Pragmatic metadata management for integration into multiple spatial data infrastructure systems and platforms

Karl Benedict * Soren Scott †

11-Dec-2013. AGU Fall Meeting, San Francisco, CA

Introduction

> prag · mat · ic *adjective* \prag-'ma-tik\ : dealing with the problems that exist in a specific situation in a reasonable and logical way instead of depending on ideas and theories
>
> — Merriam-Webster Dictionary

Abstract

While there has been a convergence towards a limited number of standards for representing knowledge (metadata) about geospatial (and other) data objects and collections, there exist a variety of community conventions around the specific use of those standards and within specific data discovery and access systems. This combination of limited (but multiple) standards and conventions creates a challenge for system developers that aspire to participate in multiple data infrastrucutres, each of which may use a different combination of standards and conventions. While Extensible Markup Language (XML) is a shared standard for encoding most metadata, traditional direct XML transformations (XSLT) from one standard to another often result in an imperfect transfer of information due to incomplete mapping from one standard's content model to another.

This paper presents the work at the University of New Mexico's Earth Data Analysis Center (EDAC) in which a unified data and metadata management system has been developed in support of the storage, discovery and access of heterogeneous data products. This sytem, the Geographic Storage, Transformation and Retrieval Engine (GSTORE) platform has adopted a polyglot database model in which a combination of relational and document-based databases are used to store both data and metadata, with some metadata stored in a custom XML schema designed as a superset of the requirements for multiple target metadata standards: ISO 19115-2/19139/19110/19119, FGCD CSDGM (both with and without remote sensing extensions) and Dublin Core. Metadata stored within this schema is complemented by additional service, format and publisher information that is dynamically "injected" into produced metadata documents when they are requested from the system. While mapping from the underlying common metadata schema is relatively straightforward, the generation of valid metadata within each target standard is necessary but not sufficient for integration into multiple data infrastructures, as has been demonstrated through EDAC's testing and deployment of metadata into multiple external systems: Data.Gov, the GEOSS Registry, the DataONE network, the DSpace based institutional repository at UNM and semantic mediation systems developed as part of the NASA ACCESS ESLeWEB project. Each of these systems requires valid metadata as a first step, but to make most effective use of the delivered metadata each also has a set of conventions that are specific to the system. This

*University of New Mexico, Earth Data Analysis Center, University Libraries, Department of Geography - kbene@unm.edu
†University of New Mexico, Earth Data Analysis Center

presentation will provide an overview of the underlying metadata management model, the processes and web services that have been developed to automatically generate metadata in a variety of standard formats and highlight some of the specific modifications made to the output metadata content to support the different conventions used by the multiple metadata integration endpoints.

Organizational Foundation

EDAC specializes in applied geographic information technologies across multiple research and applications domains including public health; disaster planning, management and mitigation; natural resources management; broadband network distribution analysis and mapping; data archival and clearinghouse services and research cyberinfrastructure. Historically EDAC has managed work in these domains as separate projects, often built upon vertically integrated data management, discovery and access systems as appropriate for a given project. Over the last 5 years EDAC has developed a unified data management, discovery and access platform (the Geographic Storage, Transformation and Retrieval Engine - GSTORE) that is under continuous development in support of broadening its utility across multiple projects and use cases.

Background - Diversity of data

[Figure 1 about here.]

While the project contexts within which EDAC works are commonly geospatial, the types of data, their structure and format, and subject are highly variable.

Background - Diversity of systems to integrate with

The requirements of different projects have led to a need to integrate data products, services **and their metadata** into multiple systems, including

- Project specific data portals such as the New Mexico EPSCoR data portal and the New Mexico Resource Geographic Information System (NM RGIS)
- Geospatial OneStop (subsequently replaced by *geo.data.gov* and then by *catalog.data.gov*
- The Global Earth Observation System of Systems (GEOSS) Components and Services Registry
- DataONE
- CUAHSI
- LoboVault - UNM's institutional repository

Background - Diversity of data, metadata and service formats and standards

Metadata	Vectors	Rasters	Files	Services
FGDC CSDGM	SHP	GeoTIFF	ZIP	WMS
FGDC CSDGM-RSE	KML	IMG	HTML	WFS
ISO 19115-2 / 19139	GML	SID	PDF	WCS
ISO 19119	GeoJSON	ECW	DOC/DOCX	

ISO 19110	JSON	DEM	GZ
	XLS	ASCII	XLS/XLSX
	CSV		PPT/PPTX

Table 1: Data, metadata and service standards currently supported by GSTORE

High-level Architecture - Overall View

[Figure 2 about here.]

High-level Architecture - Highlighted Elements

[Figure 3 about here.]

Process Overview

[Figure 4 about here.]

The overall GSTORE metadata processing workflow consists of the following high-level steps:

- Submission of a JSON document containing an embedded XML (ISO, FGDC or GSTORE) metadata record via the GSTORE API
- Processing of the submitted JSOM document into the PostgreSQL database
- Generation and delivery of metadata in a variety of formats in response to GSTORE web service requests

Process - Ingest

[Figure 5 about here.]

The ingest proces consists of parsing the submitted JSON package and adding its contents to various locations within the GSTORE database

- If the JSON document contains valid FGDC or ISO metadata it is transformed into GSTORE metadata and stored. Otherwise, the embedded XML document is stored as the *original* metadata record
- Other elements of the JSON document are written into multiple tables in the database.

Process - Export

[Figure 6 about here.]

Supported Output Formats

- FGDC CSDGM & CSDGM-RSE
- ISO
  - 19115-2 / 19139 (MI- and DS-)
  - 19110 (vector data feature catalog)
  - 19119 (OGC WxS)
- Dublin Core
- Schema.org within HTML renderings
- Elastic Search JSON documents for indexing/search
- EML (early development)
- Project Open Data (early development)

Integration into other Systems

The system is currently operational or in testing with:

- Catalog.data.gov
- GEOSS Components and Services Registry
- DataONE Tier 1 Member Node
- ElseWEB semantic mediation system

Internal development under way:

- CUAHSI WaterML
- DataONE Tier 4 Member Node
- Catalog.data.gov collection integration (ISO DS metadata records)
- W3C PROV services

High-level View of GSTORE Metadata XML

```
<metadata>
    <original>...</original>
    <identification dataset="51349b33-92eb-4ab8-9217-81c82b5c3afa">...</identification>
    <constraints>...</constraints>
    <spatial>...</spatial>
    <quality>...</quality>
    <lineage>...</lineage>
    <attributes>...</attributes>
    <metadata>...</metadata>
    <distribution>...</distribution>
    <contacts>...</contacts>
    <citations>...</citations>
    <sources>...</sources>
</metadata>
```

Sample Requests - *Digital Geologic Map of New Mexico - Formations*

GSTORE Metadata - XML http://gstore.unm.edu/apps/rgis/datasets/51349b33-92eb-4ab8-9217-81c82b5c3afa/metadata/GSTORE.xml

FGDC CSDGM 1998 - XML http://gstore.unm.edu/apps/rgis/datasets/51349b33-92eb-4ab8-9217-81c82b5c3afa/metadata/FGDC-STD-001-1998.xml

FGDC CSDGM 1998 - HTML with Schema.org additions http://gstore.unm.edu/apps/rgis/datasets/51349b33-92eb-4ab8-9217-81c82b5c3afa/metadata/FGDC-STD-001-1998.html

ISO 19115-2 (19139) - XML http://gstore.unm.edu/apps/rgis/datasets/51349b33-92eb-4ab8-9217-81c82b5c3afa/metadata/ISO-19115:2003.xml

ISO 19115-2 (19139) - HTML with Schema.org additions http://gstore.unm.edu/apps/rgis/datasets/51349b33-92eb-4ab8-9217-81c82b5c3afa/metadata/ISO-19115:2003.html

ISO 19110 (Feature Catalog) - XML http://gstore.unm.edu/apps/rgis/datasets/51349b33-92eb-4ab8-9217-81c82b5c3afa/metadata/ISO-19110.xml

ISO 19119 (Service metadata) - XML for OGC WFS service http://gstore.unm.edu/apps/rgis/datasets/51349b33-92eb-4ab8-9217-81c82b5c3afa/metadata/ISO-19119:WFS.xml

ISO 19119 (Service metadata) - XML for OGC WMS service http://gstore.unm.edu/apps/rgis/datasets/51349b33-92eb-4ab8-9217-81c82b5c3afa/metadata/ISO-19119:WMS.xml

Lessons Learned & Next Steps

- There is a significant difference between *valid* and *useful* metadata
- Metadata development is an iterative process
- Implementation of new output formats is a relatively fast and straightforward process
- Regenerating new collections of metadata with updated autogenerated content not as fast - over 285k data objects currently managed in GSTORE

Resources

- GSTORE XSLT/XSD Github Collection: https://github.com/edac/metadata/tree/master/gstore
- GSTORE API Reference: http://gstore.unm.edu/

Key Acronyms

ISO International Standards Organization

FGDC CSDGM (RSE) Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (Remote Sensing Extension)

JSON Javascript Object Notation

XMK Extensible Markup Language

API Application Programming Interface

OGC Open Geospatial Consortium

WMS, WFS, WCS OGC Web Map/Feature/Coverage Service

Acknowledgements

This work has been funded in part through the support of:

- NASA: ACCESS-11-0018 NNX12AF49A
- NSF
  - Energize New Mexico - #1301346
  - New Mexico EPSCoR RII3: Climate Change Impacts on New Mexico's Mountain Sources of Water - #0814449
  - Collaborative Research: Cyberinfrastructure Development in the Western Consortium of Idaho, Nevada, and New Mexico - #0918635
  - Collaborative Research: Western Consortium for Watershed Analysis, Visualization, and Exploration (WC-WAVE) - #1329470
  - Collaborative Research: CI-Team Diff: The Virtual Learning Commons: STEM Research Communities Learning about Data Management, Geospatial Informatics, and Scientific Visualization - #OCI-1135530
- New Mexico RGIS Program - New Mexico State Legislature

Questions?

List of Figures

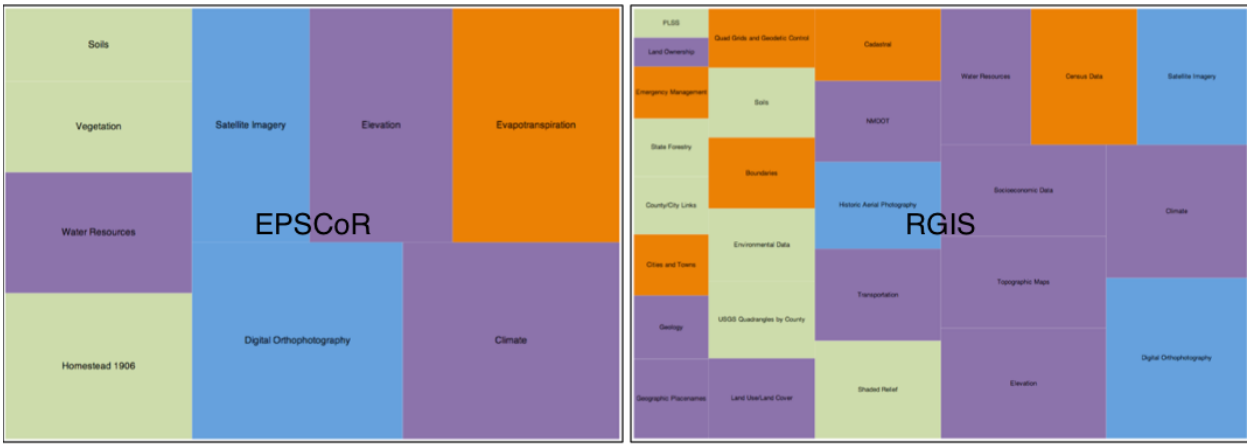

Figure 1: Illustration of the diversity of data classes in two project collections

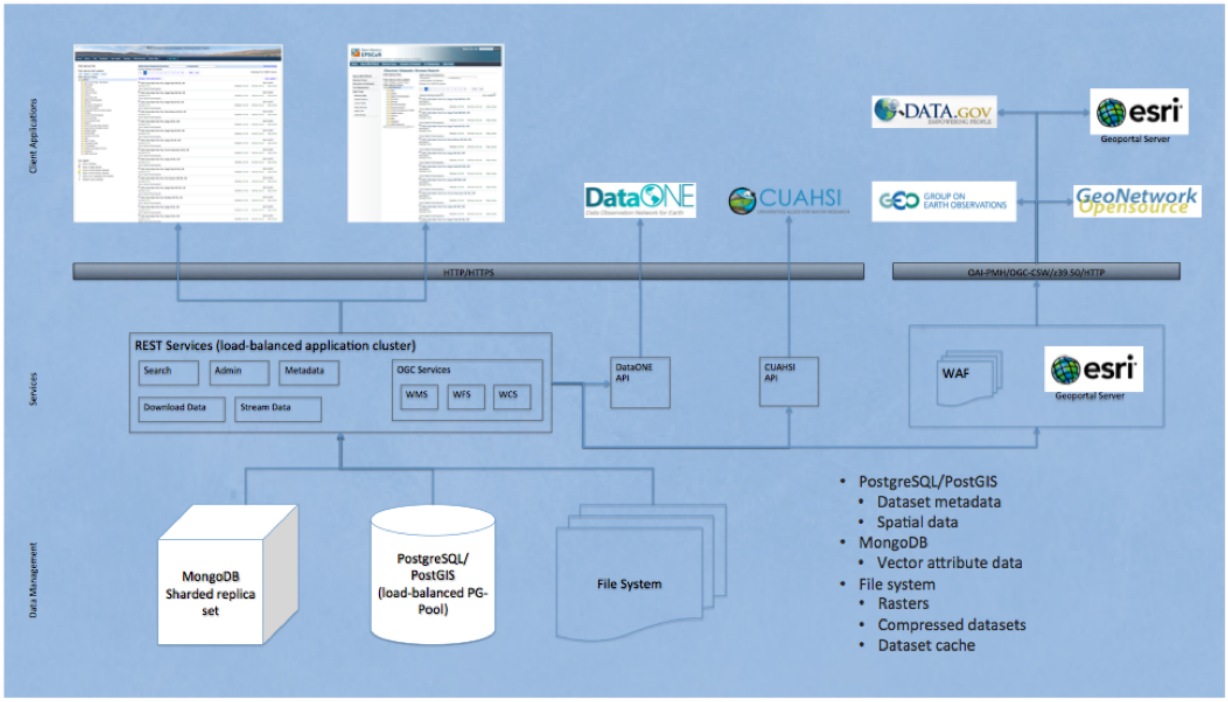

Figure 2: Overall Architectural Diagram

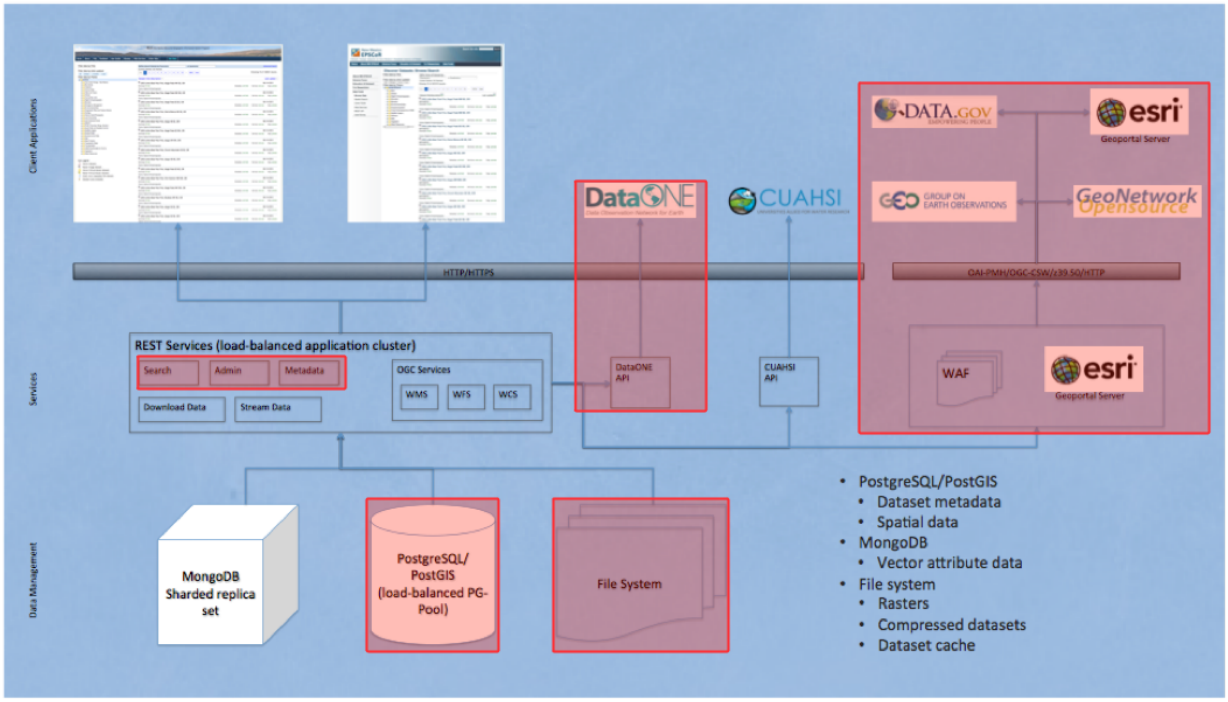

Figure 3: Highlighted Architectural Diagram

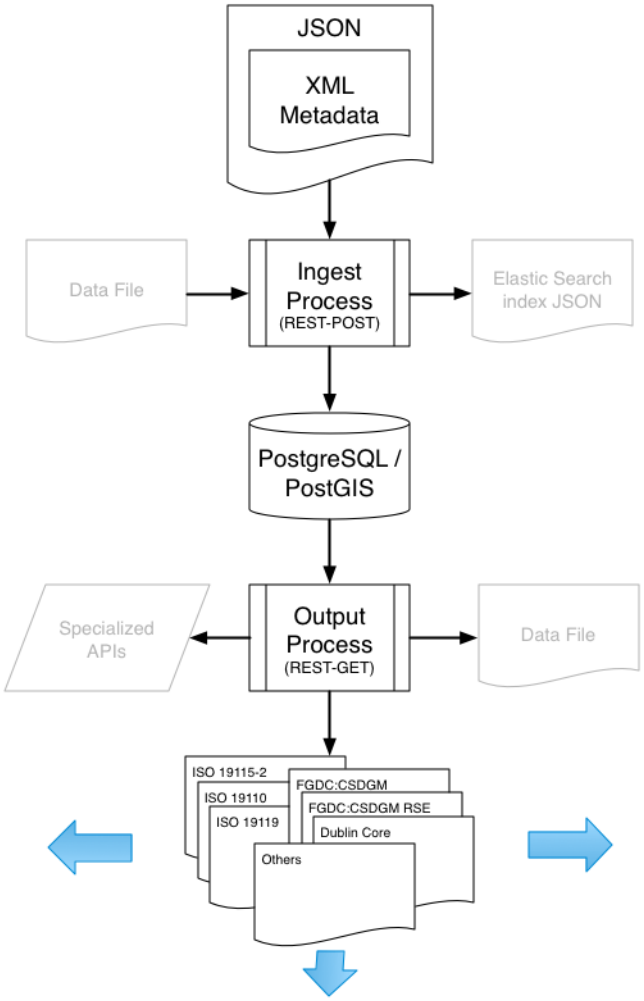

Figure 4: Overview of the Combined Ingest and Metadata Generation Process

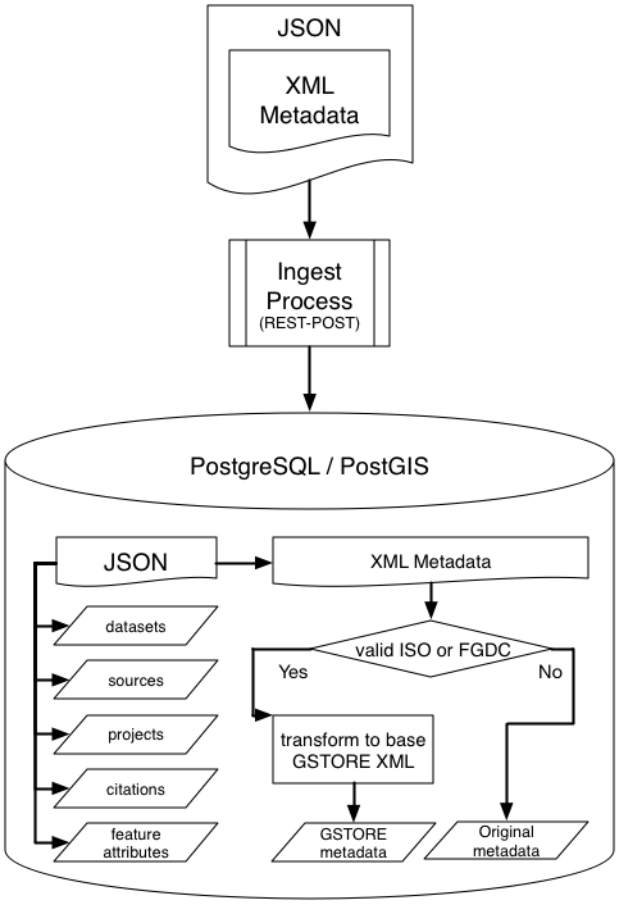

Figure 5: Detailed View of the Metadata/data Ingest Process

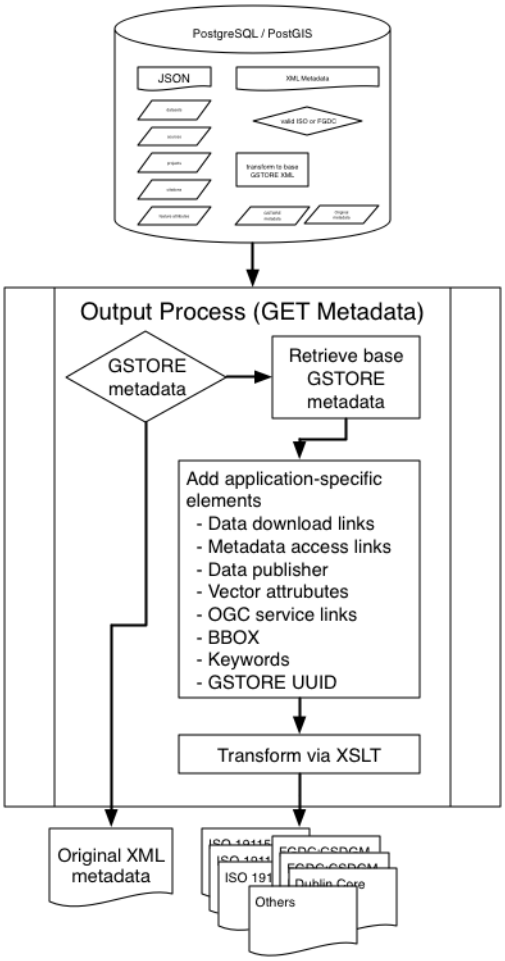

Figure 6: Overview of the Metadata Generation Process