

Metadata and the Tracking Network: An Update



Environmental Public Health Tracking Workshop

Tampa Bay, Florida

April 19, 2006

A Presentation by
The Metadata Registry Team



What is Metadata?

Metadata are “Data About Data”. They help a person to locate and understand data by describing the content, quality, condition, and other characteristics of the data.



Metadata Reveals...

- Who collected data.
- The method(s) used to collect data.
- When data was collected.
- Data gaps.



Metadata Reveals...

- How data was processed.
- When data was last updated.
- Limitations on data access.
- Data distribution information.



Frequently Asked Questions About Metadata...

- “ Creating metadata is going to tie up staff time. Is it really worth the effort? ”
- “Why do I want to take the time to create metadata? I am busy enough already.”
- “Why should our agency take the time to create metadata? We already know what we have.”



Why is Metadata Important to the Tracking Network?

- CDC said you must create metadata to obtain funding
- Allows users to discover the existence of data
- Helps users to understand data
- Protects investment in data
- Limits liability
- Can reduce staff workload (once created)

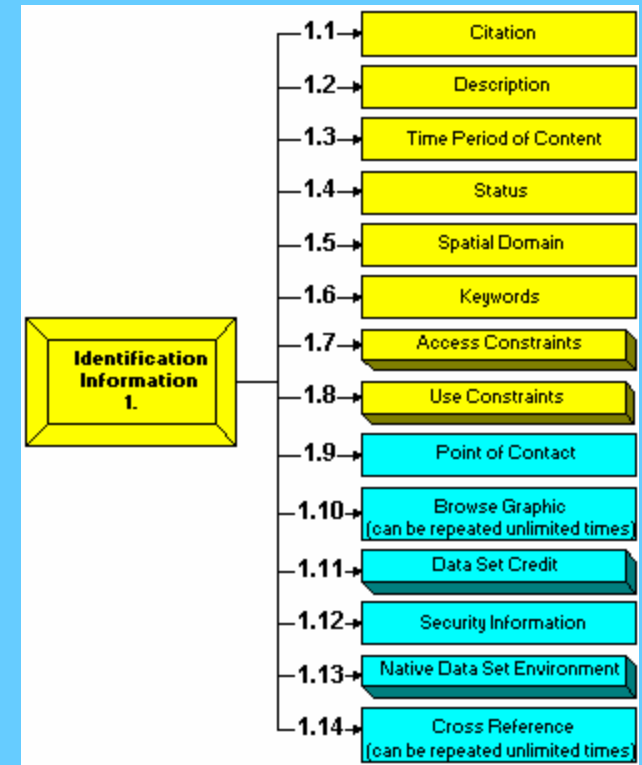
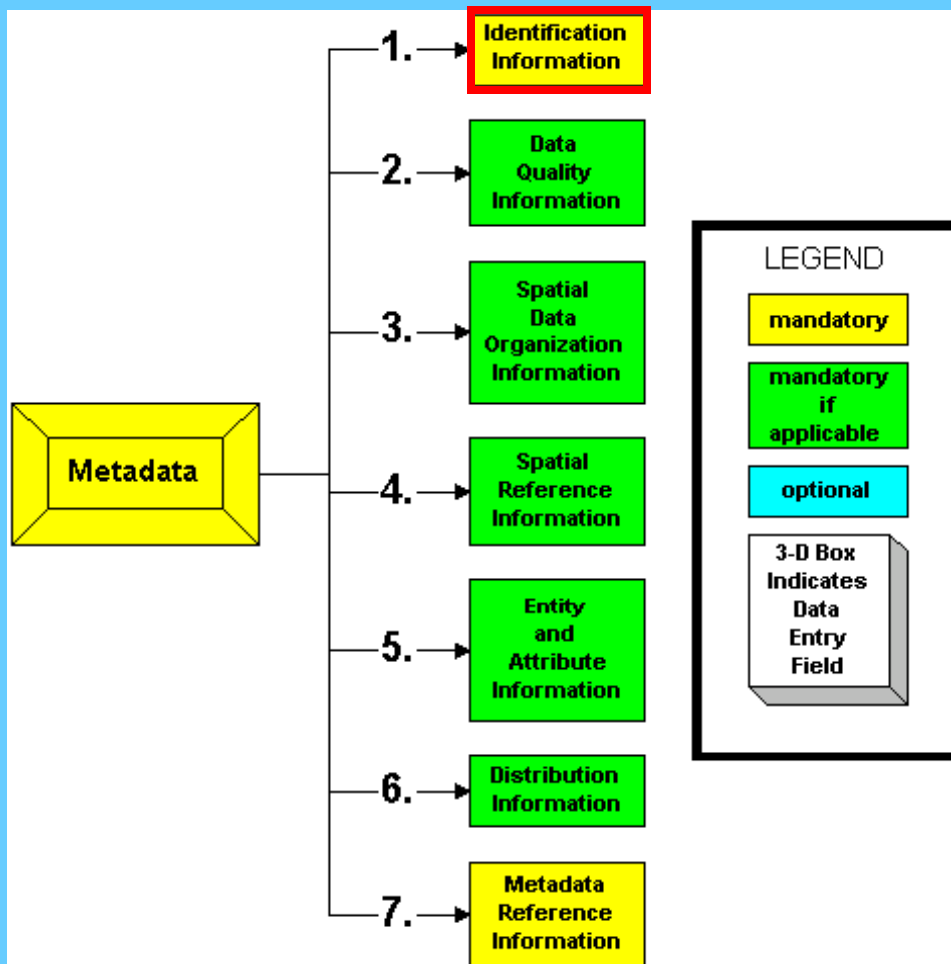


Metadata Activities and Products

Process/Activities	Products
Identified core elements from the FGDC Metadata standard	Tracking Network Metadata Template V1.0
Tested Metadata Template using freeware metadata creation tool – Tkme	Template passed the Test
Conducted requirements gathering session to identify desirable and mandatory attributes for a metadata creation tool	Tracking Network Metadata Tool Vision Document V1.0
Consolidated template with US EPA	NYC Metadata Toolset
Identified requirements for Tracking Network Metadata Registry Infrastructure	Tracking Network Metadata Registry Use Cases

FGDC Content Standard

Where we Started



Approximately 200 elements that could be filled out in the complete template



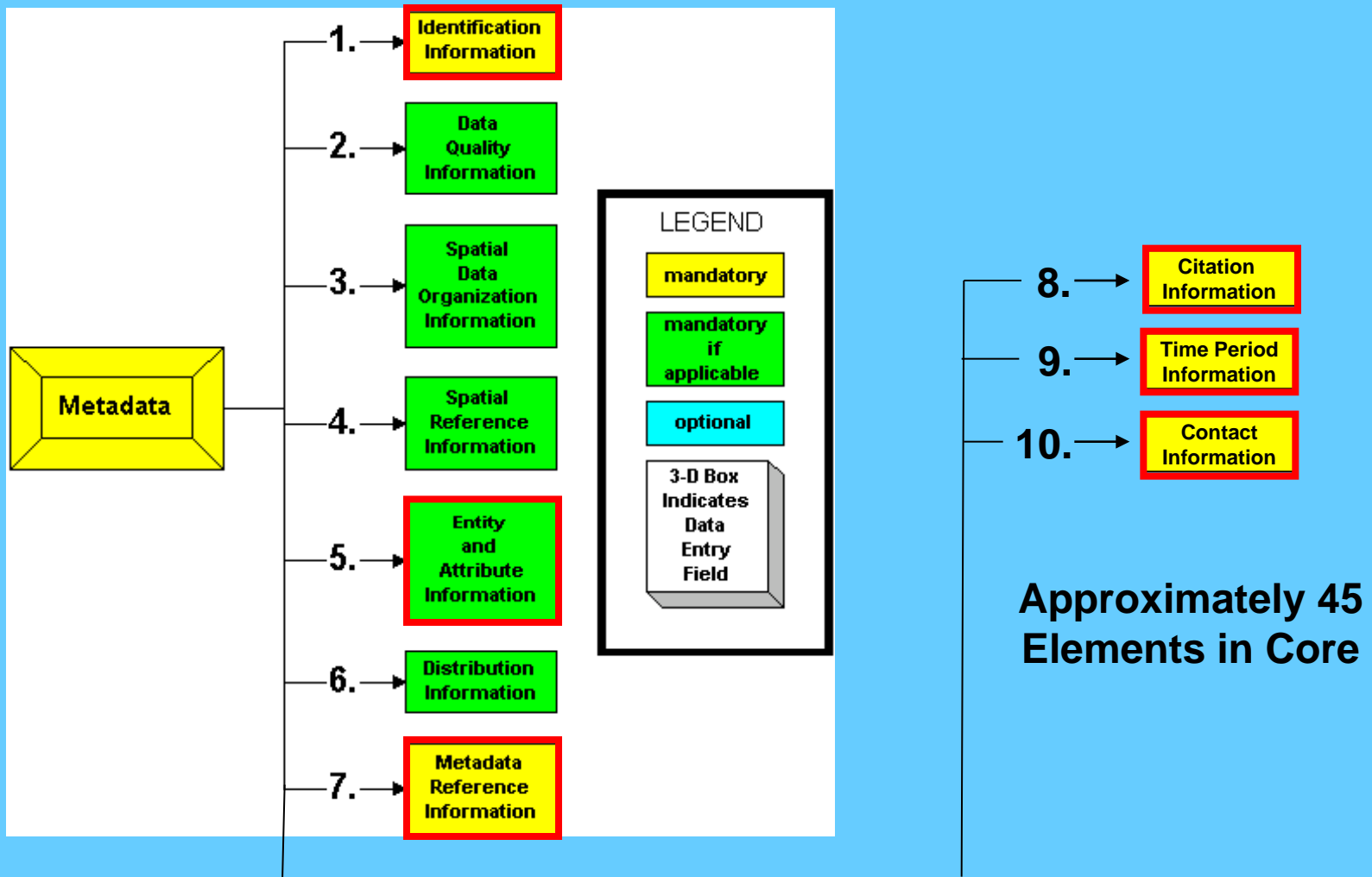
FGDC Content Standard

Making it Work

- Looked at the full FDGC Content Standard
- Investigated metadata creation tools
- Tested the initial template using the Tkme freeware
- Used feedback from test participants to create a “Core Elements” Metadata Template
- Finalized the Metadata Template with Metadata Subgroup

FGDC Content Standard

The EPHT Core Metadata Template



FGDC Content Standard

The EPHT Core Metadata Template

<i>Element</i>	<i>Field</i>	<i>Definition</i>
Section 1: IDENTIFICATION		
Citation	Citation (1.1)	Information to be used to reference the data set.
Originator	Originator (8.1)	The name of an organization or individual that developed the data set. If the name of editors or compilers are provided, the name must be followed by "(ed.)" or "(comp.)" respectively.
Publication Date	Publication_Date (8.2)	The date when the data set is published or otherwise made available for release.
Title	Title (8.4)	The name by which the data set is known.
URL	On-line_Linkage (URL) (8.10)	The name of an online computer resource that contains the data set. Entries should follow the Uniform Resource Locator convention of the Internet.
Description	Description (1.2)	A characterization of the data set, including its intended use and limitations.
Abstract	Abstract (1.2.1)	A brief narrative summary of the data set.
Purpose	Purpose (1.2.2)	A summary of the intentions with which the data
Supplemental Info	Supplemental_Info (1.2.3)	Other descriptive information about the data set.
Time Period of Content	Time_period_of_content (1.3)	Time period(s) for which the data set corresponds to the currentness reference. (Can be a single date, multiple dates, or a range of dates.).
Calendar Date	Calendar_Date (9.1.1)	The year (optionally month or month and day).
Currentness	Currentness_Reference (1.3.1)	The basis on which the time period of content information is determined.
Status	Status (1.4)	The state of and maintenance information for the data set.
Progress	Progress (1.4.1)	The state of a data set ("complete, in work, planned").
Maintenance and Update Frequency	Maintenance_and_Update_Frequency (1.4.2)	The frequency that changes are made to the data set after the initial data set is completed.
Spatial Domain	Spatial_Domain (1.5)	The geographic area covered by the data set.
West Bounding Coordinate	West_Bounding_Coordinate (1.5.1.1)	Western-most coordinate of the limit of coverage expressed in longitude.
East Bounding Coordinate	East_Bounding_Coordinate (1.5.1.2)	Eastern-most coordinate of the limit of coverage expressed in longitude.
North Bounding Coordinate	North_Bounding_Coordinate (1.5.1.3)	Northern-most coordinate of the limit of coverage expressed in latitude.
South Bounding Coordinate	South_Bounding_Coordinate (1.5.1.4)	Southern-most coordinate of the limit of coverage expressed in latitude.

FGDC Content Standard

The EPHT Core Metadata Template

<i>Element</i>	<i>Field</i>	<i>Definition</i>
Keywords	Keywords (1.6)	Words or phrases summarizing an aspect of the data set.
Theme Keyword Thesaurus	Theme_Keyword_Thesaurus (1.6.1.1)	Reference to a formally registered thesaurus or a similar authoritative source of theme keywords.
Theme Keyword	Theme_Keyword (1.6.1.2)	Common-use word or phrase used to describe the subject of the data set.
Place Keyword Thesaurus	Place_Keyword_Thesaurus (1.6.2.1)	Reference to a formally registered thesaurus or a similar authoritative source of place keywords.
Place Keyword	Place_Keyword (1.6.2.2)	The geographic name of a location covered by a data set. (Includes city, county, state, state acronym, regional descriptions and references).
Access Constraints	Access_Constraints (1.7)	Restrictions and legal prerequisites for accessing the data set. These include any access constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on obtaining the data set.
Use Constraints	Use_Constraints (1.8)	Restrictions and legal prerequisites for using the data set after access is granted. These include any use constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on using the data set.
Point of Contact	Point_of_Contact (1.9)	Contact information for an organization that is knowledgeable about the data set. Format found in Section 10.
Native Data Set Environment	Native_Data_Set_Environment (1.13)	A description of the data set, including the name of the software, computer operating system, file name, and data set size.
Section 5: ENTITY AND ATTRIBUTES		
Overview Description	Overview_Description (5.2)	Description of the entities, attributes, attribute values, and related characteristics encoded.
Entity and Attribute Overview	Entity_and_Attribute_Overview (5.2.1)	Detailed summary of the information contained in a data set.
Entity and Attribute Detail Citation	Entity_and_Attribute_Detail_Citation (5.2.2)	Reference used to the complete description of the entity types, attributes, and attribute values for the data set.

FGDC Content Standard

The EPHT Core Metadata Template

<i>Element</i>	<i>Field</i>	<i>Definition</i>
Section 7: METADATA REFERENCE		
Metadata Date	Metadata_Date (7.1)	The date that the metadata were created or last updated.
Metadata Contact	Metadata_Contact (7.4)	The party responsible for the metadata information. Format found in Section 10.
Metadata Standard Name	Metadata_Standard_Name (7.5)	The name of the metadata standard used to document the data set.
Metadata Access Constraints	Metadata_Standard_Version (7.6)	Restrictions and legal prerequisites for accessing the metadata. These include any access constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on obtaining the metadata.
Metadata Use Constraints	Metadata_Access_Constraints (7.8)	Restrictions and legal prerequisites for using the metadata after access is granted. These include any metadata use constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on using the metadata.
Section 10: CONTACT INFORMATION		
Contact Information	Contact_Information (10.0)	This section provides a means of identifying individuals and organizations, and is used by other sections of the metadata standard. This section is never used alone.
Contact Organization	Contact_Organization (10.1.2)	The name of the organization.
Contact Address	Contact_Address (10.4)	The address for the organization.
Address Type	Address_Type (10.4.1)	Address type ("mailing address"; "physical address"; "mailing and physical address").
Address	Address (10.4.2)	Contact address for organization.
City	City (10.4.3)	Contact address city.
State or Province	State_or_Province (10.4.4)	Contact address state or province.
Postal Code	Postal_Code (10.4.5)	Contact address Zip or postal code.
Country	Country (10.4.6)	Contact address country.
Contact Telephone Number	Contact_Voice_Telephone (10.5)	The telephone number by which individuals can speak to the organization.
Contact Fax Number	Contact_Facsimile_Telephone (10.7)	The telephone number of a facsimile machine of the organization.
Contact E-mail Address	Contact_Electronic_Mail_Address (10.8)	The address of the electronic mailbox of the organization.
<p>Template is based on a subset of the Federal Geographic Data Committee's "<i>Content Standard for Digital Geospatial Metadata</i>" <i>Version 2 - 1998</i>. (FGDC-STD-007 <i>June 1998</i>). The complete FGDC standard can be viewed at www.fgdc.gov/metadata/contstan.html</p> <p>*Areas in gray are descriptive headings. Data entry elements are in white.</p>		



Collaboration with EPA

- Region II EPA Mission: Create a toolset for storing, searching and retrieving metadata.
- CDC Tracking Mission: Create a toolset for storing searching and exporting metadata useable by network Members.
- Common Goal: Create a common data model for storing metadata based on FGDC.
- Result: 49 linked tables available in multiple formats, including Oracle, SQL Server, and Sybase. Includes most of the FGDC fields.



Results

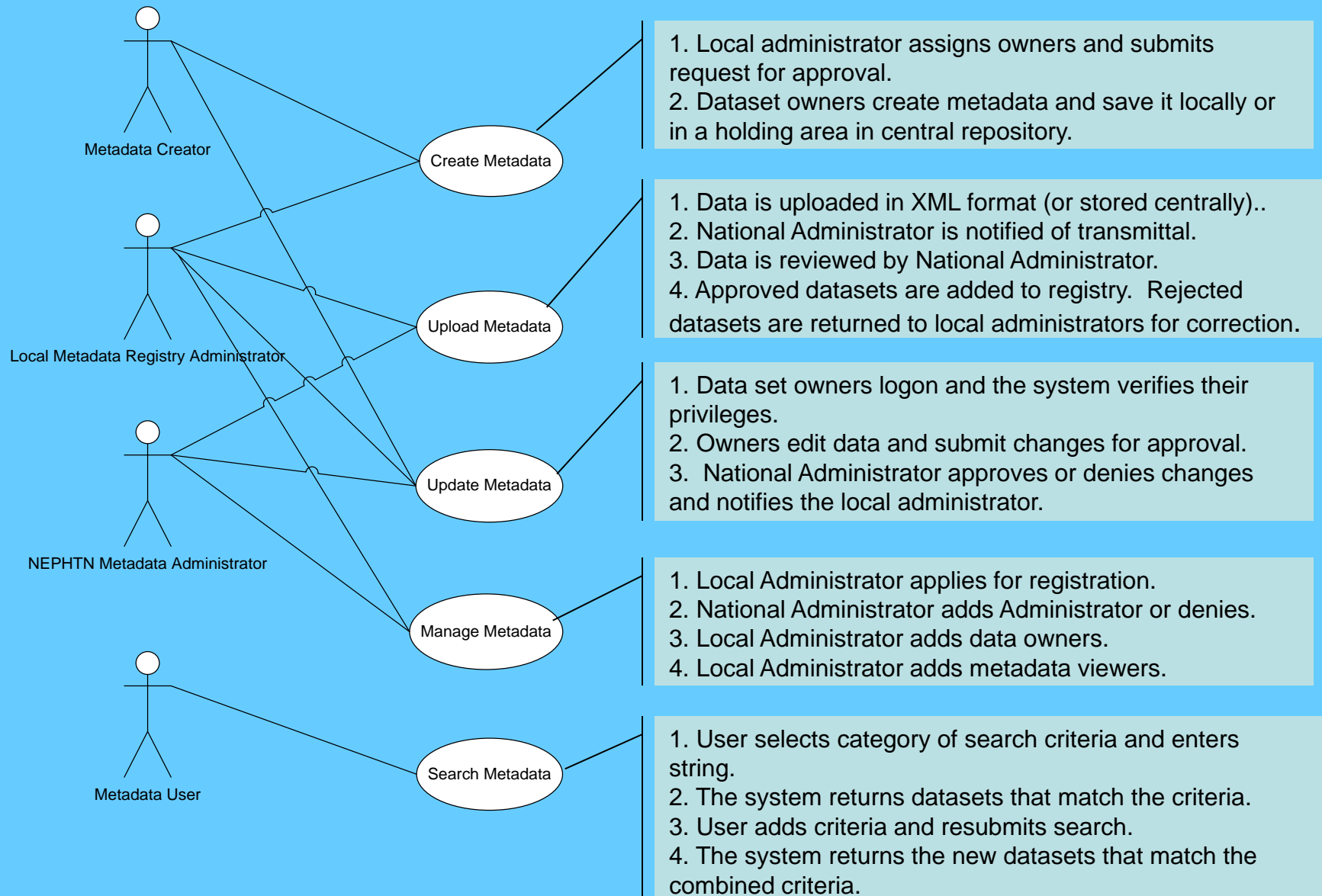
- An integrated logical data model that reflects the majority of the FGDC attributes.
- DDL has been generated for SQL Server and Oracle.
- The model is part of the toolset specifications for CDC and EPA in Washington has taken the model and used it as a basis for its own tool development.



Use Cases

- A Use Case is a technique for capturing the potential requirements of a new system or software change.
- Each use case provides one or more scenarios that convey how the system should interact with the end user or another system to achieve a specific business goal.
- Use cases typically avoid technical jargon, preferring instead the language of the end user or domain expert.

Metadata Process Flow Overview





Metadata Infrastructure Requirements

- The chosen metadata solution:
 - Needs to scale
 - Needs to be standardized
 - Needs to be interoperable across the Tracking Network
 - Needs access controls



Architecture Options

- Centralized Metadata Registry
 - Metadata Stored centrally
 - Each grantee manages their own metadata
 - Search from one central location
- Distributed Metadata Registry
 - Metadata stored locally by each grantee
 - Metadata registries are synched
 - Search from one central location or at each metadata registry
- Hybrid Metadata Registry
 - Some Metadata stored centrally, some locally
 - Some of the metadata in local registries are synched
 - Some metadata searchable centrally, some only locally



Recommendations

- Grantees begin to create metadata using the EPHTN core metadata standard
- Identify/develop a metadata creation tool that can incorporate the metadata standard
- Identify/develop a metadata registry tool that meets the identified requirements
- Phased deployment of the metadata infrastructure
- Develop process for adding additional FGDC elements to the core metadata template



Available Resources

- Metadata Template
- Metadata Tool Vision Document
- Metadata Use Cases
- NYC Metadata Toolset



Questions?

