Data Models for Long-Term Preservation for Diverse Information Objects

Shalini R. Urs
International School of Information Management
University of Mysore
Mysore, India
shalini@isim.ac.in

Outline

- Digital preservation issues –
  - why the paranoia ?
  - general issues
- OAIS Reference Model
- Preservation Metadata and PREMIS
- Data Model and data dictionary :
  - Semantic Units
- PREMIS Implementation survey and case studies

Digital Preservation- preamble

- Increasing quantum of digital information :161 exabytes of information in 2006, and growing @ 6 times a year
- Information management is not only critical, but has to contend with diverse and difficult issues of preservation
- The horror of digital dark ages haunts everyone
- Digital preservation - extending the life and afterlife of digital materials, encompasses two dimensions – bit stream maintenance and content accessibility.

Preservation Metadata

- It is hard to discuss information management topics today, without encountering the term *metadata* (Lavoie and Gartner, 2005)
- Digital preservation strategies for technological hazards may encompass - migration, emulation, Technology preservation, Encapsulation, Digital archeology and others
- Most strategies require some information to be collected and stored. And this is achieved by using metadata

PREMIS

- International Digital preservation efforts have also necessarily focused on the preservation metadata.
- ‘Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO)
- Preservation Metadata: Implementation Strategies (PREMIS), an international working group was established in 2003 by OCLC and RLG.

Preservation Metadata

- PREMIS WG defines “preservation metadata” as *the information a repository uses to support the digital preservation process*.
- PM is a self documentation tool.
- PM regulates, standardizes and automates the process of digital preservation management.
- PM is an essential component of digital preservation strategy and tool
- PM becomes a part and parcel of digital preservation strategy.

Scope of PM

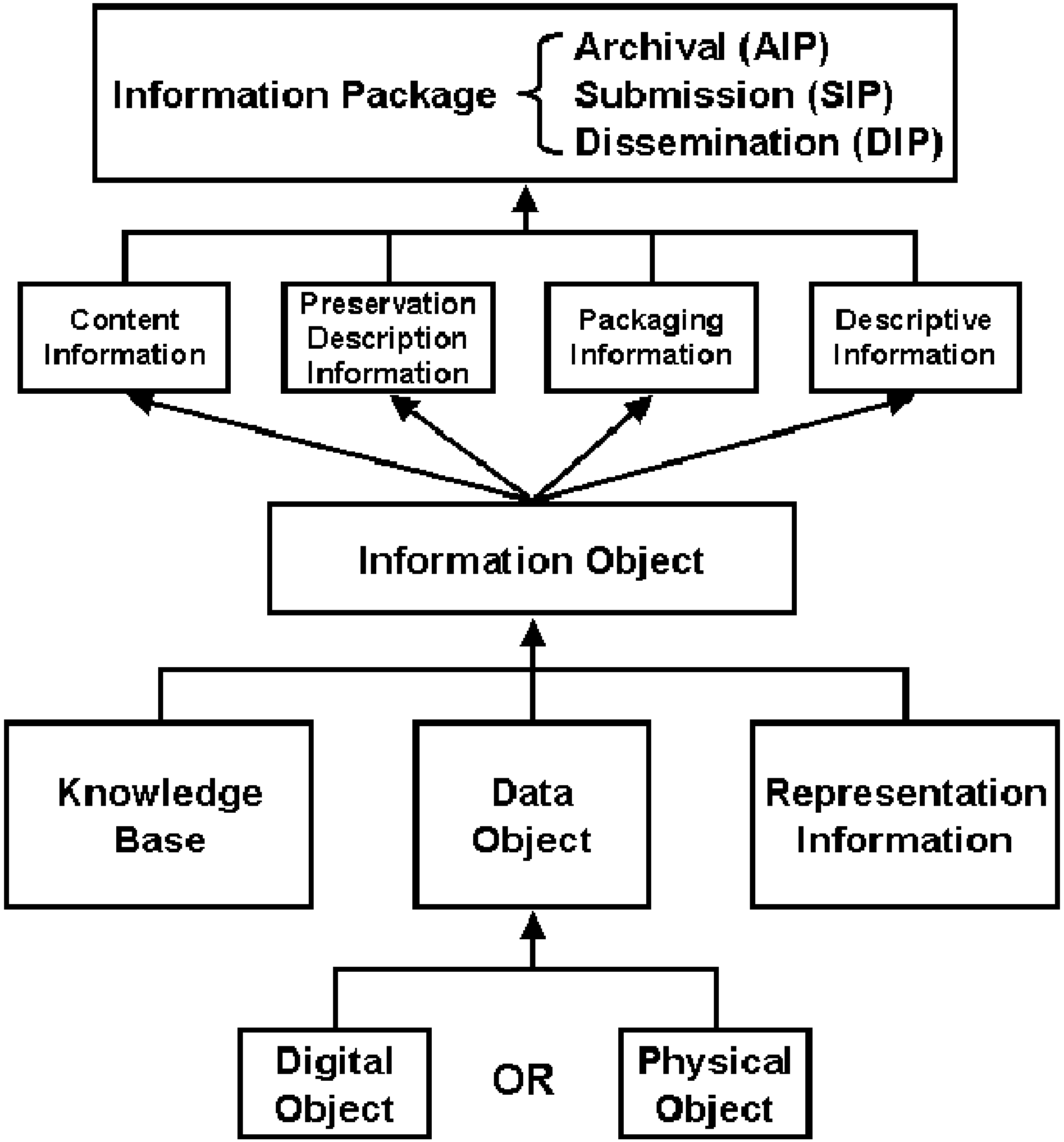
- Provenance
- Authenticity
- Preservation Activity
- Technical Environment
- Rights Management

The OAIS Information model

Defines four different Information Objects –

- Content Information – the information that requires preservation
- Preservation Description Information (PDI) – any information that will allow the understanding of the content information over an indefinite period of time.
- Packaging Information – the information that binds all other components into a specific medium
- Descriptive Information – the information that helps users to locate and access information of potential interest.

OAIS Information Model



THE PREMIS Data Model

- Implementation independent - the core elements define information that a repository needs to know, regardless of how, or even whether, that information is stored
- Because the emphasis is on the need to know rather than the need to record or represent in any particular way, the PREMIS WG preferred to use the term “semantic unit” rather than “metadata element.”
- The Data Dictionary names and describes semantic units, and the properties of entities
- Data Model consisting of three semantic units – ***entities***, ***relationships***, and ***properties,*** serves as the next level of building block for the digital preservation initiatives

Entities

- Intellectual Entities-a coherent set of content that is reasonably described as a unit
- Objects or Digital Object - a discrete unit of information in digital form.Object entities are described in three sub-types-
  - Bit stream
  - File
  - Representation
- Events - an action that involves at least one object or agent known to the preservation repository.
- Rights - or Rights Statements, are assertions of one or more rights or permissions pertaining to an object and/or agent
- Agents - a person, organization, or software program associated with preservation events in the life of an object

Relationship

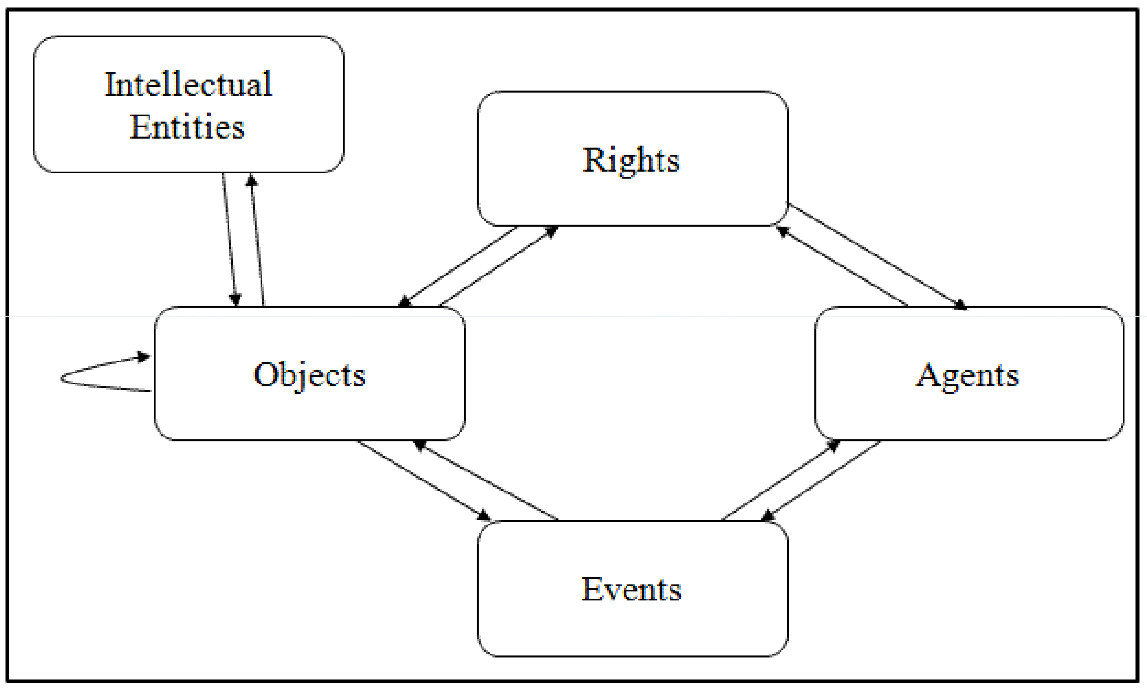
Statements of association between instances of entities. Relationships may be of 3 types

- *Structural relationships* showing relationships between parts of objects
- *Derivation relationships* resulting from the replication or transformation of an Object.
- *Dependency relationship* exists when one object requires another to support its function, delivery, or coherence of content

Properties

- The Semantic units are the properties of an entity.
- In some cases a semantic unit can be a container that groups a set of related semantic units and the grouped subunits are called semantic components of the semantic unit.

PREMIS 2.0 Data Model



- The data model includes five entities
  - Intellectual Entities
  - Objects
  - Events
  - Rights
  - Agents

PREMIS and METS

- The Metadata Encoding and Transmission Standard (METS) - a data encoding and transmission specification, expressed in XML, provides the means to convey the metadata
- The METS XML schema created in 2001, received NISO registration in 2004, is supported by the Library of Congress as its maintenance agency, and is governed by the METS Editorial Board.
- The PREMIS schema has been endorsed by METS as an approved extension schema for METS.
- The METS schema is widely used by digital repositories as a packaging mechanism for objects and their associated metadata.
- PREMIS is registered as a recognized metadata scheme to be used as administrative metadata with METS

PREMIS implementation survey

- The major findings and observations of the study are -
- Trends such as storing preservation metadata in either XML structures or relational database management systems (RDBMS) and allowing the design of systems to be able to incorporate multiple strategies for digital preservation continue to hold true.
- Very few off-the-shelf tools are being used for implementing preservation metadata. The main three tools (primarily relating to technical metadata creation) in use are –
  - DROID/PRONOM (Digital Record Object Identification and format registry)
  - JHOVE (JSTOR/Harvard Object Validation Environment)
  - The National Library of New Zealand Metadata Extraction Tool
- Many implementation methods are being developed in-house for repositories as part of ingest workflow

- The Study noted that most repositories identified well with the PDD data model, implementing equivalent metadata entities for intellectual entities, object entities at representation, file and (less commonly) bitstream levels, event entities and agent entities. Rights entities were only occasionally created

PREMIS implementation case studies

- SHERPA Digital Preservation Project
- Florida Digital Archive Project
- PARADIGM (Personal Archives Accessible in Digital Media)
- MathArc (Ensuring Access to Mathematics over Time
- NGDA (The National Geospatial Digital Archive)

SHERPA Digital Preservation Project

- **The purpose of SHERPA DP is to create a collaborative, shared preservation environment for the SHERPA project framed around the OAIS Reference Model**
- **The project brings together the SHERPA institutional repository systems with the preservation repository established by the Arts and Humanities Data Service (AHDS)**
- **A subset of the PREMIS entities and metadata were considered essential by the SHERPA DP-they are**
- **Object:** Information about an asset or file
- **Event:** Information that describes important events in the lifecycle of the digital object, such as migration, transfer of an object between repositories, or deletion from the preservation archive

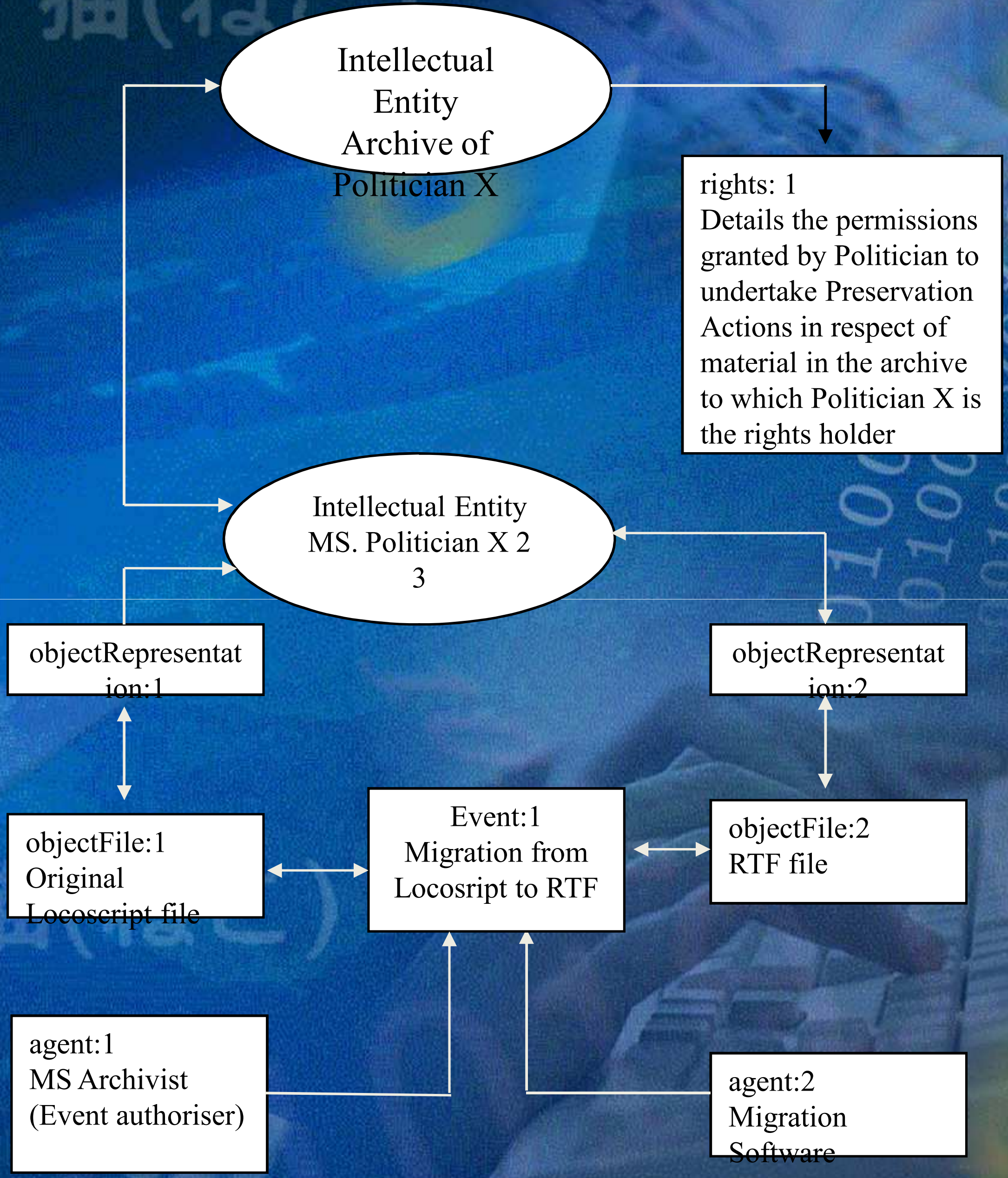
Florida Digital Archive Project

- The Florida Digital Archive is based on DAITSS (Dark Archive in The Sunshine State), a preservation repository management application, available as open source software.
- The mission of the FDA is to provide a cost-effective, long-term preservation repository for digital materials in support of teaching and learning, scholarship, and research in the state of Florida
- The PREMIS semantic units obtained from the information available from DAITSS as follows:
- In the DAITSS model there are Intellectual Entities and two types of Objects: files and bitstreams. The third PREMIS object – 'Representations' are not currently tracked.

PARADIGM

- The **P**ersonal **Ar**chives **A**ccessible in **Dig**ital **M**edia (PARADIGM) project brings together libraries of the Universities of Oxford and Manchester
- Since PARADIGM explores digital preservation from ‘personal’ and ‘collecting’ perspectives in the context of a ‘hybrid archive’ , they realized that putting METS AIP together for digital archivists is difficult and therefore CAIRO, a user-friendly ingest tool for archivists to ingest complex collections of born-digital materials was developed.
- CAIRO maps output to preservation metadata standards (PREMIS and object-specific) in METS package for Fedora submission via DirIngest according to content models

PARADIGM model



Intellectual Entity Archive of Politician X

rights: 1
Details the permissions granted by Politician to undertake Preservation Actions in respect of material in the archive to which Politician X is the rights holder

Intellectual Entity MS. Politician X 2 3

objectRepresentation:1

objectRepresentation:2

objectFile:1
Original Locoscript file

Event:1
Migration from Locosript to RTF

objectFile:2
RTF file

agent:1
MS Archivist
(Event authoriser)

agent:2
Migration Software

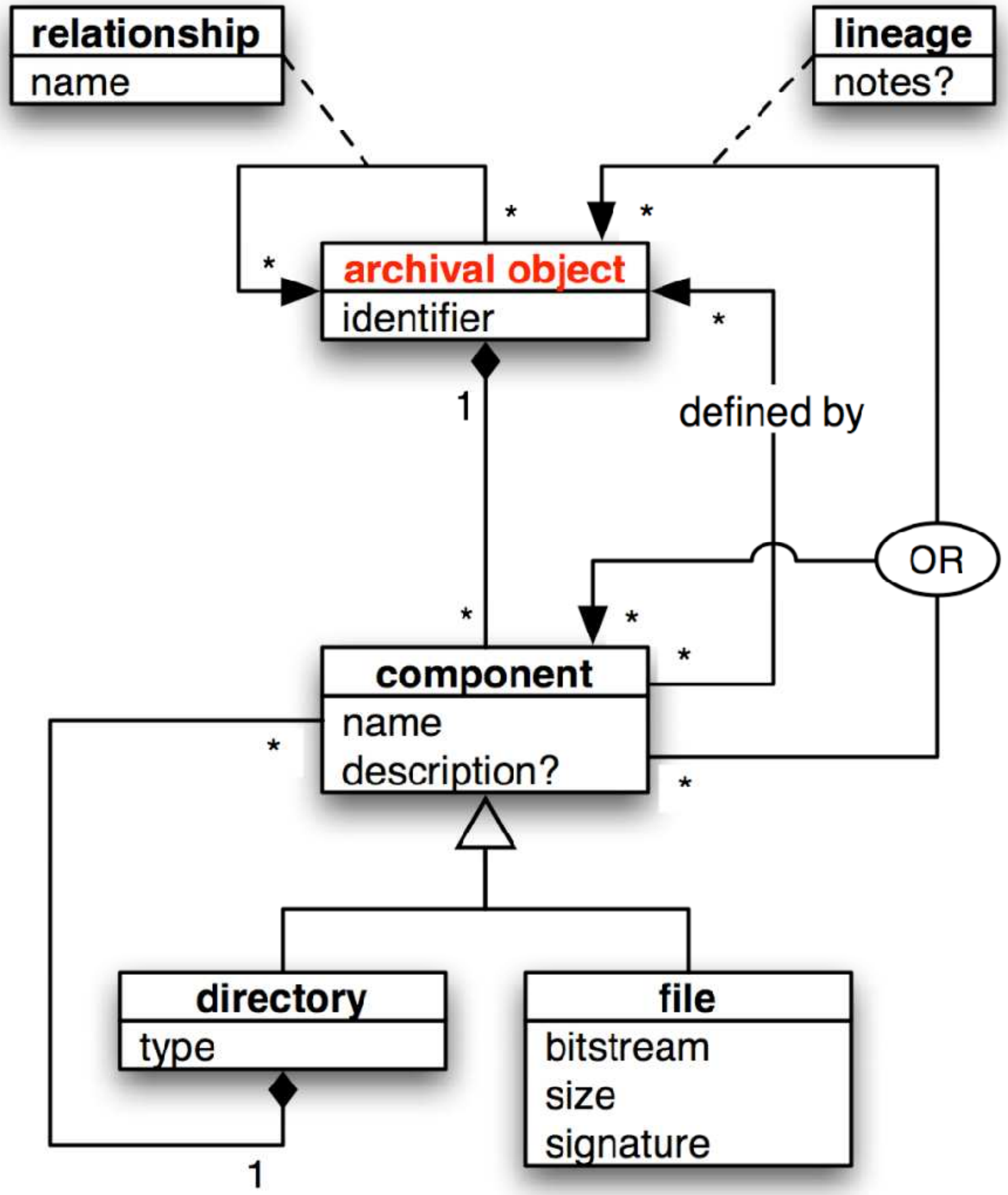
MathArc – Ensuring Access to Mathematics over Time

- MathArc - a collaborative project of Cornell University Library and Göttingen State and University Library aims at the development and maintenance of reliable digital archives of Serial Literature in the field of Mathematics
- In MathArc every AIP, SIP or DIP is defined as an “asset”- a digital representation of content which is going to be archived.
- In MathArc the following PREMIS schemas are used:
- object: contains information about an asset or a single file
- event: contains information about migration (on file level) or information about transferring or deleting a whole asset in one of the partner's archive.

The National Geospatial Digital Archive (NGDA)

- NGDA participants include the Map & Imagery Laboratory at the University of California at Santa Barbara and Branner Earth Sciences Library at Stanford University.
- The overarching goal is preserve geospatial data on a national scale and make it available to future generations.
- The architecture is based on a OAIS Reference Model and the PREMIS data model that defines a uniform representation of all information in the archive, paired with a storage API that abstracts the storage subsystem
- The storage subsystem is assumed to provide reliable, long-term storage through redundancy and active monitoring. A suite of components built on top of the data model and storage API provides ingest and access functionality

NGDA Architecture



Case Studies - Comparison

- Most repositories do comply with the Object identification and some kind of event recording with reference to preservation activity.
- The 'agent' is either not well defined or sparsely adhered to.
- Rights details are again not well attended to.
- A data archive (such as NGDA) requires deeper level of detail than information resource repositories

Conclusion

- OAIS reference model serves as a high level framework for the interaction between different elements and the PREMIS Data Model consisting of three semantic units – entities, relationships, and properties serves as the next level of building block for the digital preservation initiatives.
- Given that each type of information resource and data, whether it be text or music or Imagery or genomic data or geospatial data, brings to the preservation problem its own complications and special requirements, one of the key challenges is to develop the next level of domain specific/mission specific data models and data dictionaries

Thank you