

Syracuse University

From the Selected Works of Jian Qin

April, 2013

Functional and Architectural Requirements for Metadata

Jian Qin, *Syracuse University*

Alex Ball

Jane Greenberg, *University of North Carolina at Chapel Hill*



Available at: https://works.bepress.com/jian_qin/8/

Functional and Architectural Requirements for Metadata:

Supporting Discovery and Management of Scientific Data

Jian Qin

School of Information
Studies
Syracuse University
USA

Alex Ball

Digital Curation Centre
UKOLN
University of Bath
UK

Jane Greenberg

School of Information and
Library Science
University of North Carolina
Chapel Hill, USA

Research Data Access & Preservation Summit
Baltimore, MD, 2013

Metadata standards for scientific data



Ecological
Metadata
Language (EML)



ClinicalTrials.gov
Protocol Registration System

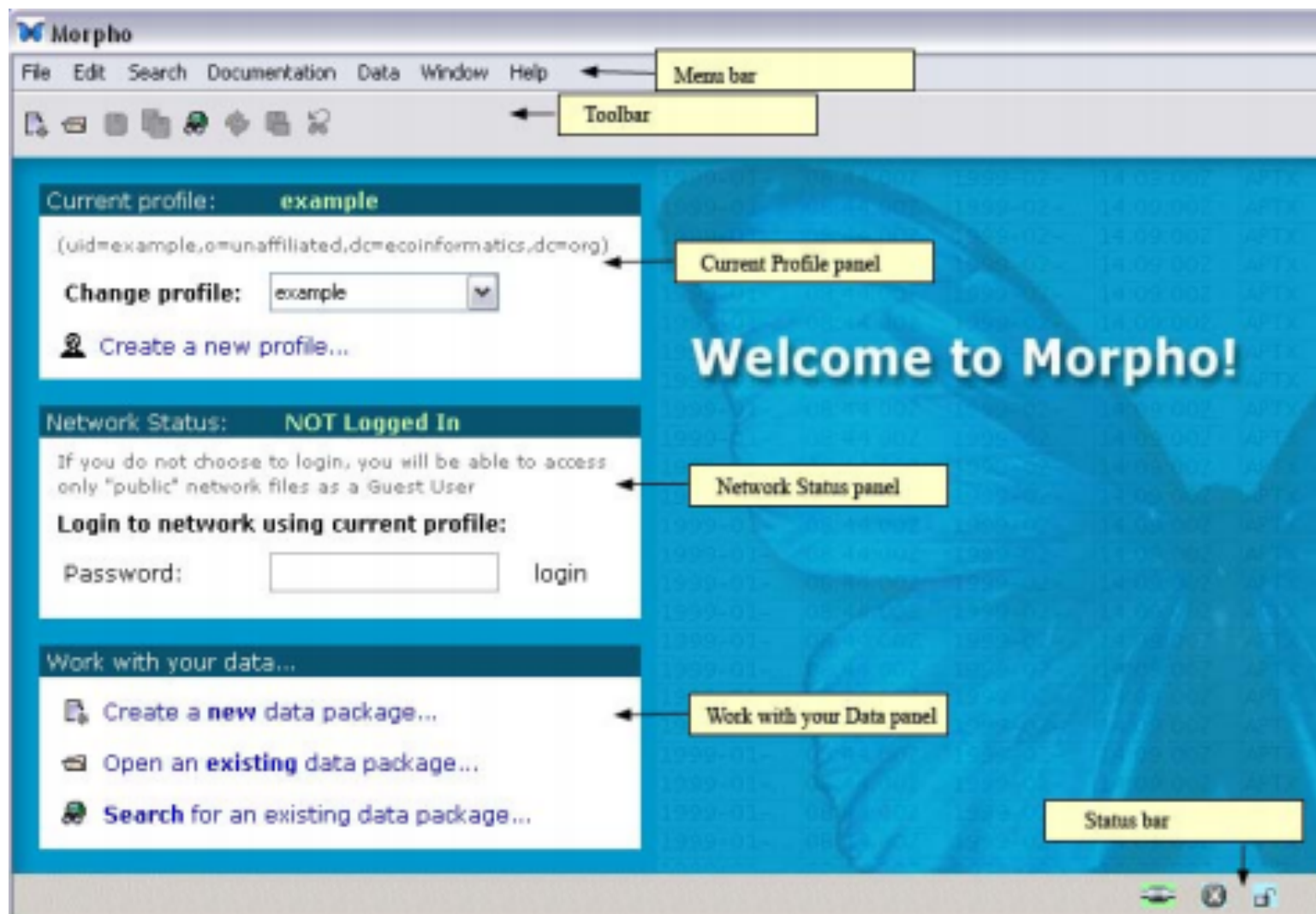


Biodiversity
Information
Standards
TDWG

Access to Biological
Collections Data –
ABCD

Darwin Core

Many tools have been developed...



Many tools have been developed...

EPA Metadata Editor v2

Basic Data Set Information | Quality, Coordinate System, and Attribute Information | Distribution & Metadata Information

Citation

* **Origin:** United States Environmental Protection Agency, Office of Environment default

* **Title:** EPAs Web Feature Service for All Sites listed in the Facility Registry System

Publisher

* **Published by:** U.S. Environmental Protection Agency default

* **Published at:** Washington, DC * **Date:** 2006 today

Online Linkage

** **Primary Linkage:** <http://intramap33.rtpnc.epa.gov/wfsconnector/com.e> ✓

** **Secondary Linkage:** <http://www.epa.gov/frs/index.htm> ✓

Description

* **Abstract:** This web feature service provides location and attribute information on facilities, sites or places subject to environmental regulations or environmental interest. There are over 1.9 million records available

* **Purpose:** The purpose of this web feature service is to provide users with access to information stored in EPA's Facility Registry System (FRS). The information in FRS may be used to track sites, facilities, or areas

Supplemental Info: This is an EPA-Produced Data Set. This data set is one in a s

Time Period

* **Date:** 2006 today * **What is data current to:** publication date default

* **Progress:** In work * **How often is data updated?** Monthly

Bounding Box

N:71.490303 E:-67.434525 S:17.305182 default

* N: 71.490303 * E: -67.434525

* S: 17.305182 * W: -173.129822

Keywords

* ISO default	** EPA default	** Place default
biota	Surface water	Alabama
boundaries	Toxics	Alaska
climatologyMe	Transportatic	American San
economy	Waste	Arizona
elevation	Water	Arkansas
environment	emergency re	California
farming	environmenta	Canada

Data Set Constraints

* **Access:** This service is provided for EPA inter default

* **Use:** None. Please check sources, scale, a default

* **Security Classification:** No Confidentiality (from meta default

Contact

☐ Primary Person ☒ Primary Organization

U.S. Environmental Protection Agency | EP default

open dbase default

Set Default Spell Check Validate EPA

YELLOW * mandatory GREEN ** mandatory if applicable BLUE optional Click on text to link to element description Save & Close Discard & Close Help

Many tools have been developed...

Results

Title: **Results:Baseline Overview: Edit Baseline Participants**

Title: Amphotericin Alone or in Combination With Fluconazole for AL... NCT00145249 ID: 03-154

Number of Participants Started First Participant Flow Period:

Total:	AmphoB standard	AmphoB+Fluc400	AmphoB + Fluc800
143	47	48	48

	AmphoB standard	AmphoB+Fluc400	AmphoB + Fluc800
Overall Number of Baseline Participants	45 NOTE : The number of baseline participants is not consistent with the number of participants who started first period.	47 NOTE : The number of baseline participants is not consistent with the number of participants who started first period.	49 NOTE : The number of baseline participants is not consistent with the number of participants who started first period.

OK Cancel

Measure Dispersion:

"Not Applicable" for other measure types.

Standard Deviation

Unit of Measure:

If the Measure Type is "Number", the Unit of Measure is typically "participants".

years

OK Cancel Delete

13

Motivation for adoption

- Standardize the format and terminology of metadata
- Enable fast and effective discovery of datasets across different data repositories
- Enable data sharing, reuse, and preservation
- Provide information for obtaining datasets from the data owners

Hindering factors

Metadata standards for scientific data have:

- large numbers of elements
- many layers in structure
 - “Unwieldy to apply”
- been created for manual data entry, rather than for automatic generation

Effects on metadata generation:

- steep learning curve
- difficult to automate metadata generation
- unnecessary duplicate data entry
- high costs in time, resource, and personnel expertise

Same entity data repeated in the same record...

Seamless Daily Precipitation for the Conterminous United States

Metadata:

[Identification Information](#)

[Data Quality Information](#)

[Spatial Data Organization Information](#)

[Spatial Reference Information](#)

[Entity and Attribute Information](#)

[Distribution Information](#)

[Metadata Reference Information](#)

Citation_Information:

Originator:

Mauro Di Luzio, Blackland Research Center, Texas Agricultural Experiment Station, Texas A&M University System, Temple, Texas

Publication_Date: 200710

Title: Seamless Daily Precipitation for the Conterminous United States

Edition: First

Data_Set_Credit:

Mauro Di Luzio Blackland Research Center, Texas Agricultural Experiment Station, Texas A&M University System, Temple, Texas

Metadata_Reference_Information:

Metadata_Date: 20071001

Metadata_Contact:

Contact_Information:

Contact_Person_Primary:

Contact_Person: **Mauro Di Luzio**

Contact_Organization:

Texas Agricultural Experiment Station, Blackland Research Center

Point_of_Contact:

Contact_Information:

Contact_Person_Primary:

Contact_Person: **Mauro Di Luzio**

Contact_Organization:

Texas Agricultural Experiment Station Center

...and they are already in...

Publication record associated with the data

LOCUS SCU49845 5028 bp DNA
DEFINITION Saccharomyces cerevisiae TCP1-bp
(AXL2) and Rev7p (REV7) genes, c
ACCESSION U49845
VERSION U49845.1 GI:1293613
KEYWORDS .
SOURCE Saccharomyces cerevisiae (baker'
ORGANISM Saccharomyces cerevisiae
Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE 1 (bases 1 to 5028)
AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE Cloning and sequence of REV7, a gene whose function is required for
DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL Yeast 10 (11), 1503-1509 (1994)
PUBMED 7871890
REFERENCE 2 (bases 1 to 5028)
AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE Selection of axial growth sites in yeast requires Axl2p, a novel
plasma membrane glycoprotein
JOURNAL Genes Dev. 10 (7), 777-793 (1996)
PUBMED 8846915
REFERENCE 3 (bases 1 to 5028)
AUTHORS Roemer,T.
TITLE Direct Submission
JOURNAL Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
Haven, CT, USA

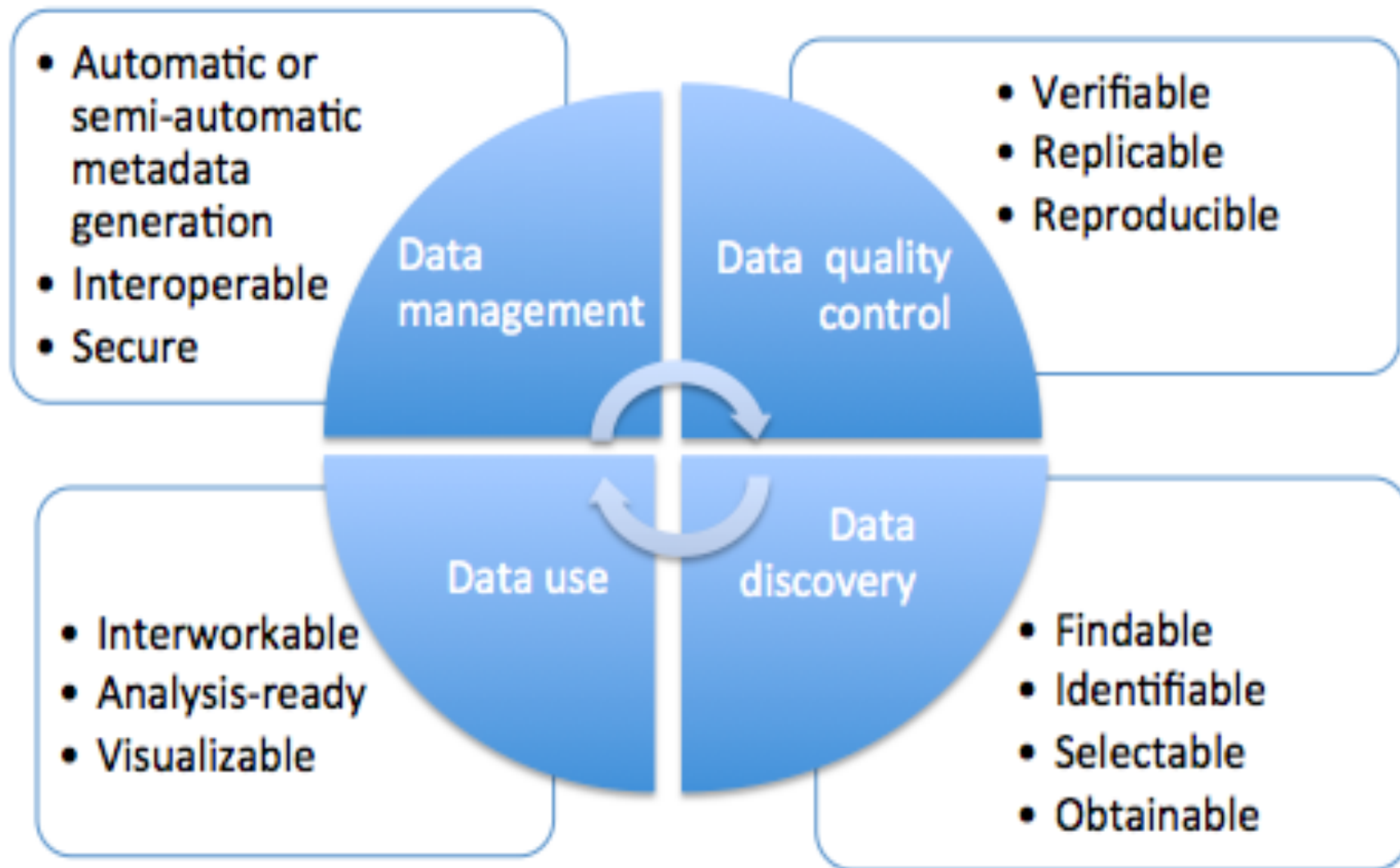
Research questions

- What functions do metadata standards for scientific data serve?
- How should metadata standards for scientific data be modeled to support these functions by meeting the associated requirements?

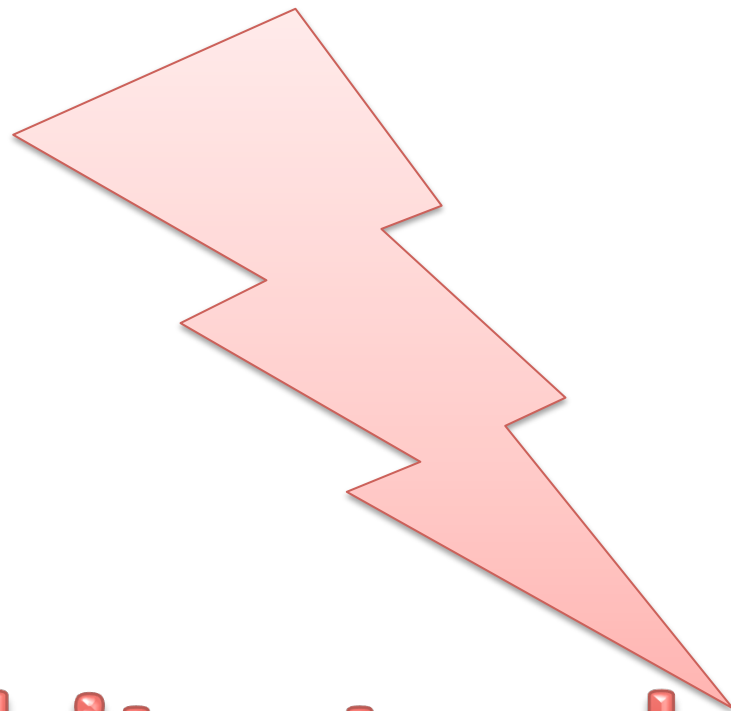
Functions expected

- Resource discovery and use,
- Data interoperability,
- Automatic and semi-automatic metadata generation,
- Linking of publications and underlying datasets,
- Data/metadata quality control, and
- Data security.

Metadata requirements for scientific data



Functional view



Architectural view

Architectural view

Identity metadata

Agent / Investigation / Publication / Data set or collection

Semantic metadata

Taxonomy / Thesauri / Classification / Ontology /
Relations

Scientific context

Workflow /
Provenance /
Parameter / Processing

Geospatial metadata

Temporal metadata

Miscellany

Identity metadata

Globally, uniquely identify entities

- Person: researcherID, URI, FOAF, ORCID
- Institution: ORCID, URI
- Data object: DOI, Handle, URI
- Associated publication: DOI

Build a metadata infrastructure service

- Name repositories
- Linked data architecture
- Customizable research group/community member name lists

Semantic metadata

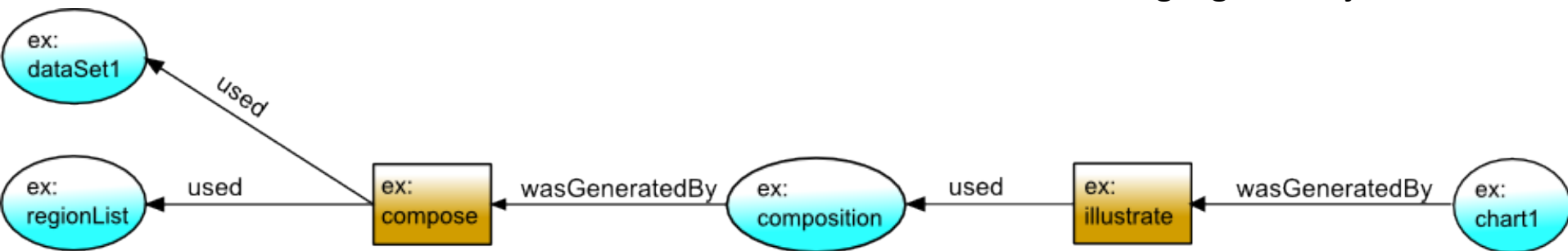
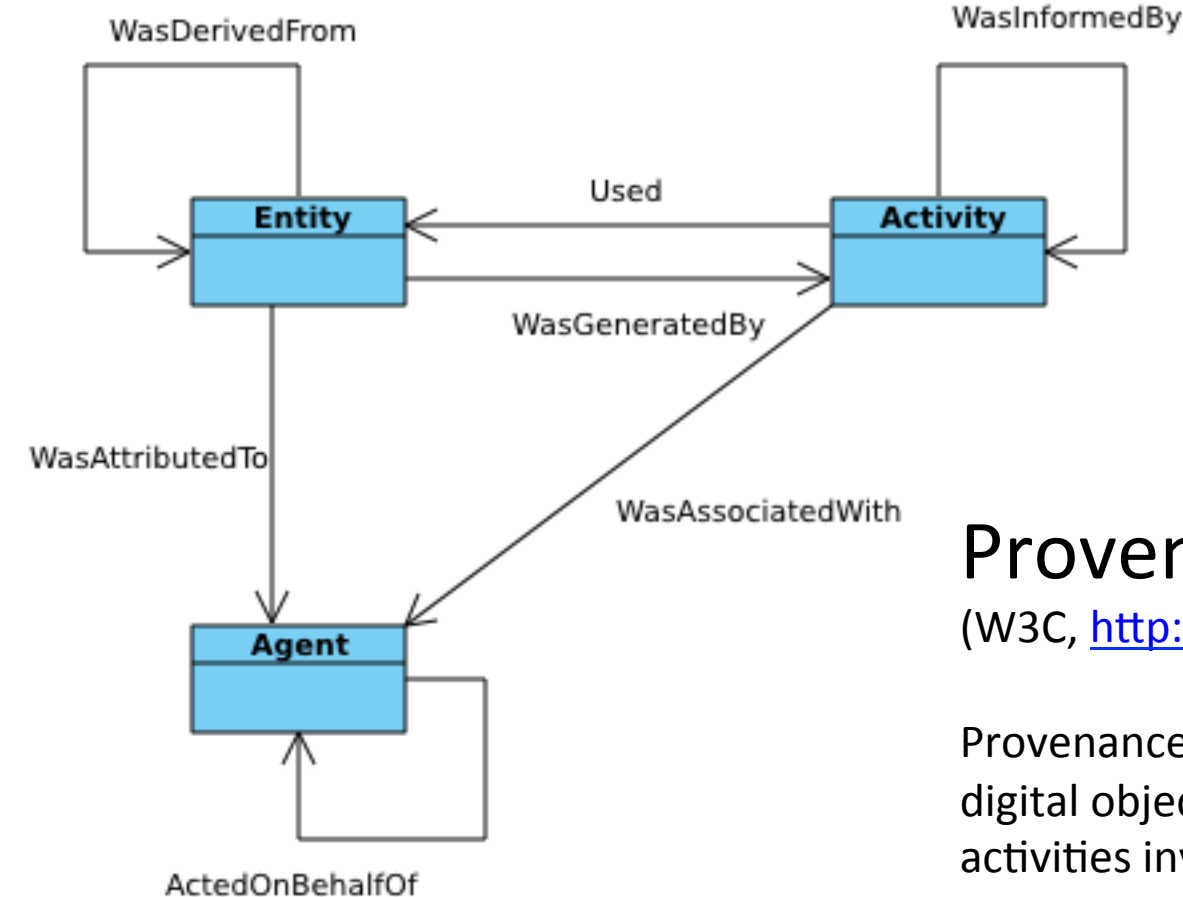
- Large semantic resources available in linked data format, but usually not suitable for representing scientific data because they are designed for publications, especially books and journals (containers)
- Format is contemporary but the content is far from it

Smaller, specialized semantic resources are necessary for automatic semantic metadata generation

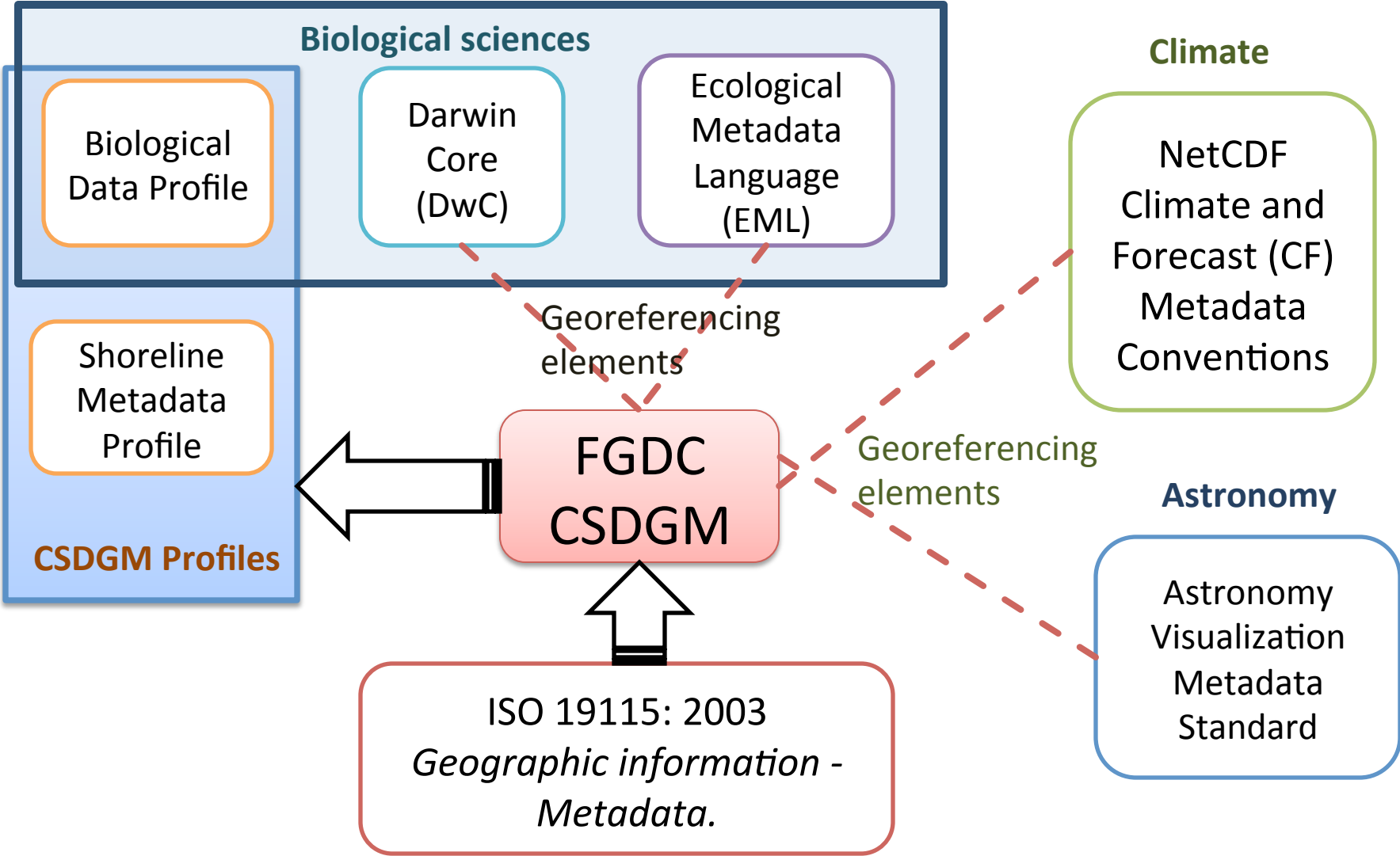
Contextual metadata

Provenance data model (W3C, <http://www.w3.org/TR/prov-primer/>)

Provenance data represent the origins of digital objects and describe the entities and activities involved in producing and delivering or otherwise influencing a given object.



Geospatial metadata



Temporal metadata

- Mean solar time
- Civil time
- GPS time
- Terrestrial time
- Atomic time
- ...
- Geologic time
- Different measurement systems result in different units and format
- Conversion between systems

Three Principles

- The least effort principle
- The infrastructure service principle
- The portable principle

TABLE 1. Mapping data user tasks with metadata functions and architectural building blocks

Data user tasks	Metadata function	Architectural building block
Discover	Descriptive metadata	Identity and semantic metadata
Identify	Descriptive metadata	Identity metadata
Select	Descriptive, technical metadata	Identity, semantic, scientific context, geospatial, temporal, miscellany metadata
Obtain	Descriptive metadata	Identify metadata
Verify	Descriptive metadata	Scientific context metadata
Analyze		Scientific context, geospatial, and temporal metadata
Manage	Descriptive, administrative, structural, and technical metadata	Identify, semantic, scientific context, geospatial, temporal, miscellany metadata
Archive	Descriptive, administrative, structural, and technical metadata	Identify, semantic, scientific context, geospatial, temporal, miscellany metadata
Publish	Descriptive metadata	Identity, semantic, scientific context, geospatial, and temporal metadata
Cite	Descriptive metadata	Identify metadata

Development potentials

- An infrastructure of metadata services
 - Entities as linked data
 - Tools for “slicing” members by research group, community, or institution to customize the entity set
 - Tools for grabbing entity data from existing resources through interoperability protocols

Application scenarios

- Cross-domain discovery and verification
- Automatically populating entity information from customized slices of entities into metadata records
- And more...

Conclusion

- Scientific data are inherently complex and diverse
- Functional metadata requirements should be translated into an effective and efficient architecture
 - Three principles for modeling metadata for scientific data
- Metadata for scientific data (or other domains at large) should adopt an infrastructure service approach
- Much to be explored, experimented, and evaluated

Questions?

Questions?