Identifiers, Persistent Identifiers, and the Evolving Demand for Standard Metadata

PID Workshop, Berlin
May 7/8, 2012
Arofan Gregory
Open Data Foundation/Metadata Technology

A Fundamental Question…

What are we doing?

The Search for Truth

Eastern Version:

The Search for Truth…

Western Version:

What are we doing?

We are supporting evidence-based decision-making.

The Data Archive Community

- Strong understanding of data and how it can be used in research
- Strong understanding of issues around data confidentiality and disclosure control
- Strong understanding of data preservation
- Not-so-strong understanding of how technology can be best utilized

A Changing Society

- We live in a society which is being profoundly affected by technology
- Data archives (and producers and disseminators) all face challenges driven by changing needs and expectations
- The question is: “How do we continue to provide a relevant, useful service in the face of technology-driven change?”

Children of the 1960s' Technology

Children of Today's Technology

What Has Changed?

- The expectation of *visibility:*
  - *If it's not on the network, it doesn't exist!*
- The expectation of *connectivity:*
  - *Everything is linked to related things*
- The expectation of *immediate usability:*
  - *When I follow the link, it allows me to interact meaningfully with the thing at the other end*
  - *Otherwise, it doesn't exist!*

What Does this Mean for Us?

- We must enable centralized, portal-based search across organizations and domains
  - The “Google” model
- We must support the linking of related resources
  - “Enhanced publications” – data, research, documentation, etc. all linked
  - Support user-enhancement of the links
- We must support useful behaviors for the data we link to
  - Both for human beings and machines

Specifically…

- Global identification, citation, and resolution
  - PIDs, URNs, URLs…
  - Must work with different technology platforms (XML, RDF, etc.)
- Rich metadata, in standard, machine-processable formats
  - DDI, SDMX, etc.
- Standard infrastructure
  - For discovery of resources by humans and machines
  - For utilization of resources by humans and machines

What is the Danger?

Disintermediation:

"In economics, **disintermediation** is the removal of intermediaries in a supply chain: 'cutting out the middleman'. Instead of going through traditional distribution channels, which had some type of intermediate (such as a distributor, wholesaler, broker, or agent), companies may now deal with every customer directly, for example via the Internet." (Wikipedia)

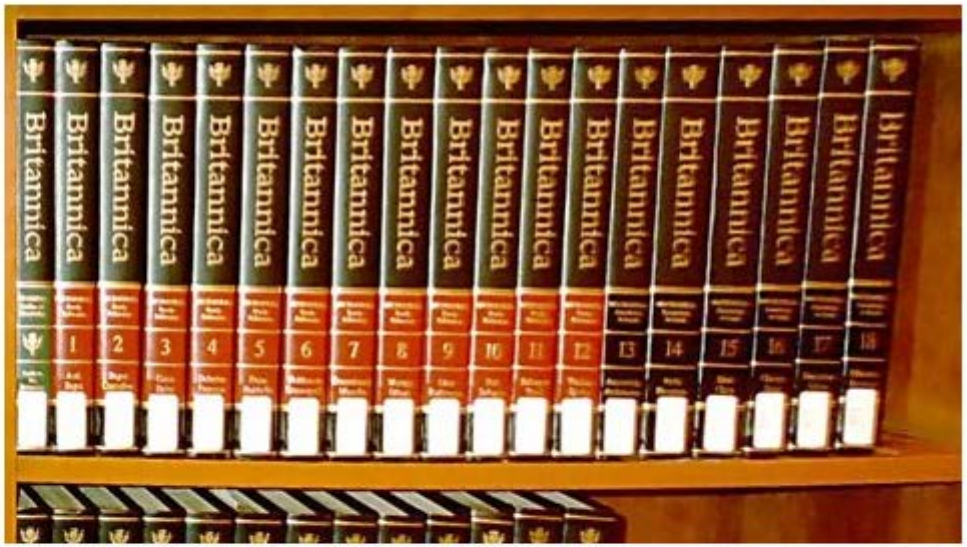
Convenience versus Quality

- At a certain point, people will use the most convenient resource, even if of lower quality
- We see examples in official statistics
  - Businesses and national statistics example
- There are pressures in terms of resources
  - Survey data collection is very expensive and slow!
  - “Aggregated” data from credit cards, cell phones, and Internet use (etc.) is sold by vendors cheap, although of unknown quality and provenance

By Joanna Stern
@joannastern

Mar 13, 2012 7:23pm

Encyclopaedia Britannica Kills its Print Edition



Encyclopaedia Brittanica. Image credit: WikiMedia Commons

The first Encyclopaedia Britannica was printed in 1768. And now, 244 years later, it has been printed for the last time. At least as a set of bound books.

Its publisher has announced that it will no longer be publishing the print

Other Important Trends

- Open Data movements
  - Driven by a mis-trust in government and a demand for increased transparency
  - Largely coming from the technology sector (W3C, Linked Data community)
  - Technology optimized for “connectivity” – RDF
  - DDI and SDMX are both proactively working with the Linked Data community: Data Cube Vocabulary, DDI-RDF for classifications, discovery metadata

Other Important Trends (cont.)

- Borders between domains are vanishing
  - Examples in medical research and social science (MIDUS example)
  - Users want data merged across domain boundaries
  - Non-survey data sources are increasing in importance (bio-markers, administrative registers, etc.)
- Increased interest in microdata access
  - Especially in official statistics
  - Driven by the financial crisis – sudden interest in labor force and household surveys among economists

Many Interesting Initiatives

- Visibility: Data without Boundaries
- Visibility: SDMX Global Registry
- Connectivity: Data Cube Vocabulary/DDI-RDF
- Usability: DDI and SDMX
- Usability: GSBPM, GLBPM, and GSIM

Data without Boundaries

- Focused on all barriers to microdata access and use within Europe
  - 7th Framework funding from European Commission
  - Includes both CESSDA archives and national statistical organizations (liaison with Eurostat and OECD)
  - Addresses many issues: legal and governance, confidentiality, technology
  - Prototyping a pan-European data portal
  - Interest in future virtual secure data access from multiple European sources

SDMX Global Registry

- SDMX ("Statistical Data and Metadata Exchange") is a well-adopted standard within official statistics
  - Driven by international organizations (IMF, BIS, World Bank, OECD, UN Statistical Division, Eurostat, ECB)
  - Focus is on public, aggregate statistics and related metadata
- Highest priority is on a single global portal for discovery of and access to all international statistical data and metadata
  - Currently prototyping, reviewing design with SDMX Technical Working Group

RDF and Linked Data: Data Cube and "DDI-RDF"

- Data Cube Vocabulary: now part of the W3C "open data" standards, in process
  - Based on SDMX's model and SKOS (among others)
  - Starting to be widely used
  - Proactive support from within the SDMX community
- "DDI-RDF" vocabularies
  - To be published by the DDI Alliance (in development)
  - SKOS-based extensions for statistical concepts and classifications
  - Discovery model for metadata about microdata
  - Integrated with Data Cube Vocabulary

DDI and SDMX

- These are the dominant metadata standards for implementation
  - SDMX for official aggregate statistics
  - DDI for research data/microdata
- SDMX provides strong infrastructure support
  - Standard web-services and registry-based architecture
- DDI is “leaking” into areas such as medical research and environmental data
  - Holistic, life-cycle approach is very useful for data management
  - Good for merging data across domains

GSBPM, GLBPM, and GSIM

- DDI Unified Lifecycle Model was used as the basis for the Generic Statistical Business Process Model (GSBPM)
  - A more detailed reference model of statistical data production
  - Published by UN/ECE METIS workshop, widely adopted
- GSBPM used as the basis for the "Generic Longitudinal Business Process Model" (GLBPM)
- The Generic Statistical Information Model (GSIM) now being produced
  - A generic view of all data, metadata, and paradata needed in statistical production
  - Uses DDI, SDMX, Neufchatel, and ISO/IEC 11179 as inputs (among others)
  - To be implemented using SDMX and DDI
- A standard for capturing upstream metadata, based on common standards, could be leveraged by archives to collect otherwise expensive metadata

The Way Forward

- More of the same!
  - Many positive directions have been identified
  - Much good work is being done
  - Continued collaboration with official statistical organizations needed
- Broader recognition of changes in the technology landscape are needed
  - The expertise of the archives needs to be pro-actively injected into technology initiatives around data
  - Expertise of technologists needs to be injected into the archives
- This goes beyond issues of persistent identification
  - If resources are not immediately useable, they don't exist!
  - This requires rich metadata in standard, machine-actionable formats (repeat as needed!)

The Challenge…

Are we doing this?

Or this?

Questions?