Developing data management project (3): Data organization and preservation

IST400/600

Jian Qin

Review: Science data life cycle

Metadata: descriptions about datasets
- Identification information
- Subject and coverage
- Lineage: processing history
- Distribution information
- ...

Collect
Present
Store
Search
Retrieve

Use
Organization of data

- Organizing datasets
  - Data files have formats
  - Data files have multiple components
  - Storage media
  - Physical locations (file directories)

- Organizing information about datasets
  - Metadata for datasets
  - Metadata schemas
  - Metadata records
  - Storage of these records
  - Retrieval of datasets by using metadata

Datasets need to be associated with metadata so that they can be found and used.

Metadata is also data and needs to be stored in databases for retrieval.

Storage technology

- Disk systems

- Tape systems

- Storage area network
Metadata for datasets

- Determine the metadata schema
  - Adopt a standard or develop an in-house schema?
  - Community-based metadata description conventions
  - Understand how scientists search and use data
- Create metadata records
  - Who will create the records? Scientists? Technicians? Data management staff?
  - When will metadata records be created? At the time when data are collected or after?
  - If manual or semi-automated, what tools will be used?

An example

- Institutionalize Metadata Before It Institutionalizes You
A 5-minute quiz (1)

- What are the benefits of metadata for institutions?
- What are the benefits of metadata for individuals?
- ____________ is a good way for metadata creation.
- What metadata information is involved in data planning?
- Give two examples of metadata tools.

A 5-minute quiz (2)

- What are the major responsibilities of managers in developing metadata procedures and policies?
- Give an example of metadata production obstacles and the corresponding recommendation
Once metadata schema, procedures, and policies are defined, what next?

**IMPLEMENTATION OF METADATA APPLICATIONS**

**Conceptual model vs. implementation model**

- Identify entities, attributes, and relationships
- Define data fields, data types, data entry rules and controls
- Specify constraints, rules, operations
- Design queries, triggers, and stored procedures, and user interfaces
Conceptual modeling tools

- E-R Diagram
- Unified Modeling Language (UML)
- XML/RDF (Resource Description Framework)
- Web Ontology Language (OWL)

E-R modeling

- Entities
- Attributes
- Relationships
UML modeling (1)

- Use case diagrams
  - Actor who initiates the events involved in the task
  - Communication: connection between actor and use case


UML modeling (2)

- Class diagrams:
  - Provides an overview of a system by showing its classes and relationships among them

Types of relationships:
- Association
- Aggregation
- Generalization

---

IST400/600 Scientific Data Management 13
IST400/600 Scientific Data Management 14
UML modeling (3)

- Packages
  - Package: a collection of logically related UML elements
  - Dependencies: one package depends on another if changes in the other could possibly force changes in the first

UML modeling (4)

- Object diagrams
  - Contains instances instead of classes
UML modeling (5)

- Component and deployment diagrams
  - Component: a code module
  - Deployment diagram: shows physical configurations of software and hardware.

UML modeling (6)

- Other diagrams in UML:
  - Sequence diagrams
  - Collaboration diagrams
  - Statechart diagrams
  - Activity diagrams
  - Component diagrams
## Example: Metadata fields mapped to data flow

<table>
<thead>
<tr>
<th>Data Development Stage</th>
<th>Metadata Information</th>
</tr>
</thead>
</table>
| Data Planning          | Identification Information  
|                        | title, originator, abstract, purpose, keywords, time period  
|                        | Data Organization  
|                        | point, raster, vector  
|                        | Spatial Referencing  
|                        | coordinate system and datum  
|                        | Entity and Attributes (planned) |
| Data Processing        | Data Quality  
|                        | completeness, positional accuracy, geoprocessing steps |
| Data Analysis          | Data Quality  
|                        | attribute accuracy, analysis steps  
|                        | Entity and Attributes (results)  
|                        | Metadata Reference |

## UML modeling for the example

- What would a use case diagram look like?
- What would a package diagram look like?
- What would an object diagram look like?