Metadata and the US Geoscience Information Network

Stephen M. Richard Arizona Geological Survey and USGIN steve.richard@azgs.az.gov

Interoperable Search

- Users should be able to perform wide-ranging inter-disciplinary searches across the geoscience domain using the same search tools.
- Registration and search operations should use published, non-proprietary protocols and interchange formats such that anyone can implement new clients for search, or new servers for publication of metadata accessible to the network.
- Metadata should be viewed as a public resource. Catalogs should
  - o enable harvest of metadata for resources registered with them,
  - o be free to harvest from other catalogs in the system

Requirements:

Documented service protocols define operations and request syntax, e.g. OpenSearch, OGC CSW, OAIPMH

Standard metadata interchange format:

A. Content model (ISO19115, FGDC, Dublin Core...)

B. Information encoding- ISO19139, ATOM, csw:record

Goal: Reusable components

- Decouple components of system
  - Repository
  - Metadata registry
  - Clients for searching
  - Search service
  - Access service

The Catalog

What is the catalog

- A collection of metadata records describing resources available in the information network
- Records are
  - Structured – consistent information content, arranged to enable automated extraction/location of particular information
  - Shared – each record 'advertises' a resource; they are meant to be distributed widely and easily accessible
  - Owned by the resource steward

What's it for?

- Find, evaluate, get resources
  - Bounding box, keywords, author, title, date…
- Resource– identifiable item of interest
- Data vs metadata: Catalog is 'high level'
  - Have metadata record in catalog
    - Documents, physical samples, datasets, services, software, files, databases
  - Don't have metadata record in catalog
    - Individual database records, polygons in a GIS dataset, individual chemical analyses, individual temperature measurements
  - Domain specific description information in domain-specific services (not the catalog)

Metadata Content

- Title
- Description
- Originator(s)
- Date
- Geographic Extent
- Access instructions
- Distribution contact
- Metadata
  - Date
  - Contact
  - Specification
  - Identifier
- Citation
- Keywords
- Language
- Resource ID
- Vertical Extent
- Temporal extent
- Quality
- Lineage
- Access constraints
- URL

Structured vs. unstructured metadata:

Structured-- fields or elements contain semantically constrained information granules, using controlled vocabularies where feasible

Unstructured --
free text narrative describing resource, scraping information by analyzing text inside the resource

What is role of structured vs free text metadata

structured metadata--fielded search, separate content from presentation, machine parsing to extract specific information necessary for processing

free text-- available with little effort from text documents; more natural for people to write

Find → Select → Get → Use

Filter the universe of resources to identify a subset of interest

Look at description of resources to select candidates for use

Obtain the resource in a format meeting application requirements

Establish trust in information obtained and conclusions drawn from it; depends on nature of application

Discovery

Documentation

Find: Text indexing approaches to finding work well (Google, Lucene...).

Structured metadata is unnecessary for many search scenarios

Select: Initial resource evaluation and selection is normally done by a person reading some descriptive text

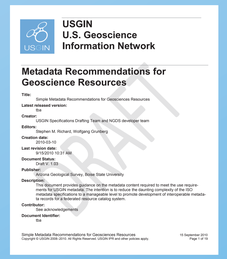
Unstructured text description of resources solves most requirements for evaluation

Get: For web-accessible documents, simple links do the job

But... to enable software assisted access to web applications, databases, or web services, machine-actionable (structured!) links are necessary (see discussion to Right)

Use: Structured complex provenance, data structure, and quality descriptions are useful for presenation, data filtering, and may be necessary as input for data processing.

http://repository.usgin.org/uri_gin/usgin/dlio/335



USGIN
U.S. Geoscience Information Network

Metadata Recommendations for Geoscience Resources

Title: Simple Metadata Recommendations for Geoscience Resources
Latest release version: 1.0
Creator: USGIN Specifications Drafting Team and NGDS developer team
Editors: Stephen M. Richard, Wolfgang Grunberg
Creation date: 2010-03-10
Last revision date: 8/10/2010 10:37 AM
Document Status: Draft C 0.83
Publisher: Arizona Geological Survey, Boise State University
Description: This document provided guidance on the metadata content required to meet the use requirements of the Geoscience Information Network. It provides a minimal set of requirements for metadata specifications to a sufficient level to promote development of interoperable metadata records for a federated resource catalog system.
Contributor: See acknowledgments
Document Identifier: file

Simple Metadata Recommendations for Geoscienceis Resources
Copyright © USGIN 2006-2010. All Rights Reserved. USGIN IPR and other policies apply.
15 September 2010
Page 1 of 10

http://lab.usgin.org/USGIN-ISO-metadata-v1-1

USGIN.ORG
USGIN
U.S. Geoscience Information Network

Use of ISO metadata specifications to describe geoscience information resources

Version 1.1

Title: Use of ISO metadata specifications to describe geoscience information resources
Latest released version: http://lab.usgin.org/profiles/usgin-iso-19139-profile
Creator: USGIN Standards and Protocols Drafting Team
Editors: Stephen M. Richard and Wolfgang Grunberg
Creation date: 9/18/2009
Last revision date: 11/18/2010 3:20 PM
Document Status: Draft 1.1
Publisher: Arizona Geological Survey
Description: This document describes recommended practices for using ISO19139 xml encoding of ISO 19115 and ISO 19119 metadata to describe a broad spectrum of geoscience resources. The recommendations are based on the Energistics Energy-Industry profile (EIP) specification and enable interoperability of catalog service clients with multiple servers conforming to this profile.
Contributor: See acknowledgements
Document Identifier: gin2010-009

usgin_iso_metadata_1.1.4
Copyright © USGIN 2006-2010. All Rights Reserved. USGIN IPR and other policies apply.
19 November 2010
Page 1 of 136

Service vs data set metadata

most searches are focused on finding information; how the information is accessed/distributed is generally a second order consisderation

Virtually all service protocols offer a self-description operation

metadata should provide links to service-specific self-description (WSDL, WADL, GetCapabilities...)

Metadata for RPC type services vs. Resource-oriented services

Machine actionable links:

Does a computer know what to do with your URL?
A ubiquitous problem for next generation web

property	scope notes
linkage	URI identifies the resource that is the target of the link. (syn: href, targetURI)
rel	Semantics of link. (might need multiple values...)
title	Free text to label link in GUI
type	File format and optionally the native software application environment (MIME).
function	Term tells client why they'd use this link. e.g. 'download', 'browsing', 'fileAccess', 'service'.
protocol	URI protocol e.g. http, ftp, dns, nfs, See IETF registry at http://www.rfc-editor.org/rfcxx00.html. typically encoded as URI prefix
serviceType	URI that identifies a service protocol e.g. CSW, WFS, WMS, OpenSearch, OpenDAP, etc.
outputScheme	Identifier for schema of message retrieved by link; URI for xml schema or JSON scheme; Clients look at this to pick the link that will get a representation they can use;

Other properties that may be useful (all optional)

altTitle	String that encodes title value in a different character set, and/or contain language information as per [RFC5987].
descriptionurl	Detailed text description of what the online resource is/does.
media	Indicates intended destination medium or media for style information e.g. 'screen', 'tty', 'print', 'braille', 'aural'... Default to 'screen', and it is anticipated that other values would be only rarely required.
length	Indicates an advisory length of the linked content in octets; it is a hint about the content length of the representation returned
hreflang	Describes the language of the resource pointed to by the link
behavior	A comma separated list of properties specifying behavior expected in client when link is actuated or loaded; e.g. nofollow, showreplace, showembed, actuateOnLoad, actuateOnRequest, noreferrer, prefetch

Relation between USGIN Profile and Energistics Energy-Industry profile (EIP).

USGIN is based on ISO19115 (2006)

EIP is based on ISO19115-1 (DIS 2011)

USGIN v2 for ISO19115-1 will harmonize with EIP

What is USGIN? :

-A collection of service definitions, interchange formats, and vocabularies

-Catalogs that lead users to resources

-A community of users: providers and consumers of data, services, applications

- Collaboration between state geological surveys and the U.S. Geological Survey

http://lab.usgin.org http://stategeothermaldata.org

This material is based upon work supported by the **Department of Energy** under Award Numbers DE-EE0002850 and DE-EE0001120 and by **National Science Foundation** grant EAR-0753154