata to high level aggregated indicators and are collected from individuals, households, hospitals, enterprises and other sources using surveys or through administrative systems. How such information is compiled by agencies and made accessible to researchers involves a set of long and complex processes. However, the resulting outputs and indicators are crucial for decision makers to formulate national and global policies and to monitor progress or impact, as well as contributing to short, mid-term and long range policy, programming, and planning efforts

At the same time, societal factors and Internet technology are combining to create an unprecedented demand for data. Economic conditions are driving policy makers and researchers to seek for better measures of relevant economic phenomenon while citizens and watchdog organizations call for greater transparency in government, and access to government data - as evidenced by the "data.gov" initiative and elsewhere. Combined with these societal factors, software tools have emerged for discovering, accessing, processing and visualizing data.

The mere existence of data is, however, far from sufficient to ensure the needs of decision makers and other users are met, and that the data are effectively used. Measures of data quality are an essential requirement, a need that has been extensively documented. It is commonly agreed that "usefulness" is a good generic high-level definition of data quality. Data alone have no purpose. To be useful, data first must be accessible to researchers. It must also be usable, i.e. comprehensive documentation describing these data must be available.

While significant progress has been made towards improving overall SBE data quality, accessibility, documentation, exchange, and harmonization across sources remain significant weaknesses. Solutions to address these issues are now available thanks to the emergence in the past decade of SBE specific metadata specifications and their underlying information technology. The existence of such solutions is only a first step, however. The adoption and effective use of standards requires endorsement by key agencies, the availability of supporting tools and infrastructure, the existence of training and reference materials, the active management of changes in practice, and integration into existing processes.

The Standards and Technology Revolution

Rapid progress achieved in the field of information science led by the rise of the Internet has completely reshaped the way we discover, access and manipulate data. This revolution has lead to the availability of technologies that allow for the effective exchange of information between agencies, independent of their geographical location or proprietary infrastructure. Central to this phenomenon are the extensible markup language (XML), web services, and service-oriented architecture. While the SBE community has,

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

compared to other fields, been slow in leveraging such technologies, the gap is rapidly closing – due to the emergence of standards and best practices in this area.

Two core XML complementary metadata specifications are being widely adopted internationally for the management of SBE data:

- The Data Documentation Initiative (DDI) for the management of microdata and low-level administrative data
- The Statistical Data and Metadata Exchange standard (SDMX) for aggregated data, indicators, time series and similar official statistics

In addition, several other metadata specifications have been identified to complement SDMX and DDI, such as:

- ISO/IEC 11179 Information Technology -- Metadata Registries (MDR)
- The Dublin Core Metadata Initiative (DCMI)
- Open Archive Initiative (OAI)
- Metadata Encoding & Transmission Standard (METS)
- ISO 19115 for geographic information

The DDI is an internationally recognized standard for describing social science data. In existence since 2000, it facilitates documenting microdata as well as related tabulated outputs. It can capture information at the survey, file and variable levels to surround statistical data with comprehensive documentation. The latest iteration of the DDI has extended its mandate to the entire life cycle of SBE sciences datasets, encompassing conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.

Focusing on high level aggregated data and time series, SDMX is the result of a decade-long joint effort by several international organizations to meet the needs of international, national and sub-national reporting requirements. It was recognized by the UN Statistical Commission as the preferred standard for the exchange and sharing of data and metadata[1]. The Commission also encouraged implementation by national and international statistical organizations.

As aggregated data often finds their source in microdata, SDMX and DDI have been designed to work hand in hand, complementing each other to ensure that relevant linkages are maintained and that metadata can be carried over or reused from one space to the other.

In addition to providing common structures and language for managing data and metadata, one of the most important outcomes of SDMX and DDI is the ability to leverage on XML technology for what is termed "metadata-driven processing" by which collection, dissemination, management and other operations can be automated. This can include for example generating codebooks or questionnaires, ingesting and processing datasets, quality assurance operations, managing data warehouses, or creating self-configuring websites for querying and viewing data. This type of technology offers major gains in

[1] http://sdmx.org/?p=106

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

efficiency, timeliness and overall quality, and is one of the major goals of standards such as SDMX and DDI.

Recently, there has been expressed interest in the use of the Linked Data ("Semantic Web") technologies for publishing statistical and research data as an alternative to XML technologies. Both SDMX and DDI are being examined as the basis for expressions of the data and metadata using such approach. Ultimately, data producer using SDMX / DDI may publish data and metadata both in XML and Linked Data forms, thus maximizing the usefulness of the data for the end users. The interest in this approach is driven by the "data.gov" initiatives, notably in the US and the UK. These initiatives focus on "open data", to provide a higher degree of transparency and accountability in government. Because RDF and other Linked Data technologies are mainstream Web tools that facilitate data discovery of data, their use in combination with generally accepted specifications such as DDI and SDMX would allow for maximum reusability and comparability of this data by all types of users.

The term "open data", in and off itself, can take on various meanings, and needs to be clearly defined. From our perspective, open data means well-documented statistical information that can be easily discovered, accessed for a legitimate research and statistical purpose, and used in accordance with statistical principles and legislation. It is naïve to think that the simple existence of data in the public domain will make it practically available to researchers for use. There are many impediments to the use of data, and the legal right to use the data is only one of them. This is not a simple issue – data confidentiality is critical in an age when the Internet and other technologies raise significant concerns about privacy. There are guidelines and legislation aimed at protecting respondents, and we feel that these are necessary protections for respondents.

Benefits

The use of modern technologies for data management, coupled with standard models, provide a wide range of benefits. This begins with comprehensive documentation and overall quality enhancements. Going beyond such well-understood advantages however is where the full potential of the approach unfolds itself. This includes:

- The use of machine-actionable metadata to automate processes such as workflows management, code generation, data processing or analysis, quality assurance, publication, and others. This leads to improvement in timeliness, accuracy and integrity of the information;
- Combining standards-based metadata with neutral data storage formats such as ASCII text to ensure long term preservation in a non-proprietary format;
- The reuse of metadata across projects, institution, or national boundaries to foster harmonization and comparability;
- The ability to deliver custom content to users to meet their particular views or needs;
- Leveraging industry standard practices such as service-oriented architecture to facilitate the publication and exchange of information over the web;
- Ease of integration with the Internet and semantic web like technology; and
- The power to carry metadata across domains or lifecycle stages through the inherent transformation capabilities of XML.

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

More importantly, metadata standards and technologies provide a common language and unified framework in a SBE environment dominated by a diversity of practices, software, data storage formats, proprietary databases, and legacy IT infrastructure. They provide a mean for agencies to expose their information to the rest of the world in a well-understood fashion, without the need to completely alter their internal framework. This is particularly crucial in the United States where the national statistical system is highly decentralized and composed of over a hundred agencies.

Recommendations

New IT tools, standards, and best practices for use with SBE data provide potential to maximize the utility of our data – making it easier to discover, use, understand and compare. However, the adoption of these technologies, standards, and practices will require a more disciplined approach on the part of researchers and data producers. The overall process of managing data will be impacted, requiring significant changes.

We see the path forward as including a number of different actions on the part of various players in the SBE community:

- Recognize and endorse SDMX, DDI and other relevant specifications as preferred standards for the management, exchange and sharing of SBE data and metadata;
- Support the implementation and deployment of standard driven IT infrastructures for data and metadata management;
- Establish and enforce policies that require the use of recommended practices and standards for the management and publication of SBE data;
- Maintain public metadata registries to facilitate the discovery and access to data;
- Support the development of open source and commercial standard based tools and technologies;
- Establish a community of practices among data managing agencies that foster the use of standards, both in terms of technology and harmonization of metadata;
- Foster the adoption of best practices among researchers who are themselves secondary data producers;
- Support virtual communities for knowledge capture and exchange, leveraging social networking technologies;
- Support the production of guidelines, best practices and training materials for metadata management;
- Advocate and support the use by agencies of industry standard XML and services oriented IT infrastructure;
- Promote collaboration with domain experts and agencies; and
- Leverage Rich Internet Applications, social networking, virtual spaces, and other emerging web technologies to engage with the researchers, general public, policy makers and foster participation and contributions.

It is important to recognize that adoption challenges are not only technical but also organizational and political in nature. A comprehensive approach must therefore combine technical expertise with strong leadership and sound change management techniques. These challenges are however similar to those that

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

have been overcome in many other domains, and are faced by any industry looking at effectively leveraging advances in technology and the existence of open standards.

Conclusions

What we need is not more SBE data but more useful SBE data, i.e., high-quality data surrounded by relevant metadata that leverages internationally recognized standards and technologies to ensure safe and effective access and use of the data, knowledge capture, and sharing. The metadata should be comprehensive and ready for both human and machine consumption, to maximize potential benefits. Access to data should be as “open” as possible but in accordance with agreed upon statistical principles centered on data security and respondent confidentiality. With increases in efficiency and better data management techniques, it becomes possible to responsibly meet the growing demand for data, to the benefits of society in general.

Technological progress and significant efforts by key players in the SBE community have made new tools and techniques available to the broad community. Their adoption and integration into the existing practice of data manager or users is a significant challenge presented to the next decade.

Further reading:

- "Metadata", Arofan Gregory (ODaF), Pascal Heus (ODaF), German Council for Social and Economic Data Working Paper no. 57/2009, March 2009, http://www.ratswd.de/download/workingpapers2009/57_09.pdf
- “Combining Metadata Standards: Approaches and Benefits”, Arofan Gregory, Open Data Foundation, Work Session on Statistical Metadata (METIS) (Geneva, Switzerland, 10-12 March 2010), http://www.unece.org/stats/documents/ece/ces/ge.40/2010/wp.3.e.pdf
- “The Common Metadata Framework”, http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework