Twenty Years of Metadata: Lessons from the First Two Decades of the Web

Stuart Weibel

University of Tsukuba Visiting Scholar

May 13, 2011

OUTLINE

- The Context
- Dublin Core in the Metadata Matrix
- What we did right
- The major impediments
- A few words about models
- What about the future?

Image: Carved figures (Morikawa Toen), Tokyo National Museum

THE CONTEXT

- When I started working at OCLC in 1985:
  - I was 4 years away from my first email address
  - A PC hard drive wasn't large enough to store a single high resolution digital image.
    (which was ok, because…)
- Cameras still used film
- Cell phones were suitcase-sized
- MARC Cataloging stood alone as the discovery tool for intellectual assets of libraries
- No end-user access to the global library catalogs

me… circa 1994

AND NOW?

- A cell phone has more computing power than the Space Shuttle
- An iPod will hold WorldCat
- Bandwidth is more important than computing power
- The library is still mostly mired in MARC
- There are many metadata standards (mostly struggling for traction)
- People (mostly) find things with Google
- but….

METADATA IS MORE THAN JUST SEARCH

Metadata-dependent actions
Describe
Access
Encode/Render
Preserve
Rights Management
Administer
“Bind” digital pages in digital books

50 YEARS OF METADATA

MARC standards (library metadata)

OCLC founded (shared library cataloging)

ARPANET Operational - forerunner of the Internet

Networking diffuses throughout academia

The Web begins...

my first email address

FRBR work begins

First Dublin Core Workshop

DCMI established

Google is founded

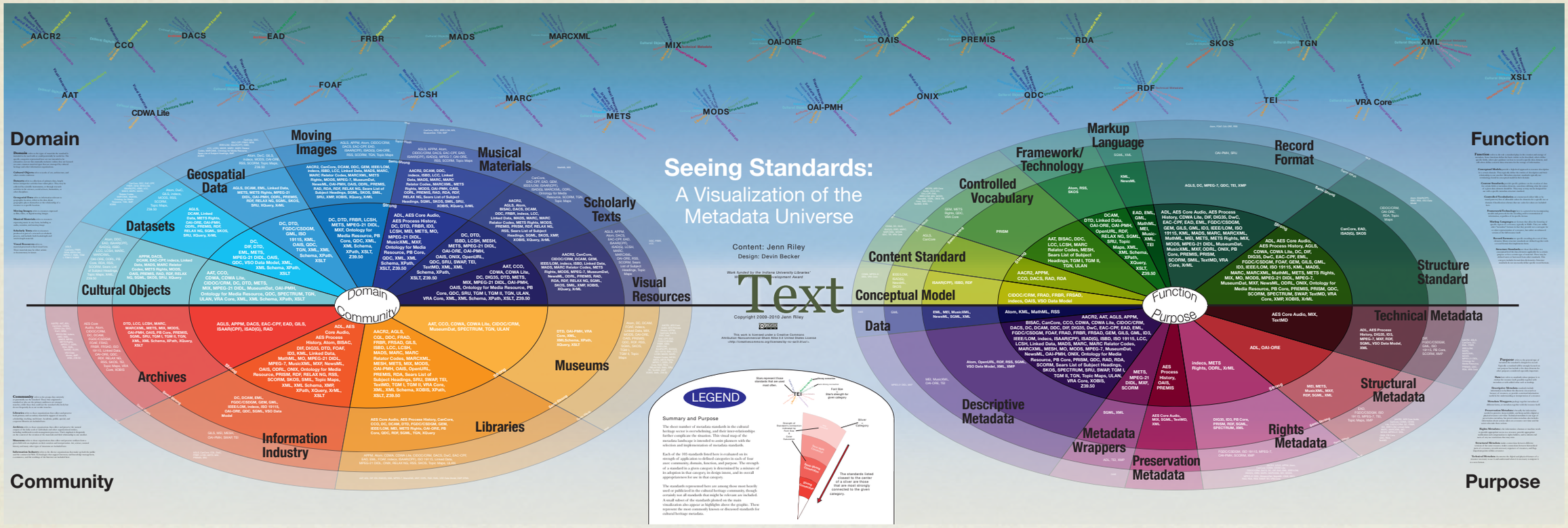
First Dublin Core Conference (Tokyo)

WorldCat introduced

RDA introduced

1960s | 1970s | 1980s | 1990s | 2000s

THE CONFUSION: HOW BAD IS IT?



"This visual map of the metadata landscape is intended to assist planners with the selection and implementation of metadata standards."

http://www.dlib.indiana.edu/~jenlrile/metadatamap/

JENN RILEY'S METADATA MAP

- 105 standards
- 30 most common across the top (3 predate the Web)
- some share common models… most do not
- much overlap
- many work together
- Who among us can choose rationally from the array of standards, platforms, technologies?
- Will the results have any reasonable expectation of interoperability?

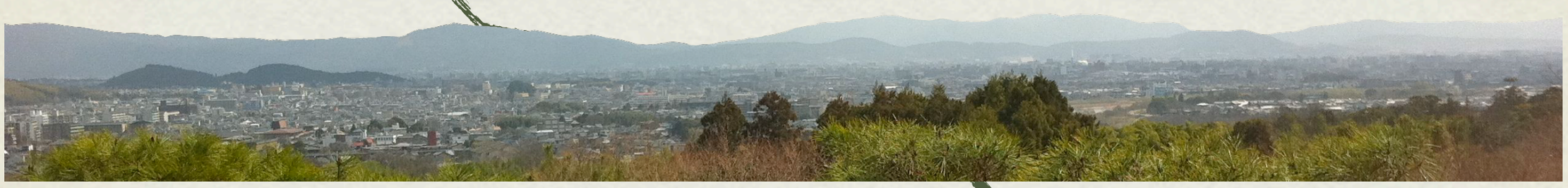
THE REAL WORLD IS NOT STANDARDS-CENTRIC

Metadata-dependent actions	Standard
Describe	MARC, DC, MODS, RDA, LCSH, MeSH….
Access	HTTP, FTP….
Encode/render	RDF, media-type dependent (many)
Preserve	PREMIS
Rights Management	CC licenses, eCommerce systems
Administer	METS, MARC….
“Bind” digital pages in digital books	METS, eBook standards

Information Entities (ex.)
Agents (persons, corporate entities, devices)
Events
Time intervals or eras
Concepts
Collections
Media-types
Structured data type

THE MAP IS MUCH MORE COMPLICATED

"This visual map of the metadata landscape is intended to assist planners with the selection and implementation of metadata standards."



"selection and implementation of metadata standards requires a clear understanding of the information entities, the standards, and the functional requirements of the system under design"

Image: Kyoto horizon from above the Tenru-ji Temple

DUBLIN CORE IN THE METADATA MATRIX

- The first metadata standard for the Web
- General and cross-disciplinary
- Simple starting place, but extensible
- International and multilingual
- Consensus-driven (bottom-up, rather than top-down)

Image: Jomon Pottery, Tokyo National Museum,

THINGS WE DID RIGHT

- We didn't call it '*cataloging*' (*Web*, not libraries)
- A hybrid of technical engineering and social engineering
- International - Major events on 5 continents, element definitions in 20+ languages (maintained in Tsukuba)
- Separated syntax and semantics
- Built a community of practice
- *About* the right level of complexity for a core element set

Image: Harajuku train station platform, Tokyo

IMPEDIMENTS THAT TRIPPED US UP

- Too many syntaxes to support (HTML, XML, RDF-XML)

- No common data model
  but we tried hard:
  data model group,
  architecture group,
  abstract model,
  Singapore Framework…

- Without a data model, the story we told was not consistent: confusion resulted

- Without a data model, details of implementation become arbitrary (and less interoperable)

Image: Netsuke, Tokyo National Museum

DATA MODELING: WHAT IS IT?

- Entity-relationship model defines the important concepts or things (entities), and the relationships among them
- A model is a model, not reality
- Designed to solve a problem, not to emulate the real world
- The complexity of the model should be mapped to the ***problem***, not to ***reality***
- Identifying the right level of abstraction is an art

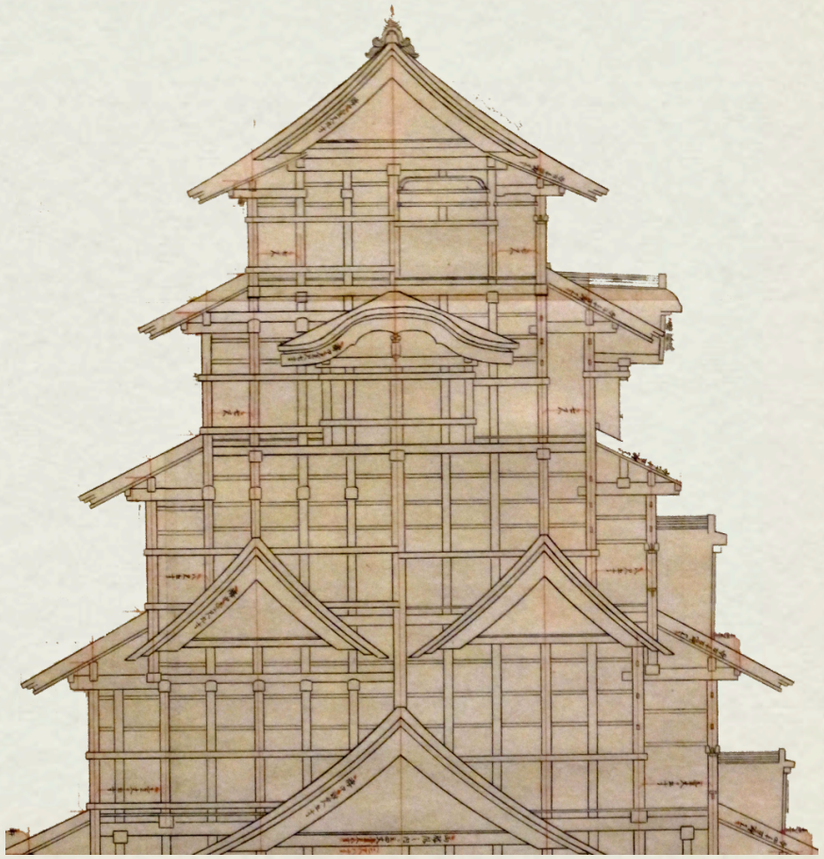


Image: Edo Museum

DATA MODELING: WHY IS IT NECESSARY?

- Without a shared understanding of the important entities, and the relationships among them, systems will not interoperate easily

- Cross-walks become necessary: clumsy, inaccurate, inefficient

Changing rail car 'bogeys' on the China/Mongolia border

AN EXAMPLE OF MODELING MISMATCH

Citation information

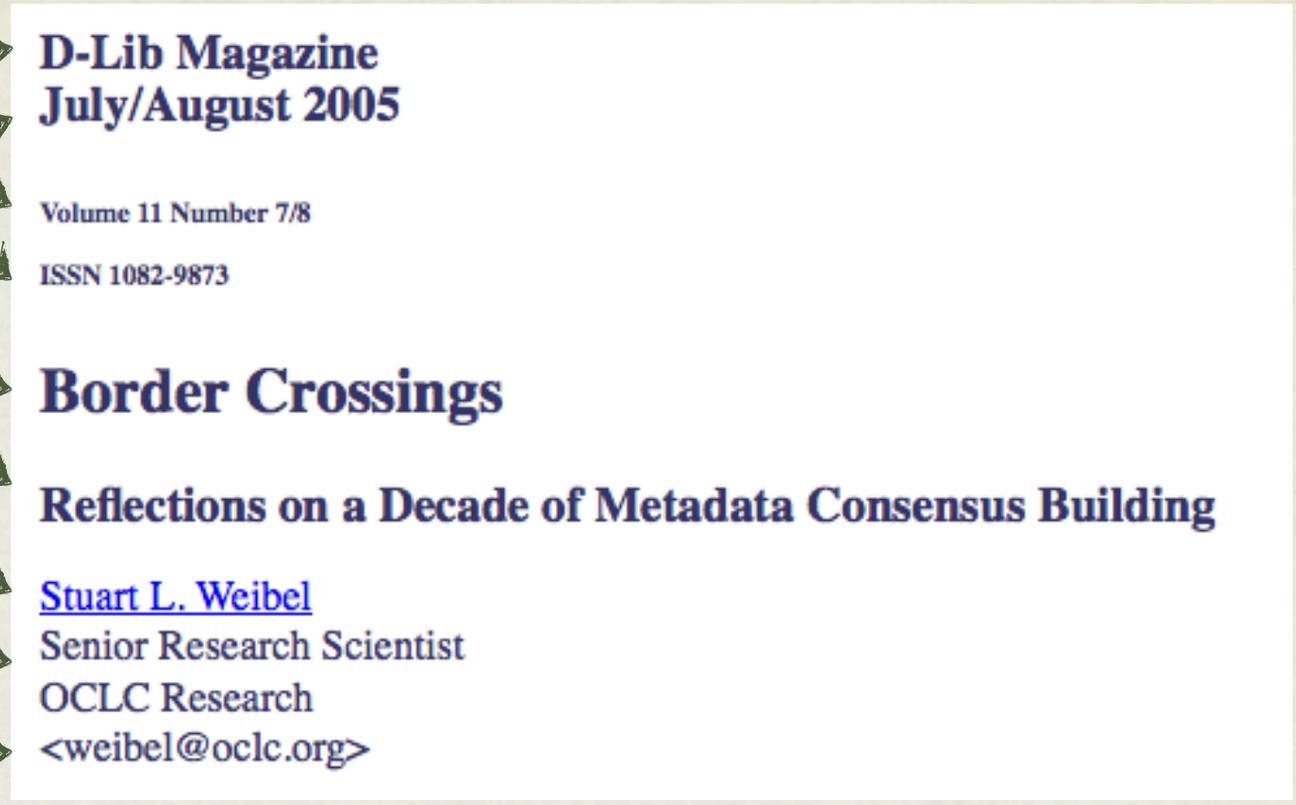
Date

Title

Author

Affiliation

Email address



D-Lib Magazine
July/August 2005

Volume 11 Number 7/8

ISSN 1082-9873

Border Crossings

Reflections on a Decade of Metadata Consensus Building

Stuart L. Weibel
Senior Research Scientist
OCLC Research
<weibel@oclc.org>

- Which of the attributes are Dublin Core?
- Is “email address” an attribute of the resource, or the person?
- Should there be a distinction between Title and Subtitle?

IS DUBLIN CORE WELL-MATCHED TO THE PROBLEM OF BIBLIOGRAPHIC DESCRIPTION?

- It is too simple to capture the precision of detailed bibliographic description
- BUT… It is ***good enough*** for many purposes, including the description of most simple internet resources
- The trade-off between perfect matching of model and problem, and simplicity of use is always a compromise
- DC was intended for general resource description, not to replace MARC

THE PROBLEM WITH MODELS

- Matching the complexity of models to a diverse and evolving problem is challenging, and full of compromises
  - too much complexity leads to failure (*creeping elegance*)
  - too little complexity leads to failure (insufficient richness to solve the problem)

- HOW DO YOU KNOW WHEN IT IS RIGHT?

Image: figures from a model in the Kyushu National Museum

CONCEPTUAL MODELS IN THE LIBRARY WORLD

FRBR and FRAD	The dominant models for bibliographic and authority data
OAIS	Reference model for Open Archive Information Systems
CIDOC CRM	Conceptual Reference Model for cultural heritage documentation
Dublin Core Abstract Model	Largely unintelligible data model for Dublin Core instance data
Singapore Framework	A vague framework describing levels of metadata interoperability

THE NEXT CHAPTERS IN THE WEB METADATA STORY…

- …are being written in the W3C Incubator Group on Library Linked Data (http://www.w3.org/2005/Incubator/lld/)
- Many questions:
  - Will the data be open?
  - Who will maintain it?
  - Is semantic web infrastructure stable?
  - Can existing metadata be integrate seamlessly into the web?
  - Can a model be agreed upon?
  - Will we ever have interoperability across domain silos?

Image: Stone Monk in the Nezu Museum Garden

誠にありがとうございます

stuart.weibel@gmail.com

http://weibel-lines.typepad.com

@stuartweibel on twitter

stuartweibel on Facebook

all photographs by the author

Image: Lantern overlooking the Irises in the Nezu Museum Garden