Metadata Exposed

Unraveling the Mysteries of Data About Data

A Presentation By:

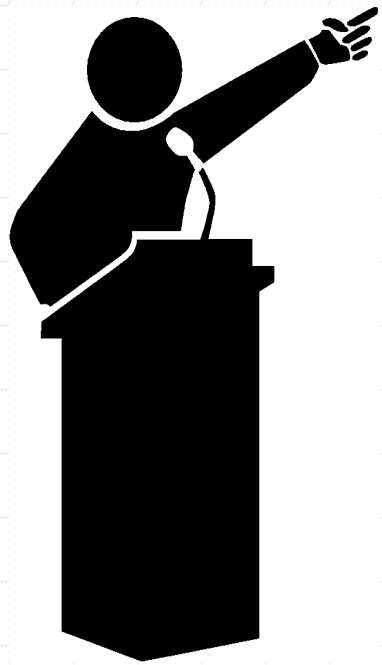
The Metadata / Data Quality Sub-Group

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.

What Users Can Learn From Metadata

- What method(s) were used to collect data.
- When were data collected?
- How were data processed?
- When were data last updated?
- Are there any data gaps?

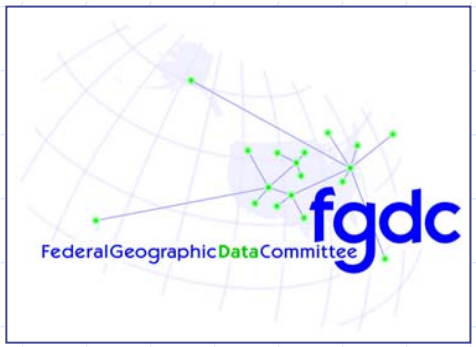


Why is Metadata Important?

- Protects investment in data
- Helps users to understand data
- Allows for users to discover the existence of data
- Limits liability
- Can reduce staff workload (once created)

Commonly Used Metadata Standards

- FGDC
- Dublin Core
- ISO 11179
- ISO 19115
- ISO 19139 *(currently undergoing review)*



Federal Geographic Data Committee (FGDC)

- Standard created for documenting geospatial datasets.
- Presidential Executive Order 12906 establishes that geospatial datasets created by Federal agencies must include FGDC-compliant metadata.
- Basic elements include:

Dataset Title	Purpose	Access Constraints
Contact Info	Citation	Time Period of Dataset
Status	Spatial Domain	Keywords
Attributes	Distribution	Metadata Reference

Dublin Core

Dublin Core is a higher level metadata standard. It consists of 16 elements and several element refiners. The elements are:

Coverage	Title	Date
Description	Audience	Format
Type	Contributor	Identifier
Relation	Creator	Language
Source	Publisher	
Subject	Rights	

International Organization for Standardization ISO 11179

- Specifies a basic set of data element characteristics necessary to share data.
- Metadata about data elements is stored in a data element registry.
- Basic attributes of data elements include:

Name	Classification Scheme	Data Type
Identifier	Keywords	Maximum Size
Version	Related Data Reference	Minimum Size
Context	Type of Relationship	Permissible Values

International Organization for Standardization ISO 19115

- ISO 19115 incorporates the FGDC standard.
- Allows for the documenting of both geographic and non-geographic data.
- Will be superceded in the United States by ISO 19139

International Organization for Standardization ISO 19139

- Based on ISO 19115
- Extensible Markup Language (XML) model
- Currently undergoing review
- Technical specification designation Winter 2004-2005

Importance of Metadata to the EPHT Network

Metadata / Data Quality Subgroup formed as part of Standards and Network Development Workgroup to:

- Develop a metadata template using a controlled vocabulary of EPHT Network datasets that will identify a core set of information that is needed to adequately document a dataset and its limitations for potential users.

- Develop a means to describe data using common ways to document datasets to facilitate data searches.

Actions Taken by Metadata / Data Quality Subgroup

- Talked with individuals involved with Public Health Information Network (PHIN) and the EPA Exchange Network

Actions Taken by Metadata / Data Quality Subgroup

- Talked with individuals involved with Public Health Information Network (PHIN) and the EPA Exchange Network
- Reviewed data inventories of the grantees to determine common elements

Actions Taken by Metadata / Data Quality Subgroup

- Talked with individuals involved with Public Health Information Network (PHIN) and the EPA Exchange Network
- Reviewed data inventories of the grantees to determine common elements
- Evaluated common elements against the currently accepted metadata standards

Results of Element Mapping

- Dublin Core too general to meet the requirements for describing an electronic dataset

Results of Element Mapping

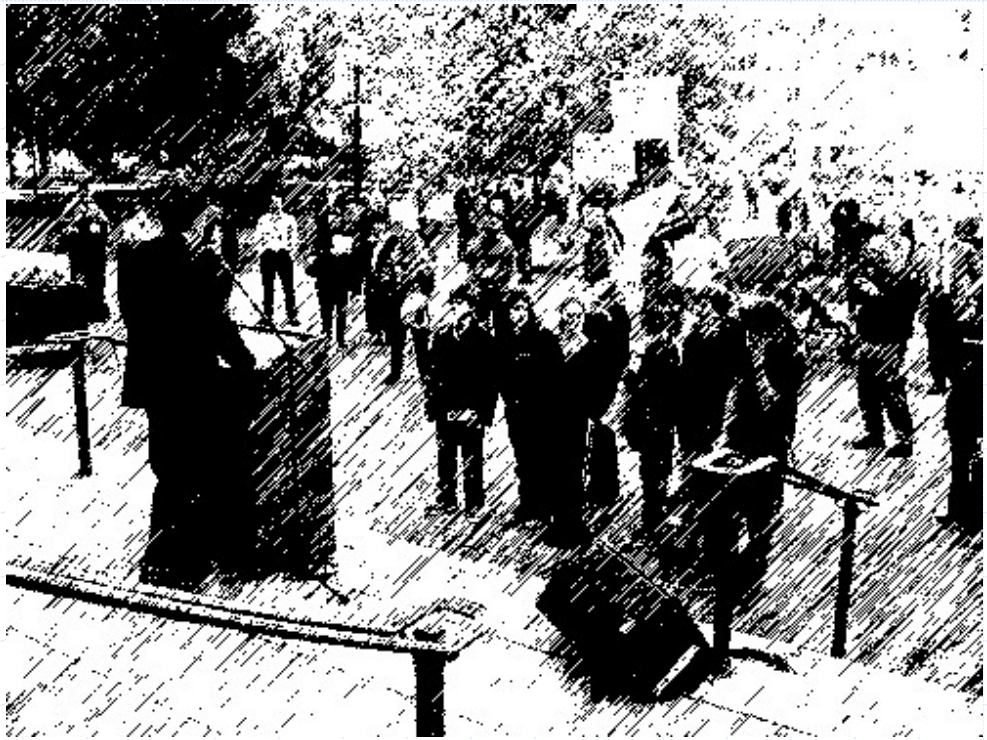
- Dublin Core too general to meet the requirements for describing an electronic dataset
- FGDC and ISO 19115 addressed most of the identified elements

Results of Element Mapping

- Dublin Core too general to meet the requirements for describing an electronic dataset
- FGDC and ISO 19115 addressed most of the identified elements
- FGDC recommended as the standard for the Network, until superceded by ISO standard.

Metadata Template Overview

- Demonstration of the Current Metadata Template

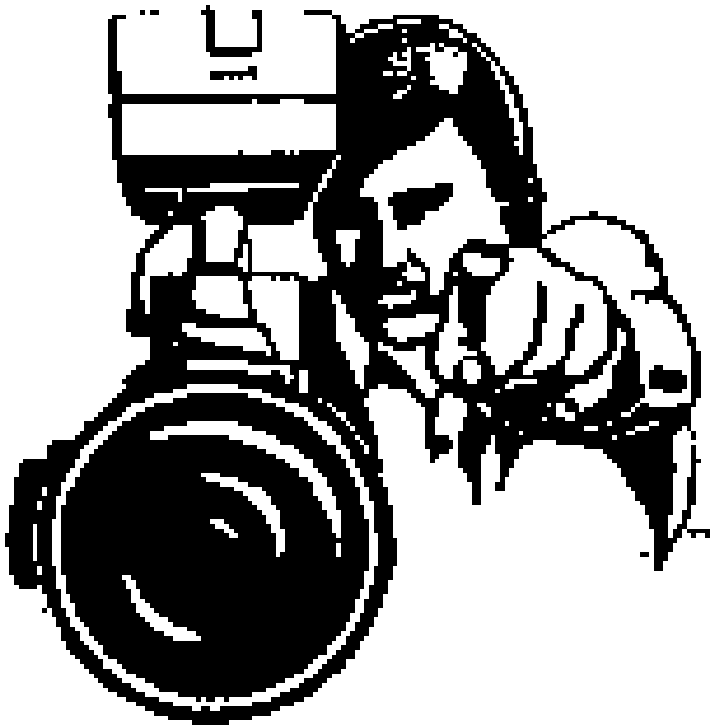


Element	FGDC Element	FGDC Definition
Section 1: IDENTIFICATION		
Citation	Citation (1.1)	Information to be used to reference the data set.
Originator	Originator (8.1)	The name of an organization or individual that developed the data set. If the name of editors or compilers are provided, the name must 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.
Abstract	Abstract (1.2.1)	A brief narrative summary of the data set.
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.)
Currentness	Currentness_Reference (1.3.1)	The basis on which the time period of content information is determined. (For example: an orthophotograph may have been compiled and delivered in June (publication date) but flown in February (ground condition).)
Keywords	Theme_Keyword_Thesaurus (1.6.1.1)	Reference to a formally registered thesaurus or a similar authoritative source of theme keywords.
Place	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)	Contant information for an individual or organization that is knowledgeable about the data set.
Credits	Data_Set_Credit (1.11)	Recognition of those who contributed to the data set.
Native data set environment	Native_Data_Set_Environment (1.13)	A description of the data set in the producer's processing environment, including items such as the name of the software (including version), the computer operating system, file name (including host-, path-, and filenames), and the data set size.

Section 2: DATA QUALITY		
Attribute Accuracy Report	Attribute_Accuracy_Report (2.1.1)	An explanation of the accuracy of the identification of the entities and assignments of values in the data set and a description of the tests used.
Logical Consistency Report	Logical_Consistency_Report (2.2)	An explanation of the fidelity of relationships in the data set and tests used.
Completeness Report	Completeness_Report (2.3)	Information about omissions, selection criteria, generalization, definitions used, and other rules to derive the data set.
Positional Accuracy Report	Positional_Accuracy Report (2.4)	An assessment of the accuracy of the positions of spatial objects.
Process Step	Process_Step (2.5.2)	Information about a single event (Can be a single collective description or individual process steps based upon: stages of processing, incorporation of sources, or project milestones)
Process Contact	Process_Contact (2.5.2.6)	The party responsible for the processing step information.
Section 3: SPATIAL DATA ORGANIZATION		
Indirect Spatial Reference	Indirect_Spatial Reference (3.1)	Name of types of geographic features, addressing schemes, or other means through which locations are referenced in the data set. (such as Geographic Names Index System (GNIS) place names, Public Land Survey System (PLSS) locations, and Federal Information
Section 4: SPATIAL REFERENCE		
Horizontal coordinate	Horizontal_Coordinate_System_D	The reference frame or system from which linear or angular quantities are measured and
Section 5: ENTITY AND ATTRIBUTES		
Detailed description	Detailed_Description (5.1)	Description of the entities, attributes, attribute values, and related characteristics encoded.
Section 6: DISTRIBUTION INFORMATION		
Distributor contact	Distributor_Contact (6.1)	The party from whom the data set may be obtained.
Distribution liability	Distribution_Liability (6.3)	Statement of the liability assumed by the distributor. (A legal-like secion that may: deny liability if the data are incorrect, incomplete, or misused, or limit 3rd party distribution of the data set.)

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.
Metadata standard name	Metadata_Standard_Name (7.5)	The name of the metadata standard used to document the data set.
Metadata access constraints	Metadata Access_Constraints (7.8)	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_Use Constraints (7.9)	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 Example



Keywords

Theme: Lead., TRI, Toxic Release Inventory

Place: Missouri

Description

Abstract

Facilities in Missouri that filed reports with the EPA's TRI (toxic release inventory) for lead releases. This information was downloaded from the EPA's web site and contains the data for years 1988-2002. From 1988 - 2000, reporting was required for all facilities who used 10,000 lbs of lead or produced 25,000 lbs of lead. For 2001, the thresholds changed to require all facilities which used or produced 100 lbs of lead. Facilities are required to report if they meet these criteria, regardless of whether or not they had any releases.

Purpose

Used for environmental regulation, these points indicate facilities that use or produce lead in Missouri.

Supplementary Information

Some facilities that are required to report will have 0 releases. From 1988 - 2000, if a facility's use or production was below the reportable levels, they were not required to report any releases.

Status of the data

Complete

Data update frequency: Annually

Time period for which the data is relevant

Date and time: 08-24-2004

Description:
publication date

Publication Information

Who created the data: Missouri Department of Health and Senior Services

Date and time: 2004

Data storage and access information

File name: TRI_LEAD

Type of data: vector digital data

Location of the data:

- \\DHSSNAS\VOL1\DATABASE\GISDATA\Environ\TRI\TRI_LEAD.shp

Data processing environment: Microsoft Windows 2000 Version 5.0 (Build 2195) Service Pack 4; ESRI ArcCatalog 8.3.0.800

Accessing the data

Size of the data: 0.682 MB

Data transfer size: 0.682 MB

Constraints on accessing and using the data

Access constraints: None

Use constraints:

None

Details about this document

Contents last updated: 20041006 at time 13305100

Contents last reviewed: August 24, 2004

Contents to be reviewed: Summer 2005

Who completed this document

Jeff Patridge

Missouri Department of Health and Senior Services

physical address:

930 Wildwood

Jefferson City, Missouri 65109

United States

573-522-8330 (voice)

573-751-6417 (fax)

patrij1@dhss.mo.gov

Hours of service: 7:30 a.m. - 4:00 p.m. M-F

Standards used to create this document

Standard name: FGDC Content Standards for Digital Geospatial Metadata

Standard version: FGDC-STD-001-1998

Time convention used in this document: local time

Metadata profiles defining additonal information

- ESRI Metadata Profile: http://www.esri.com/metadata/esriprof80.html

Horizontal coordinate system

Projected coordinate system name: NAD_1983_UTM_Zone_15N

Geographic coordinate system name: GCS_North_American_1983

Details

Grid Coordinate System Name: Universal Transverse Mercator

UTM Zone Number: 15

Transverse Mercator Projection

Scale Factor at Central Meridian: 0.999600

Longitude of Central Meridian: -93.000000

Latitude of Projection Origin: 0.000000

False Easting: 500000.000000

False Northing: 0.000000

Planar Coordinate Information

Planar Distance Units: meters

Coordinate Encoding Method: coordinate pair

Coordinate Representation

Abscissa Resolution: 0.001024

Ordinate Resolution: 0.001024

Geodetic Model

Horizontal Datum Name: North American Datum of 1983

Ellipsoid Name: Geodetic Reference System 80

Semi-major Axis: 6378137.000000

Denominator of Flattening Ratio: 298.257222

Altitude System Definition

Resolution: 0.000010

Encoding Method: Explicit elevation coordinate included with horizontal coordinates

Bounding coordinates

Horizontal

In decimal degrees

West: -95.243710

East: -89.367274

North: 40.360127

South: 36.208617

In projected or local coordinates

Left: 309418.220676

Right: 808747.375189

Top: 4467729.293927

Bottom: 4012557.081791

Spatial data quality

Horizontal positional accuracy

Latitude and longitude generated using Centrus Geocoding software. Locations that were originally centroid based due to missing address information were cross checked using EPA FRS (facility registry system), USGS quads, and aerial photography (date range 94-00)

Vertical positional accuracy

None

Spatial data description

Vector data information

ESRI description

TRI_LEAD

SDTS description

Description | Spatial | Attributes

Details for TRI_LEAD

Type of object: Feature Class

Number of records: 1012

Attributes

- FID
- Shape
- Add_std
- City_std
- State_std
- Zip_std
- Zip4_std
- Latitude
- Longitude
- MATCH_CODE
- LOC_CODE
- YEAR_REL
- FACILITY
- ADDRESS
- CITY
- COUNTY
- STATE
- ZIP
- TRI_FID
- FUG_AIR_
- STACK_AIR

  Alias: STACK_AIR

  Data type: Number

  Width: 9

  Definition:

  Air releases through confined air streams such as stacks, vents, ducts, or pipes

  Definition Source:

  Environmental Protection Agency

- TOTAL_AIR
- SURF_WATER
- UND_I
- UND_II_IV
- RCRA_LAND
- OTHER_LAND
- LAND
- SURF_IMP

Future Efforts for the Metadata / Data Quality Subgroup

- Recommend tools for creating metadata

Future Efforts for the Metadata / Data Quality Subgroup

- Recommend tools for creating metadata
- Establish guidelines for the creation of metadata registries on the Network

Future Efforts for the Metadata / Data Quality Subgroup

- Recommend tools for creating metadata
- Establish guidelines for the creation of metadata registries on the Network
- Establish guidelines for creating metadata for analytical, conversion, and geospatial tools utilized by the Network.

Future Efforts for the Metadata / Data Quality Subgroup

- Recommend tools for creating metadata
- Establish guidelines for the creation of metadata registries on the Network
- Establish guidelines for creating metadata for analytical, conversion, and geospatial tools utilized by the Network.
- Establish a process to address emergent metadata requirements not covered by existing metadata standards.

Conclusion

Thanks for Listening