

# Recent Documentation Efforts for Data Preservation at the NOAA NCDC

**Philip Jones**

**STG, Inc.**

**National Climatic Data Center, Archive Branch**

2012 NOAA EDM Conference

College Park, MD

16 May 2012

# OAIS RM

**Long Term Preservation:** The act of maintaining information, in a correct and Independently Understandable form, over the Long Term.

**Preservation Description Information (PDI):** The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information.

**Independently Understandable:** A characteristic of information that has sufficient documentation to allow the information to be understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.

## **Data Center Perspective:**

- A Data Center must ensure that **supporting information is supplied** with the data before taking responsibility of stewarding a dataset (Appraisal and Submission Agreement phases)
- **Standards are necessary** for sustaining independently understandable information in data management

# NCDC Base Reference for Collection Metadata

- Defines the required set of minimum metadata needed to **support the discovery and access** of NCDC data holdings (i.e., collections)
- Base Reference Model contains **functional and non-functional requirements** for collection-level metadata and its use for discovery
- Meets NOAA Documentation Directive
- **Facilitates understanding** of the metadata and standards for authors through **Best Practice guidelines**
- Requirements are a means of **assuring consistency and interoperability** of the metadata information and management across systems

# Data Granularity

## Types of information (OAIS)

Granularity ↓

	Reference Identification	Representation Information	Context	Provenance
Collection/Series	Dataset Title Data Center ID Journal Reference DOI	Format Specification Auxiliary data Navigation Calibration	Science Paper Validation Study Related Datasets Mission	Originator/PI Contributing Collections Platforms Instruments ATBD
File/Object	File name UUID  (checksum for integrity)	File Format File Structure Companion File	File content statistics Parameter thresholds Percentages	Input files Processing time Modified date Storage and handling
Parameters	Variable name Standard name	Data Type Units Scale Offset Time reference	Quality Flag Quality Indicator	Input source Update Flag

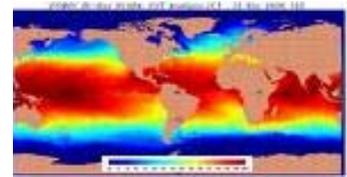
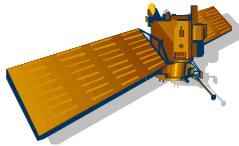
**Self-describing file formats** like netCDF are effective at documenting data at the file level and are **preferred for long-term archives**

# Discovery

- Data discovery is essentially the act of **finding a collection** of data
- It is a **multi-step process** in which a user conducts a broad search, browses returned search results, and locates collections of interest through a discovery service
- Discovery **does not cover** finding specific data within a collection (a subsequent process requiring a more granule level of metadata pertaining to that collection or related type of data)
- Discovery search results provide sufficient information for a user to resolve links for more information, or obtain the resource (through a search, download, web service, or other means)
- Standard metadata documenting data collections supports the discovery search and search results display.

# Standard Metadata

- Standard that provides a structure for describing and exchanging geographic data with the broad user communities
- Defines metadata elements, provides a schema and establishes a common set of metadata terminology, definitions, and extension procedures (normally represented in XML)
- Facilitates data discovery, access, high-level evaluation, and utilization
- Related Metadata Standards:
  1. **NASA DIF (basic)**: used by Paleo-climate group and GCMD
  2. **FGDC CSDGM ERSM (more complete)**: legacy standard used by NCDC and supported by portals
  3. **ISO 19115 & 191\*\* (most complete)**: planned NCDC baseline and compatible with FGDC



<b>ID:</b>	gov.noaa.ncdc:C00366	gov.noaa.ncdc:AVHRR	gov.noaa.ncdc:C00846	gov.noaa.ncdc:C00785
<b>Title:</b>	CRN Raw Observations	AVHRR Level 1B	Daily OISST v2	Keyed East India Co. Logs
<b>Originator:</b>	CRN	OSPO	NCDC	NCDC
<b>Spatial Extent:</b>	-172.0, -66.0, 72.0, 18.0	-180.0, 180.0, 90.0, -90.0	-180.0, 180.0, 90.0, -90.0	-180.0, 180.0, 70.0, -50.0
<b>Temporal Extent:</b>	2001-10-01 to Present	1978-11-05 to Present	2001-11-01 to Present	1789 to 1834
<b>Theme Keywords:</b>	AIR TEMPERATURE, ...	VISIBLE RADIANCE, ...	SEA SURFACE TEMP, ...	WINDS, SEA STATE, ...
<b>Platforms:</b>	CRN Stations	NOAA-15, NOAA-16, ...	NOAA-15, NOAA-16, ...	EIC Ship Voyages
<b>Instrument:</b>	RAIN GAUGES, ...	AVHRR	AVHRR	WIND VANES, ...
<b>Processing Level:</b>	Level 0	NOAA Level 1B	NOAA Level 3	NOAA Level 1
<b>File Format:</b>	Binary	Binary	netCDF	ASCII
<b>Links:</b>	Documentation	Data, Documentation	Data, Documentation, Code	Data, Documentation

**Standard Metadata** documents diverse data collections using common attributes in a standard format

# Why Define a Minimum Set?

- ISO 19115 (and FGDC CSDGM ERSM) standards define an **extensive set of metadata** elements; however **only a subset of elements are typically used**

- **Minimum Extreme:** Only eight (8) fields from ISO 19115 are required for compliance – *for example:*

1. Metadata Language: **eng**
2. Hierarchy Level: **nonGeographicDataset**
3. Metadata Contact Organization: **NCDC**
4. Metadata Contact Role: **role=originator**
5. Metadata Date Stamp: **20120515**
6. Title: **Data Preservation at the NOAA NCDC**
7. Publication Date: **20120516**
8. Maintenance and Update Frequency: **notPlanned**

- Conversely, there are **hundreds of ISO 19115 elements** in several packages/classes that can be (re-)used in many different ways

- **Provides Guidance** - because the Data SMEs who are not metadata experts need to write metadata too!

# Assumptions & Approach (1 of 1)

- Look at use cases: metadata should benefit the needs of known use cases (e.g., discovery portals, HTML page views)
- The Information is the target for support
- Baseline should support information from multiple metadata standards (where we are now)
- Minimum required elements should be common across all collections of various data – can feasibly be documented for any and all data collections
- Register and maintain a metadata record per collection / major version
- Use standard vocabularies and code list values (e.g., GCMD Keywords and ISO code lists)

# Assumptions & Approach (1 of 2)

- Creation of some metadata content can't be automated – requires human input - need technical input from knowledgeable and responsible parties
- Metadata must be simplified for Data SMEs who have the knowledge and understanding of the data
- Serve as a comprehensive resource for users of the dataset – though only include the things that need to be associated with the dataset, without replicating the contents of those things
- Include links to resolve more information – browse to related documentation and metadata for related datasets (HTTP/FTP accessible links)
- Ongoing Dependency: will need to adapt with changes in metadata standards, tools, and discovery services (uses/needs)

# Minimum Metadata Summary

**Total minimum metadata elements: 57**

*not counting the use of repeating elements*

**Total contains:**

- default value available: **28**
- valid domain code lists selections: **11**
- date fields: **4**
- other unique input required: **14**

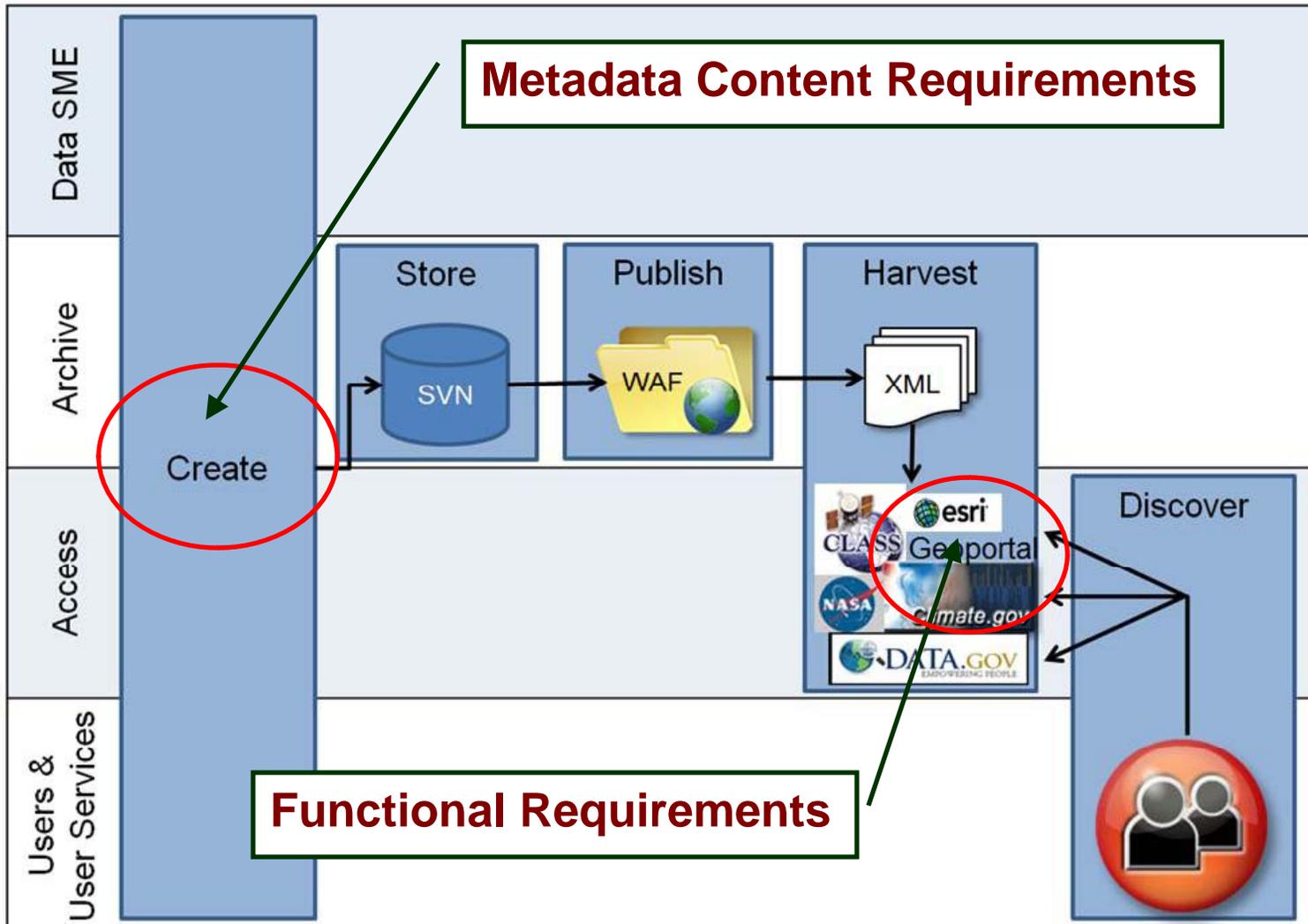
# Minimum Metadata Elements

- NCDC Required Minimum Set includes:
  - Metadata ID
  - Metadata Contact
  - Dataset Citation (Title, Originator, Publish date, etc.)
  - Abstract
  - Purpose
  - Data POC
  - Browse Graphic
  - Theme Keywords and Thesaurus
  - Constraints
  - Temporal and Spatial Extents
  - Distributor
  - Distribution Format and Link

# Beyond the Discovery Minimum

- Additional requirements for Highly Recommended Metadata Content
- Metadata authors are **encouraged to provide more applicable information** through additional metadata, including:
  - Dataset Version
  - Additional Keywords (Place, Stratum, Instrument, Platform, Project, Resolution)
  - Processing Level
  - Browse Graphic
  - Cross Referenced data and resources
  - Attribute Accuracy
  - Consistency
  - Completeness
  - Processing Description
  - Processing AlgorithmCitation
  - Input Source Citation
  - Spatial Reference System
  - Format Version
  - File Decompression Technique
  - Aggregation Information
  - Related Data/Documentation
  - Lineage Description
  - Acquisition Information

# Metadata Workflow



- Multiple groups contribute to complete metadata creation
- Requirements needed for both metadata content and its use

# Functional Requirements

## Functional Requirements for Metadata Publishing and Data Discovery

- Common, authoritative metadata repository and identifiers
- Harvest frequency
- Metadata Standards Supported
- Xpaths of elements to index for searching
- Dynamic hyperlink labels by service type (from 2 to 12)

#	applicationProfile value	Search Results Display Label
1	Web Browser	Information
2	Download	Download
3	REST	Web Service
4	SOAP	Web Service
5	WMS	WMS
6	WFS	WFS
7	WCS	WCS
8	KML	KML
9	Map Search Application	Map Search
10	Data Search Application	Data Search
11	THREDDS	THREDDS
12	Video	Video

# Extending Requirements to Metadata Tools

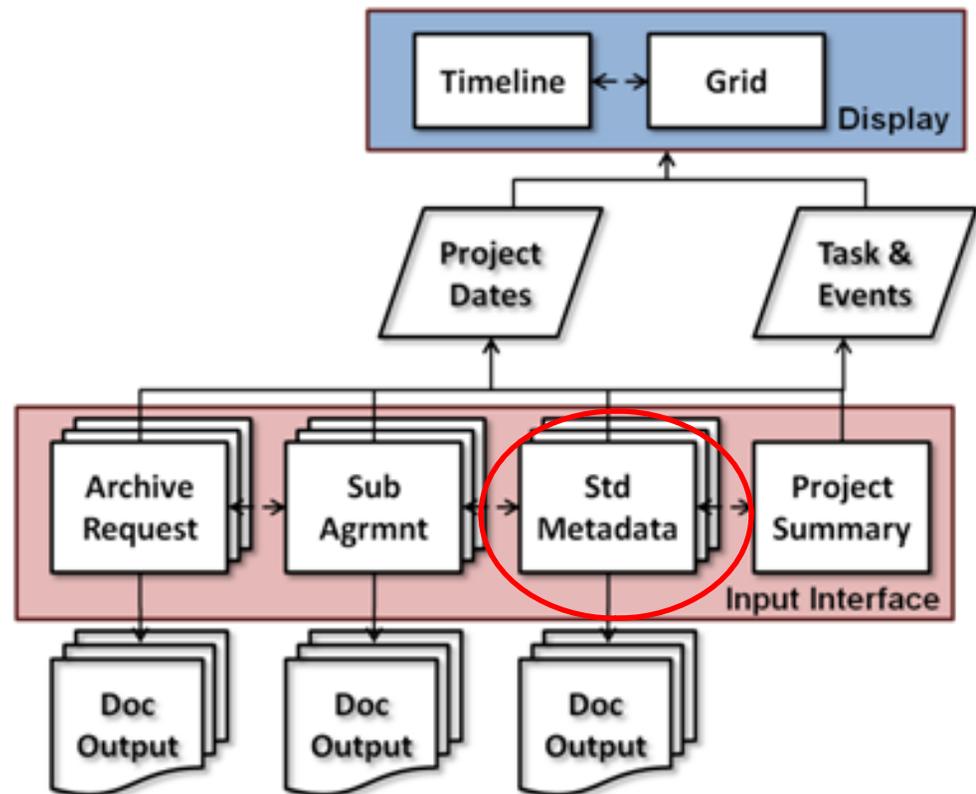
•The Advanced Tracking and Resource tool for Archive Collections (ATRAC) provides a common interface for users to enter and display information on archiving projects at the NOAA National Data Centers

## ATRAC Metadata Form (June 2012) –

Collects information through user-friendly questions that can be understood without detailed knowledge of standards and formats

### Features:

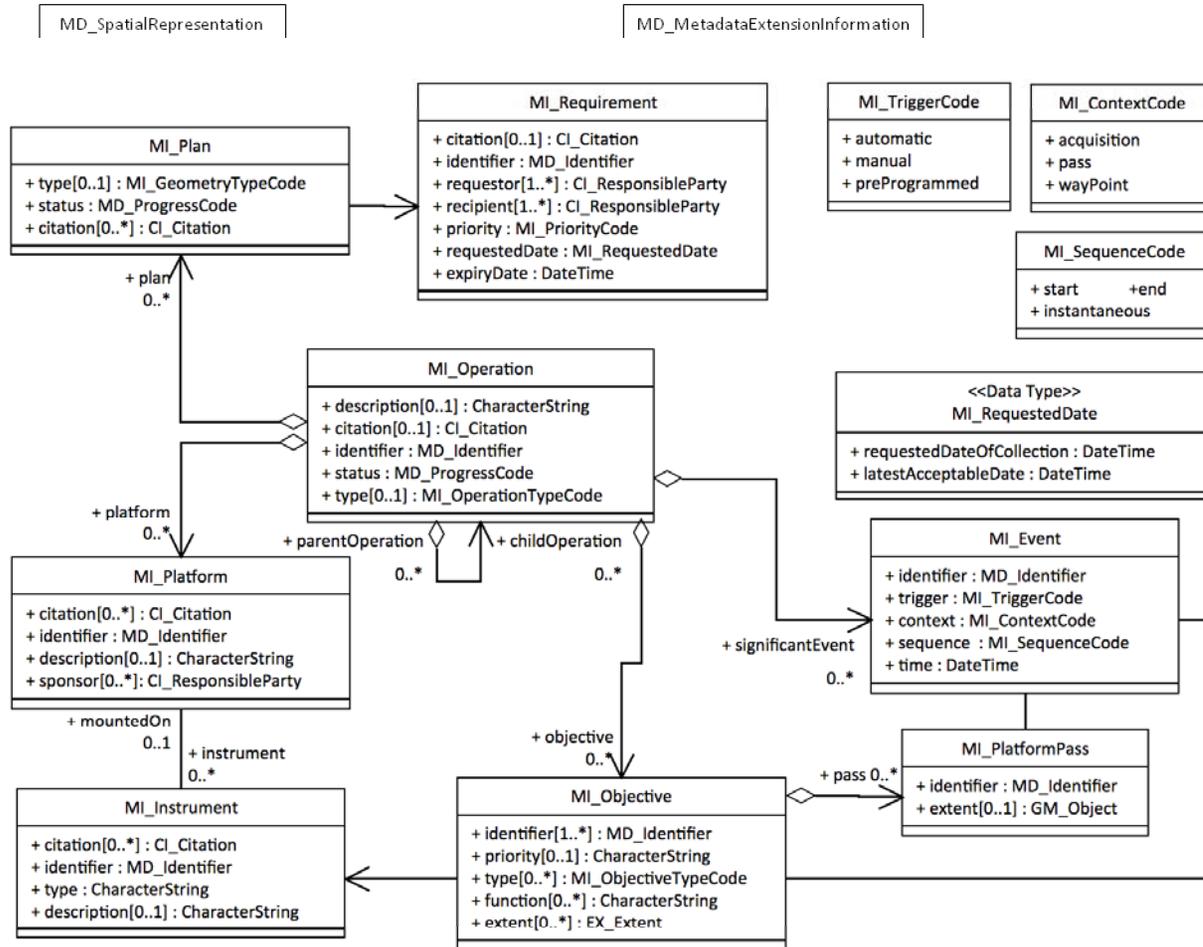
- Built-in validation for standards compliance and logical accuracy
- Web accessible with shared access
- No software license required
- Limited revision history
- Import/export information from/to multiple metadata standards
- Knowledge of metadata standards and XML not required



# Technical Knowledge

```

</gmd:EX_GeographicBound
</gmd:geographicElement>
- <gmd:geographicElement>
- <gmd:EX_GeographicDescrip
- <gmd:geographicIdentifier
- <gmd:MD_Identifier>
- <gmd:authority>
- <gmd:CI_Citation>
- <gmd:title>
  <gco:Character!
  </gmd:title>
- <gmd:date>
- <gmd:CI_Date>
- <gmd:date>
  <gco:Date>
  </gmd:date>
- <gmd:dateTyp
  <gmd:CI_D
  </gmd:dateTyp
  </gmd:CI_Date>
  </gmd:date>
  </gmd:CI_Citation>
  </gmd:authority>
- <gmd:code>
  <gco:CharacterStris
  </gmd:code>
  </gmd:MD_Identifier>
  </gmd:geographicIdentifier:
  </gmd:EX_GeographicDescrij
  </gmd:geographicElement>
- <gmd:temporalElement>
- <gmd:EX_TemporalExtent id=
- <gmd:extent>
- <gml:TimePeriod gml:id=
- <gml:begin>
- <gml:TimeInstant gml:id="
  <gml:timePosition>1994.
  </gml:TimeInstant>
  </gml:begin>
- <gml:end>
- <gml:TimeInstant gml:id="
  <gml:timePosition indet
  </gml:TimeInstant>
  </gml:end>
  </gml:TimePeriod>
  </gmd:extent>
  </gmd:EX_TemporalExtent>
  </gmd:temporalElement>
  
```



# Form Input

## ***Also:***

- **Auto-population of Organization names and keywords**
- **Auto-population of contact information**
- **Date tool**
- **ISO code lists**

**Select**

17. Ther

\*

Top

Ter

Var

+ Ac

18. Theme Keywords usi

+ Add Keyword

19. Place Keywords usin

+ Add Keyword

36. Supplemental information to complete the dataset description.

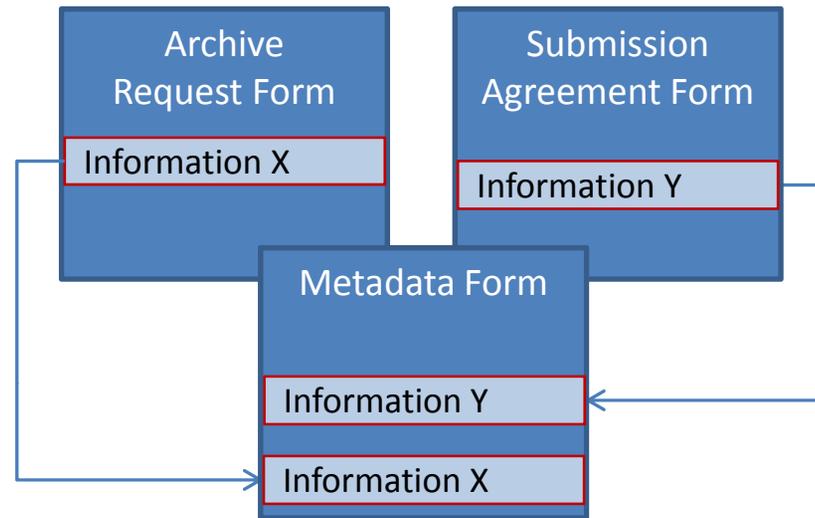
Save

Submit

# Metadata Import

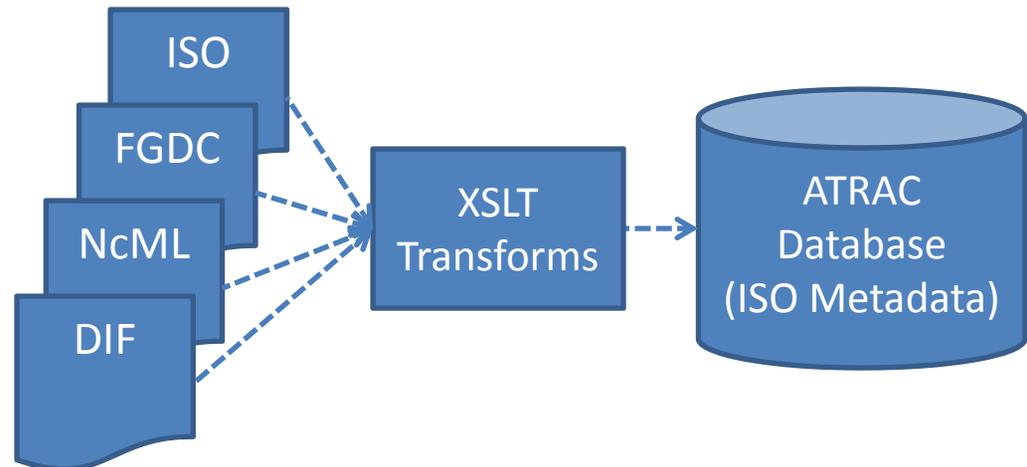
- **Metadata from Related ATRAC Forms (June 2012)**

- Imports current form information



- **Metadata from External XML Files (FY 2013)**

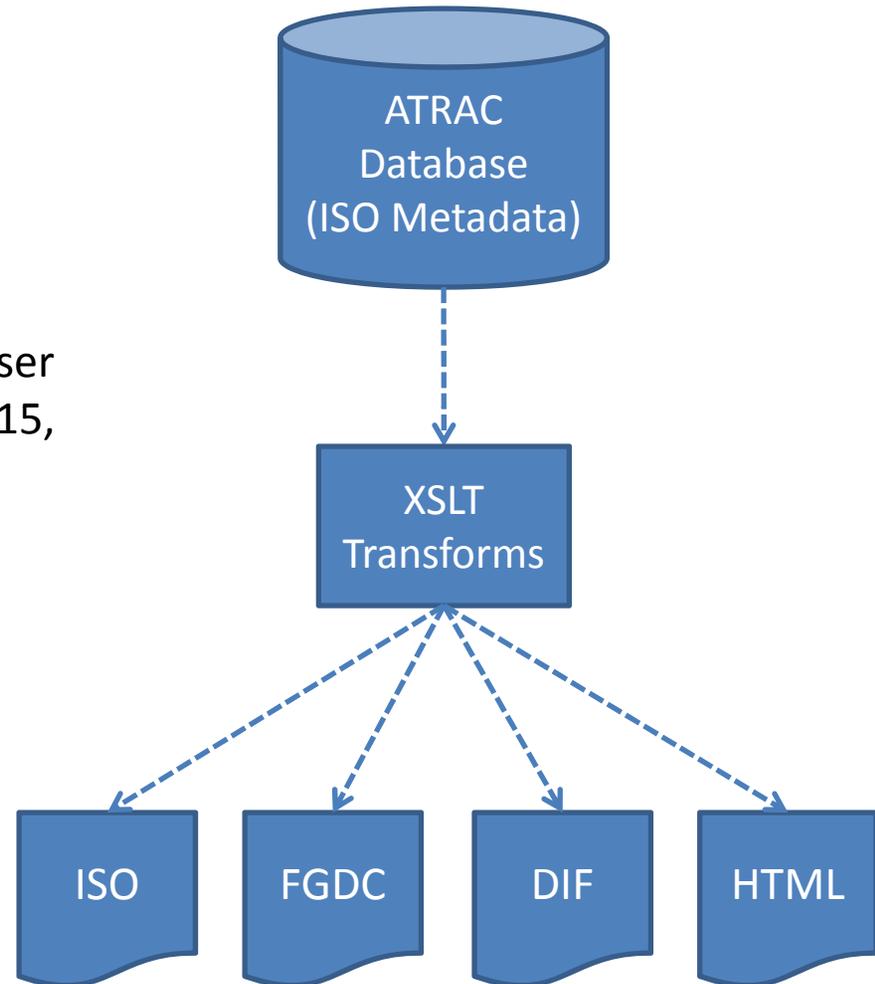
- Supports ISO 19115, FGDC CSDGM, NcML, and NASA DIF
- Extensible Stylesheet Language Transformations (XSLT translate content to ISO 19115)
- Stored in ATRAC database as a new Metadata form revision



# Metadata Export

## XML/HTML Files (June 2012)

- An XSLT converts the metadata in ISO 19115 to the requested export format
- Translated metadata is served to the user as an XML document in either ISO 19115, FGDC CSDGM, or NASA DIF format.
- Option to view metadata in a human readable HTML format



# Estimated Schedule

- **May-July 2012:**
  - Approval of NCDC Baseline Metadata Requirements
  - ATRAC v2.4 with Metadata form (beta)
  - Installation of Geoportal v1.2 with (applicationProfile)
  - Updates to existing metadata for compliance
  - Establish CM in workflow process
- **2012-2013:**
  - Identify and collect Data SME input to enhance metadata
  - ATRAC v2.5 (revised forms and import capability)
  - Improve XSLTs (XML and HTML)

# Acknowledgements

## **Metadata Baseline:**

Jeff Arnfield, Rich Baldwin, Jim Biard, Danny Brinegar, Heather Brown, Jeff Budai, Eric Freeman, Axel Graumann, Shaida Johnston, John Keck, Bryant Korzenierski, Christina Lief, Ge Peng, Ken Roberts, Valerie Toner, Anju Shah, Edward Gille and Bruce Bauer

## **Geoportal Configurations:**

Rich Baldwin, Charlie Roberts

## **ATRAC development:**

Nancy Ritchey (project lead)

Ken Roberts (developer)

Dan Kowal (SA & AR form user feedback)

# Questions?

- Email: [philip.jones@noaa.gov](mailto:philip.jones@noaa.gov)