Towards A Preservation Content Model for Numeric Data Collections: PREMIS and FEDORA

David Gewirtz
Yale Library & AM&T
Gretchen Gano
New York University

IPRES Cornell University October 2006

Towards A Preservation Content Model for Numeric Data Collections:

Overview

- **Framing the context of the Research**
  - Yale's Social Science Data Archive (SSDA)
- **Content Models for Preservation and Access**
  - Original preservation content model
    - PREMIS Implementation
- **Evolution of the Preservation Content Model**
  - Drivers, Standards, Models
  - Evolution of the content model
    - Atomic vs compound model
    - SIP to AIP Transformation and Prototype Models
- **Benefits and Consideration of the new content model**
  - Access View
  - Preservation View

SSDA at Yale

- **The Social Science Data Archive (SSDA)** is the repository and reference center at Yale for machine-readable data sources in the social sciences.

  The SSDA owns and maintains a major collection of data from academic surveys, public opinion surveys, government agencies, international organizations, and related groups.

Context: SSDA at Yale

- Statcat catalog is postgreSQL database with a PHP front end, Lucene search engine
- Records are a subset of DDI 2.1
- Includes four types of records
  - SSDA collections
  - ICPSR harvested records, which link directly to ICPSR catalog
  - Internet data sites
  - CD-ROM holdings/database links (Source OECD)

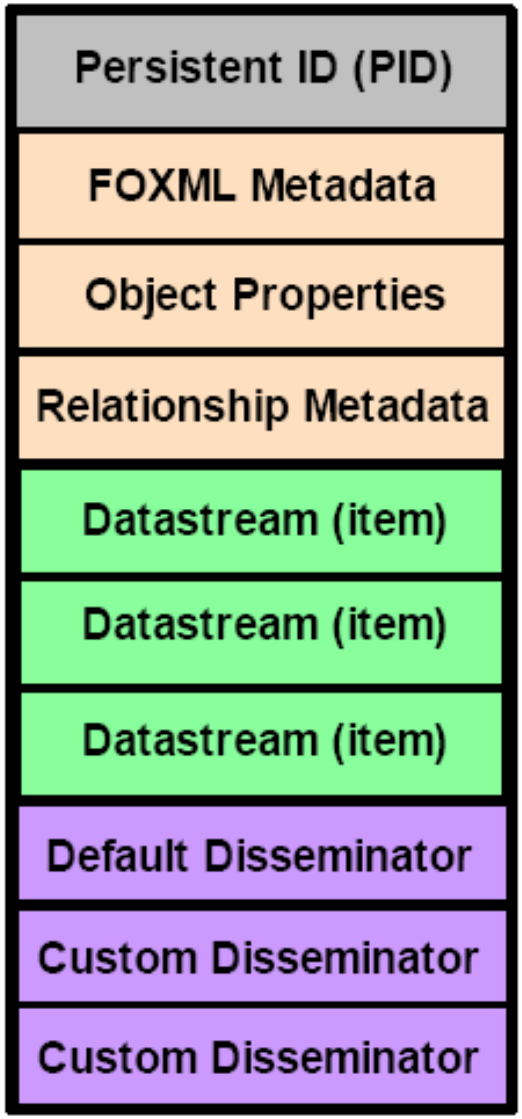
Components of Data Study

MOODAMERICA_Data_File (ascii)
MOODAMERICA_Code_Scheme (ascii)
MOODAMERICA_Questionaire (pdf)
MOODAMERICA_Setup_File
MOODAMERICA_DDI file (xml)

SSDA Development

- Enhance and deepen functionality to support resource discovery
- Support emerging standards in digital preservation
- Reconceptualize collection development for university archive
  - Unique data collections
  - Faculty research data
  - Provide avenue for contribution to national and international data archives/services
- Build capacity to support new kinds of collection development

Fedora Digital Object Model



Digital object identifier

***Internal Metadata*:** key metadata necessary to manage the object

***Item Perspective*: Set of content or metadata items**

***Service Perspective*: methods for disseminating "views" of content**

SSDA Submission Information Package

- DDI
- Dublin Core
- PREMIS representation
- PREMIS file info
- Data files
- Codebook(s) and other materials

PREMIS mark-up for the SSDA collection

- Link to XML for the PREMIS representation file
  - Key decision:
    - Study as intellectual entity
      - Codebooks, set-up files and DDI metadata described as dependencies

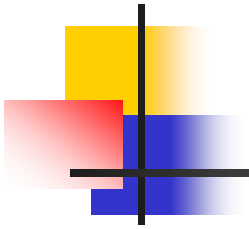
PREMIS mark-up for the SSDA collection

- File level information
  - Key points:
    - Format, autogenerated from PRONOM Format Registry
    - File Fixity information
    - Event tracking

FEDORA object model for SSDA

PID	FEDORA persistent ID
FoxML metadata	
Object Properties	
Relationship metadata	
Datastream 1	Dublin Core metadata
DS2	DDI metadata
DS3	.dat file (numeric data file)
DS4	Spss set-up script
DS5	.pdf codebook
DS6	PREMIS Representation
DS7	PREMIS file information for DS2
DS8	PREMIS file info for DS3
DS9	PREMIS file info for DS4
DS10	PREMIS file info for DS5

Prototype Design Considerations



- Fedora Object Types
- Fedora Content Model for Preservation
- Inclusion of Packaging Information in Fedora Objects
- OAI Integration – Action Assets as used in the DLF Aquifer Project

Towards A Preservation Content Model for Numeric Data Collections

- **Evolution of the Preservation SSDA Content Model**
  - Drivers: Transformation of SIP to AIP
    - The SIP represented a compound Fedora Object that was not optimized for access, storage management and preservation.
    - AIP requirements suggested that the compound model be abandoned for an Atomistic model where the design:
      - Accounts for packaging information
      - CI and RI are contained in a single Fedora Resource Object
      - PDI is contained in a separate Fedora mdPDI Object

Towards A Preservation Model for Numeric Data Collections:

- **Evolution of the Preservation SSDA Fedora Content Model**
  - Drivers: Preservation
    - Incorporation packaging information into the Fedora Object
    - Preservation requirements dictate that within the AIP content information should include representation information

Towards A Preservation Model for Numeric Data Collections

- **Evolution of the Preservation SSDA Content Model**
  - Drivers: Repository Management
    - Constrain size of Fedora Objects
    - Simplification of preservation treatments through the separation of content information from PDI or PREMIS metadata.
    - Simplification of the Archival Storage Mapping Infrastructure through the assignment of handles to Fedora Objects.

Towards A Preservation Model for Numeric Data Collections

- **Evolution of the Preservation SSDA Content Model**
  - Drivers: Access to end users
    - Atomistic model increases the discovery of objects to end users with the use of handles.
    - Atomistic model may prove to facilitate the repurposing and re-use of Fedora Objects.
    - Handles assigned to Fedora Objects usefully limits the size of the handle database.

A Preservation Model for Numeric Data Collections

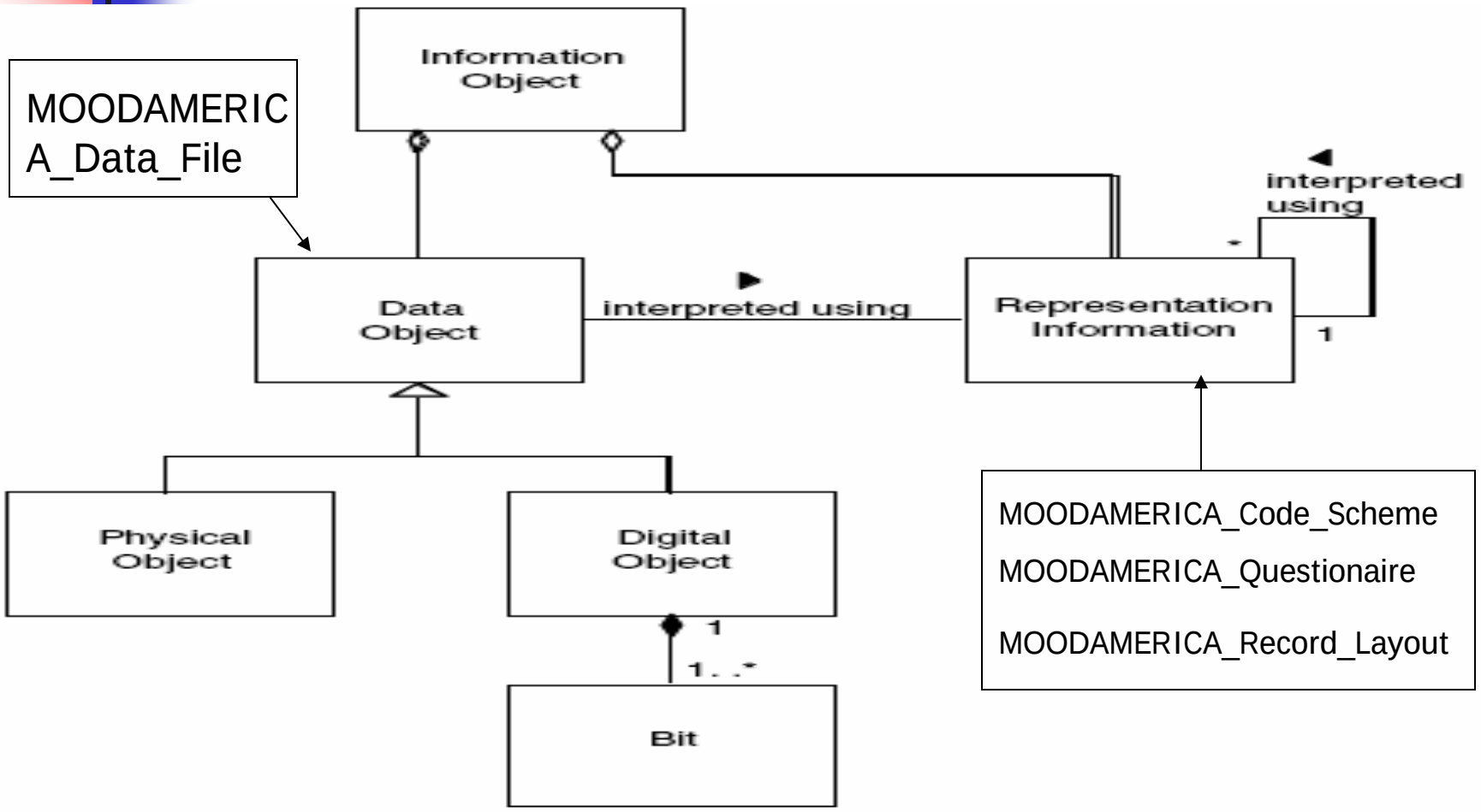
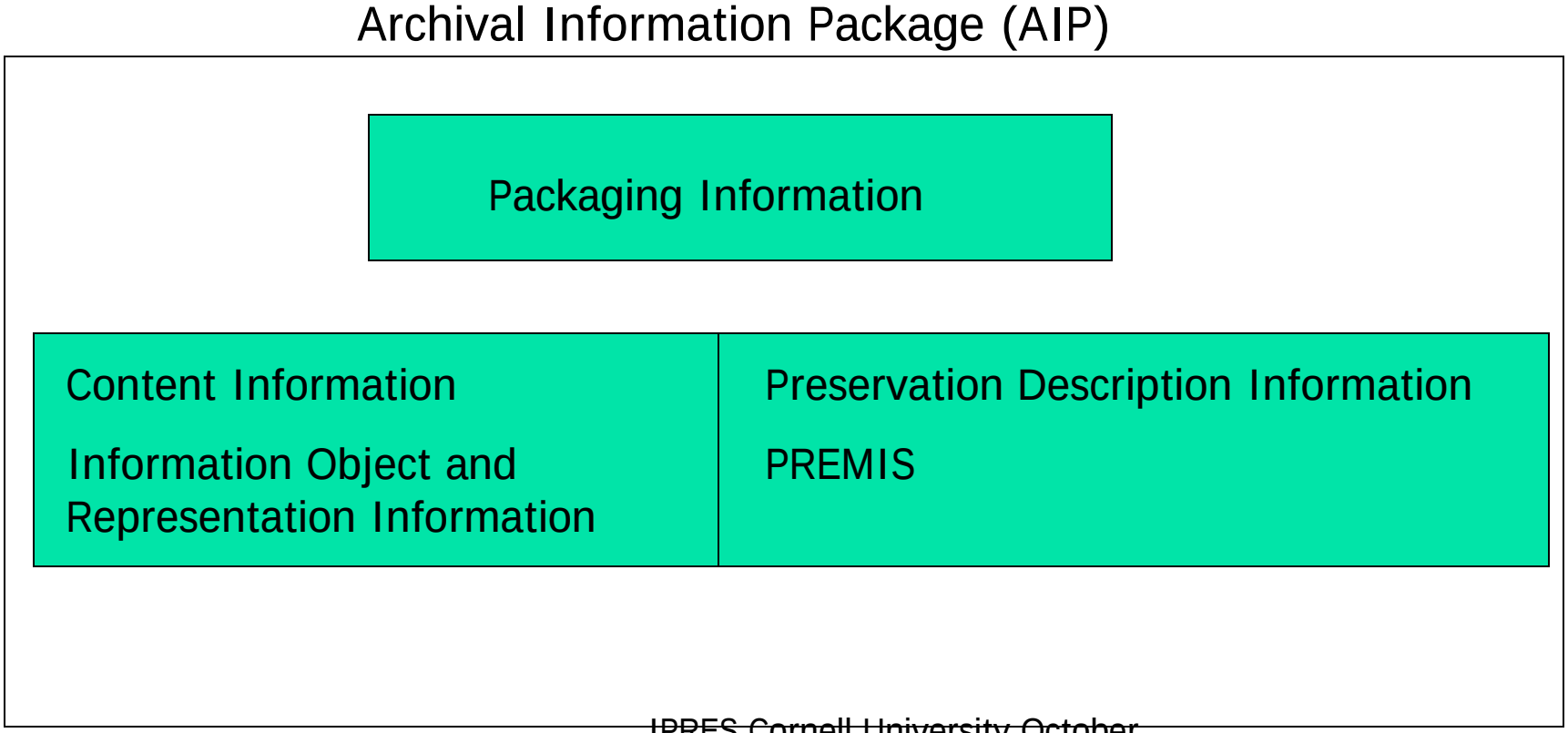


Figure 4-10: Information Object

A Preservation Model for Numeric Data Collections



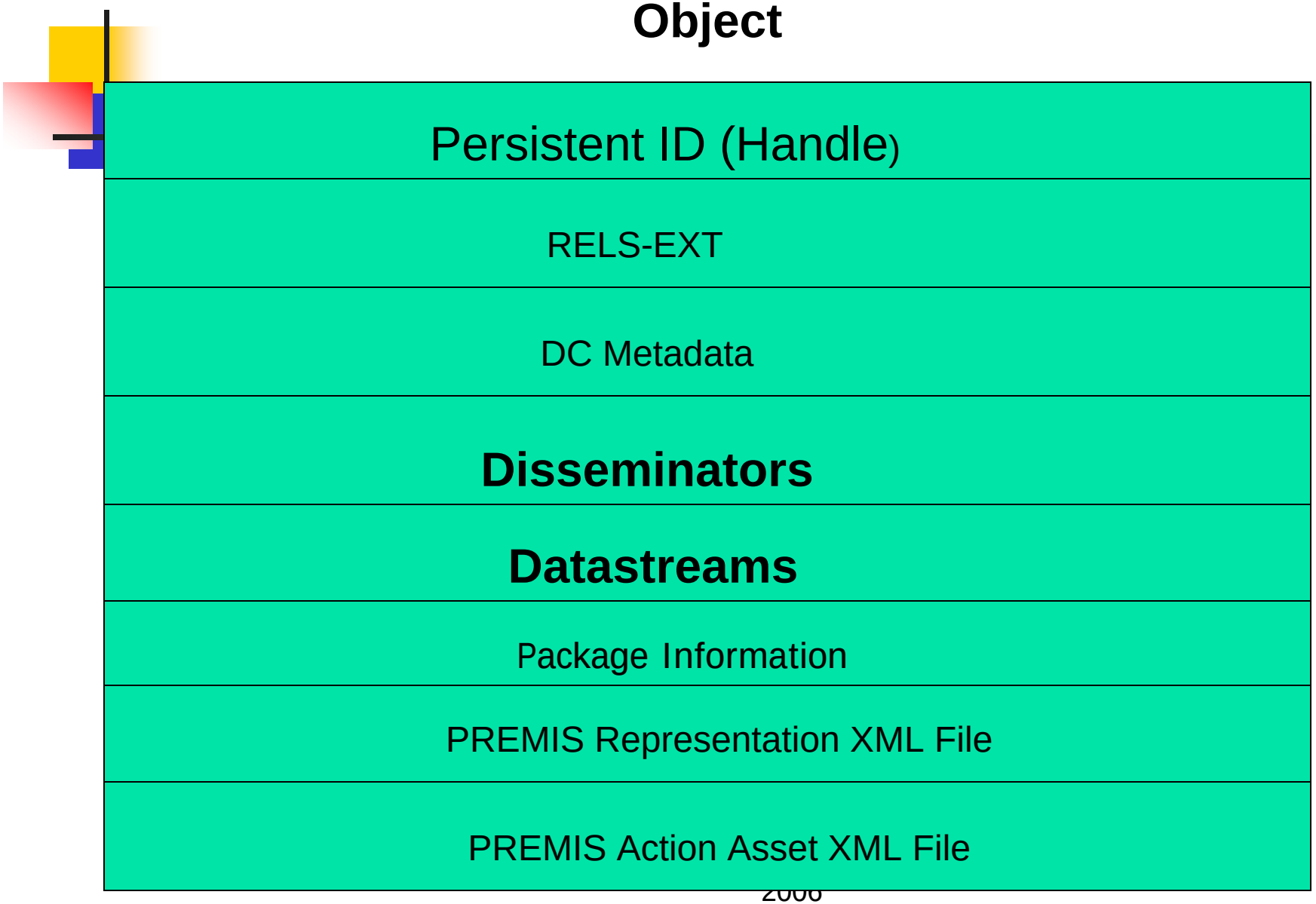
Towards A Preservation Model for Numeric Data Collections

- THE N(sdl)DR Fedora Object Types or Nodes
  - Resource
  - Metadata
  - Agent
  - Aggregation
- Extend model
  - Add mdPDI Node
  - Add mdPDI relationship type to the Fedora REL-EXT schema

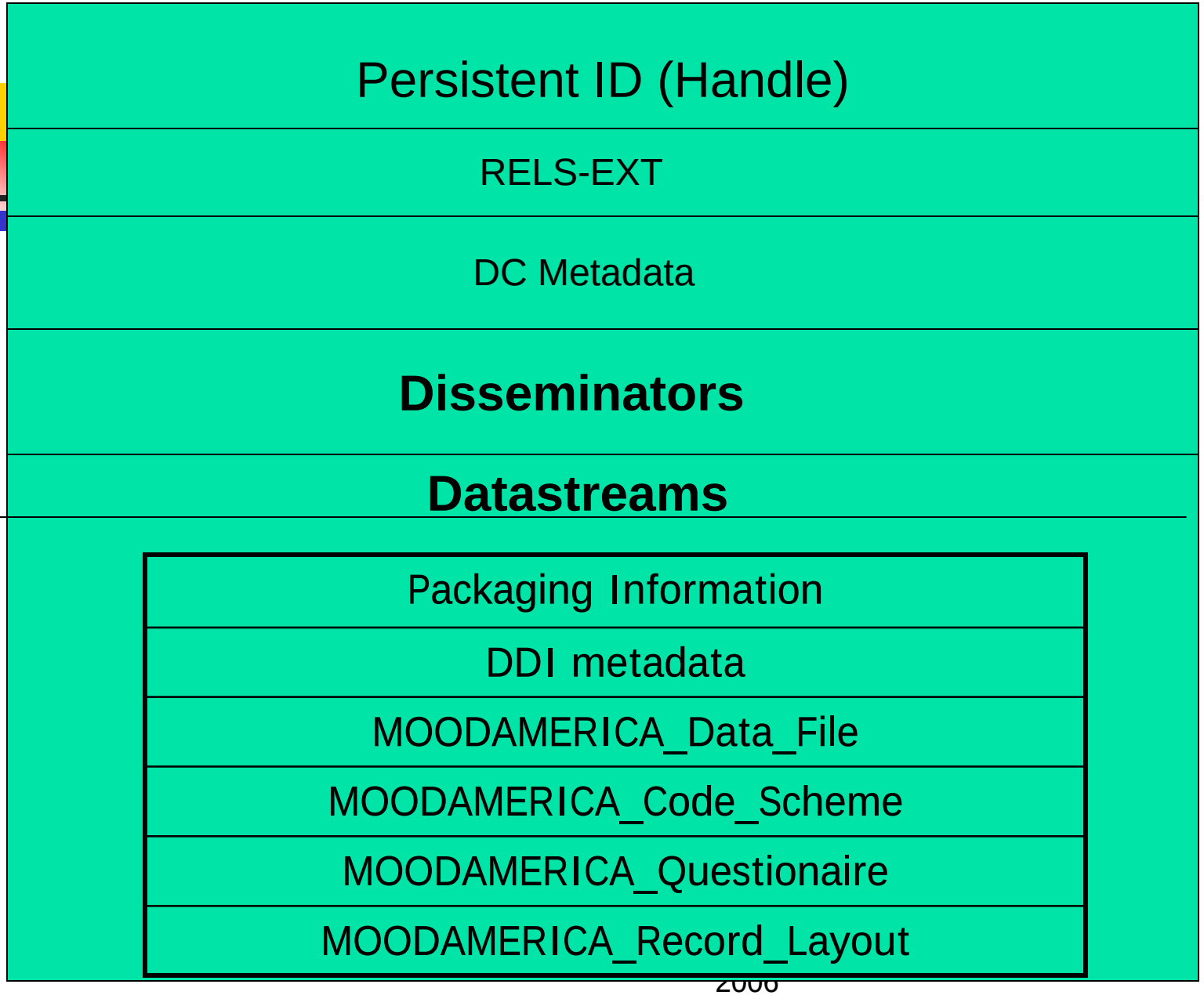
Towards A Preservation Model for Numeric Data Collections

- For each study there are three Fedora content models:
  - Representation content model
  - Data content model
  - Preservation content model

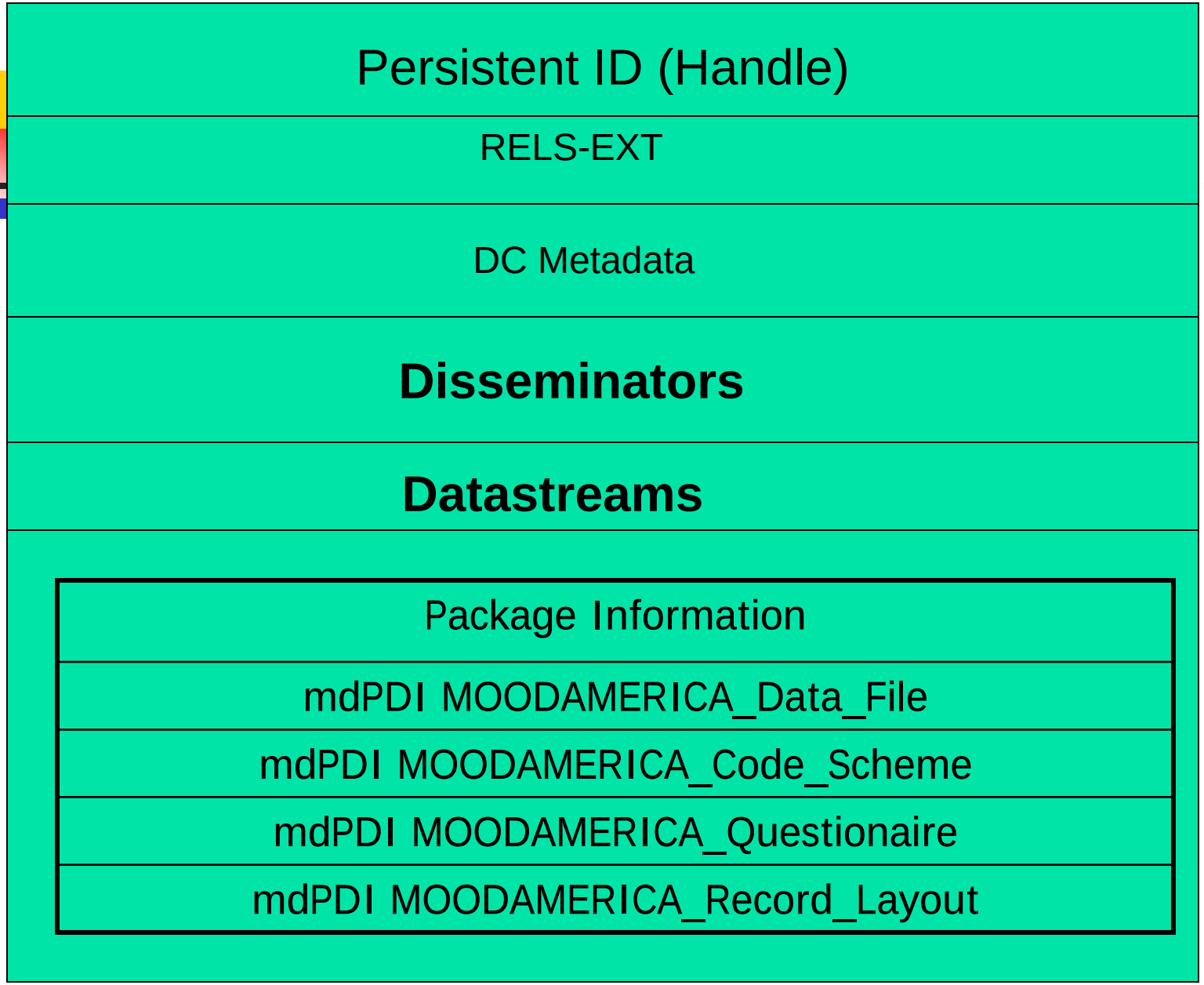
Content Model for PREMIS Representation Object



Content Model for SSDA Resource Object



Content Model for mdPDI Object



Towards A Preservation Model for Numeric Data Collections

- Benefits and Consideration of Prototype Designs
  - Access View
    - OAI-PMH interface asset actions
  - Preservation View
    - Facilitation AIP Recovery
    - Management of AIP editions and version
    - Preservation treatments on AIP components
    - Diversity and number of formats in a collection
    - Scale