Digital Preservation and Archiving

Gunter Vasold & Johannes Stigler

Center for Information Modelling – Austrian Centre for Digital Humanities
University Graz

Borås, February 2015

Timetable

09:00-09:45	Thinking sustainably in using computers (Gunter)
09:45-10:30	Digital preservation: Basics (Johannes)
10:30-10:45	Coffee break
10:45-12:00	Digital preservation: Theory and Practice (Johannes)
12:00-13:30	Lunch break
13:30-14:30	Revision control systems in creating digital editions (Gunter)
14:30-14:45	Coffee break
14:45-16:00	Digital preservation: Research and Practice (Elena)

Agenda

1. Basics
2. OAIS – Open Archival Information System Reference Model
   - Basics
   - Data Model
   - Function Model
3. PREMIS – Preservation Metadata: Implementation Strategies
4. FEDORA – An Examplary Implementation of OAIS

What is Digital Preservation?

A first approach

- Digital preservation begins with generating data
- Digital archiving is more than bitstream preservation
- Long-term preservation includes long-term availability
- Not only a technical, but an institutional infrastructure is required
- Digital preservation needs data curation

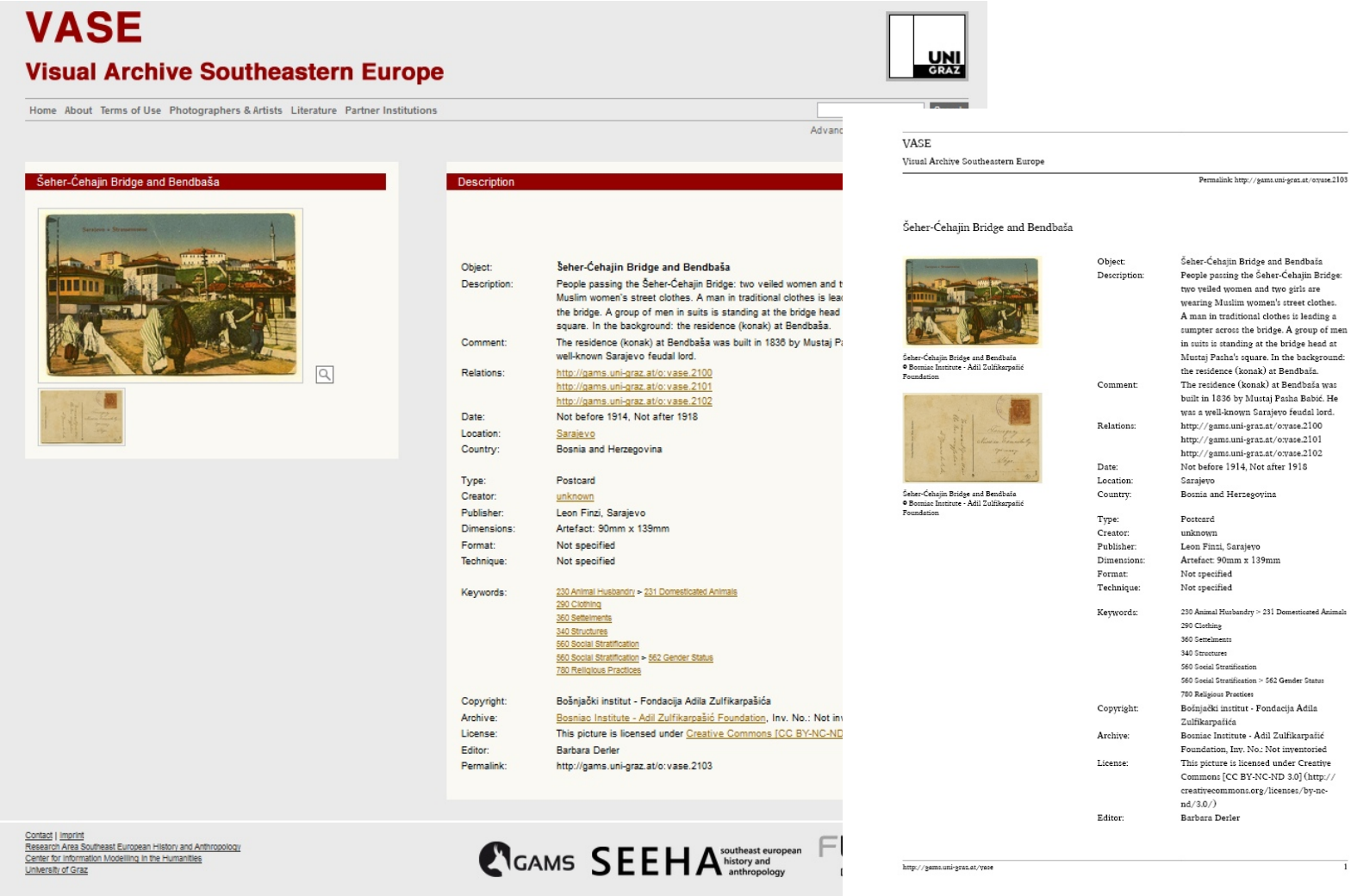
What is a Digital Archive?

A Repository provides ...

- direct access to individual information objects or information collections via a Persistent Identifier
- support of standardized protocols for data exchange and metadata harvesting, e.g. OAI-PMH etc.
- support of standardized programming languages for representation workflows
- version management strategies for datastreams
- a converter to generate long-term stable archive formats
- *knowledge* about life cycles of all objects
- a rights management system to prevent unauthorized access to informations objects

What is a Digital Archive?

Information objects ...



What is a Digital Archive?

Information objects ...

- Shows the object content in different views
  - http://glossa.uni-graz.at/o:dixit.505
- Shows the object content in different views
  - http://glossa.uni-graz.at/o:dixit.505/sdef:dfgMETS/get
- Shows the object content in different views
  - http://glossa.uni-graz.at/o:dixit.505/sdef:TEI/get
- Shows the datastreams list for the object
  - http://glossa.uni-graz.at/archive/objects/o:dixit.505/datastreams
- Shows the methods list for the object
  - http://glossa.uni-graz.at/archive/objects/o:dixit.505/methods

What is a Digital Archive?

Persistent Identifier

- A persistent identifier (PI) is a long-lasting reference to a digital object
- Persistent identifier conventions and systems:
  - Handle Systems
  - Digital Object Identifiers (DOIs)
  - Archival Resource Keys (ARKs)
  - Persistent Uniform Resource Locators (PURLs)
  - Uniform Resource Names (URNs)
  - Extensible Resource Identifiers (XRIs)

What is a Digital Archive?

Handle System

- The Handle System is a technology specification for assigning, managing, and resolving persistent identifiers for digital objects
- The system is designed to be scalable to very large numbers of entities
- The Handle System web site provides a series of implementation tools
  - Implementation of the Handle System consists of Local Handle Services that store specific handles
  - The Global Handle Registry is a unique Local Handle Service which stores information on the prefixes (also known as naming authorities) within the Handle System

What is a Digital Archive?

Handle System

An id string ...

- is not based on any changeable attributes of the entity (location, ownership, or any other attribute that may change without changing the reference identity)
- is opaque (preferably a plain number: a well known pattern invites assumptions that may be misleading, and meaningful semantics may not translate across languages and may cause trademark conflicts)
- is unique within the system (to avoid collisions and referential uncertainty)
- http://hdl.handle.net/*handle_prefix*/*identifier_string*
- hdl:*handle_prefix*/*identifier_string* (with installed browser plugin)
- e.g. `http://hdl.handle.net/11471/505.20.779`

What is a Digital Archive?

Standardized protocols for data exchange and metadata harvesting

- The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability
- Data Providers are repositories that expose structured metadata via OAI-PMH
- Service Providers then make OAI-PMH service request to harvest that metadata
- OAI-PMH is a set of six verbs or services that are invoked within HTTP
  - Archival Metadata
    - `http://gams.uni-graz.at/oaiprovider?verb=Identify`
    - `http://gams.uni-graz.at/oaiprovider?verb=ListMetadataFormats`
  - Harvesting verbs
    - `http://gams.uni-graz.at/oaiprovider/?verb=GetRecord&metadataPrefix=oai_europeana&identifier=hdl:11471/505.20.779`
    - `http://gams.uni-graz.at/oaiprovider?verb=ListIdentifiers&metadataPrefix=oai_europeana`
    - `http://gams.uni-graz.at/oaiprovider?verb=ListRecords&metadataPrefix=oai_europeana`
    - `http://gams.uni-graz.at/oaiprovider?verb=ListSets`

What is a Digital Archive?

Excursus: Standardized representation workflows – Extensible Stylesheet Language

- XSL Transformation (XSLT): An XML programming language for transforming XML documents
- XSL Formatting Objects (XSL-FO): An XML language for specifying the visual formatting of an XML document
- XML Path Language (XPath): A non-XML query language used by XSLT, and also available for use in non-XSLT contexts, for addressing the parts of an XML document
- Parser (Saxon, Xerces) are the runtime environment for XSLT

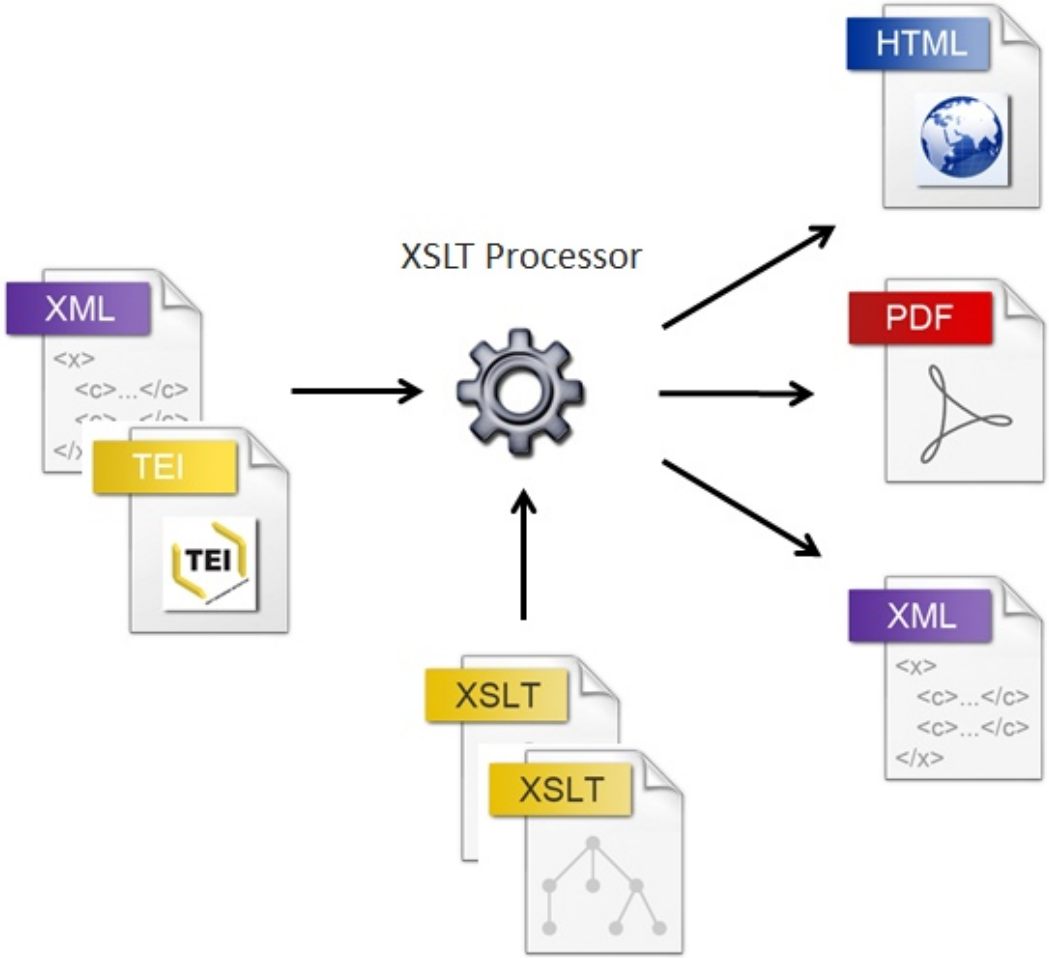
What is a Digital Archive?

Excursus: Standardized programming language – XSLT Transformation

- The XSLT processor takes one or more XML source documents, plus one or more XSLT stylesheets, and processes them to produce an output document
- In contrast to widely-implemented imperative programming languages like C, XSLT is declarative
- An XSLT program consists of template rules, defining input output scenarios
- Template rules only define how to handle a node matching a particular XPath-like pattern
- The basic processing paradigm is pattern matching

What is a Digital Archive?

Excursus: Standardized programming language – XSLT Transformation



What is a Digital Archive?

Excursus: XSLT Transformation – A XML document

```
<menu-card>
    <item>Mejillones a la Copita</item>
    <item>Bistec al Pimiento</item>
    <item>Gambas al Pil Pil</item>
    <item>Bleak Roe</item>
    <item>Bruschetta de Salmon</item>
    <item>Vol au Vent</item>
    <item>Pata Negra Bellota</item>
    <item>Queso de Cabra</item>
    <item>Mejillones Gratinados</item>
    <item>...</item>
</menu-card>
```

What is a Digital Archive?

Excursus: XSLT Transformation – A XSLT Stylesheet

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
    <xsl:template match="menu-card">
        <html>
            <head><title/></head>
            <body>
                <h1>LaCopita Menu Card</h1>
                <ol>
                    <xsl:apply-templates select="item"/>
                </ol>
            </body>
        </html>
    </xsl:template>

    <xsl:template match="item">
        <li>
            <xsl:value-of select="."/>
        </li>
    </xsl:template>

</xsl:stylesheet>
```

What is a Digital Archive?

Excursus: XSLT Transformation – A HTML result

```
<html>
   <head>
      <title/>
   </head>
   <body>
      <h1>LaCopita Menu Card</h1>
      <ol>
         <li>Mejillones a la Copita</li>
         <li>Bistec al Pimiento</li>
         <li>Gambas al Pil Pil</li>
         <li>Bleak Roe</li>
         <li>Bruschetta de Salmon</li>
         <li>Vol au Vent</li>
         <li>Pata Negra Bellota</li>
         <li>Queso de Cabra</li>
         <li>Mejillones Gratinados</li>
         <li>...</li>
      </ol>
   </body>
</html>
```

OAIS – Open Archival Information System Reference Model

- In 2000 the Research Libraries Group (RLG) and Online Computer Library Center (OCLC) discussed how both organizations could build an infrastructure for purposes of archiving digital objects.
- The resulting model guides you through building an archival information system

OAIS – Open Archival Information System Reference Model

Defines a long-term preservation terminology for

- Architectures and Operations
- Preservation strategies and techniques
- A data model

Provides ...

- a framework for an increased awareness of concepts needed for long term preservation
- a framework for describing and comparing architectures and operations of archives
- a basis for comparing data models of digital information preserved by archives

OAIS – Open Archival Information System Reference Model

Roles

- Producer – A data provider
- Administrator – A data manager
- Consumer – A data retriever

Important functions

- Ingest – Submit data to system
- Store – Preserve data in system
- Access – Retrieve data from system

OAIS – Open Archival Information System Reference Model

Roles and their responsibilities

- Producer
  - Person(s) or client system(s), who provide the information to be preserved
  - Ingest digital resource to system
- Administrator
  - Person(s) or client system(s), who manage and set the overall policy of the OAIS
  - Monitor, verify digital resource, do preservation planning, migrate digital resource, and etc.
- Consumer
  - Person(s) or client system(s), who interact with the OAIS system and services
  - Search and access digital resource in repository

OAIS – Open Archival Information System Reference Model

Data Model

- Preserved data in the system needs to be wrapped in a package
- Owing to the three important functions of OAIS (Ingest, Store, and Access), packages of preserved data are transformed into three types
  - Submission Information Package
  - Archival Information Package
  - Dissemination Information Package

OAIS – Open Archival Information System Reference Model

Data Model

- SIP – Submission Information Package
  - A form of package that is suitable for ingest to the system by the producer
  - Majorly, SIP contains Content Info and PDI
- AIP – Archival Information Package
  - A form of package that is suitable to being stored in the system
- DIP – Dissemination Information Package
  - A form of package that is suitable for dissemination to consumer
  - AIP is transformed to DIP for sharing purpose

OAIS – Open Archival Information System Reference Model

PDI – Preservation Description Information

- What is needed to preserve the Content Information
  - Provenance
    - To record why, where and how the digital resource was born
    - Including software and environment that created it
  - Context
    - To inform about original or source of content
    - To inform about history of change
    - To inform about migration process
  - Reference
    - Identifier that links to something outside system or real world resource; such as ISBN
  - Fixity
    - To provide necessary information to access and verify digital resource, e.g.Checksum, MD5

OAIS – Open Archival Information System Reference Model

Administrator

- Negotiate Submission Agreement
  - Discuss submission agreement with producer
- Manage System Configuration
  - Configure and control changes which affect system engineering of archival system
- Physical Access Control
  - Authorize access to resources
- Establish Standards and Policies
  - Manage standards and policies in order to approve migration and replication processes
- Audit Submission
  - Verify that AIP and SIP is following specification and agreement
- Activate Requests
  - To check the request of consumer is correct, then submit the request to Access
- Customer Service
  - Provide functions to manage user account

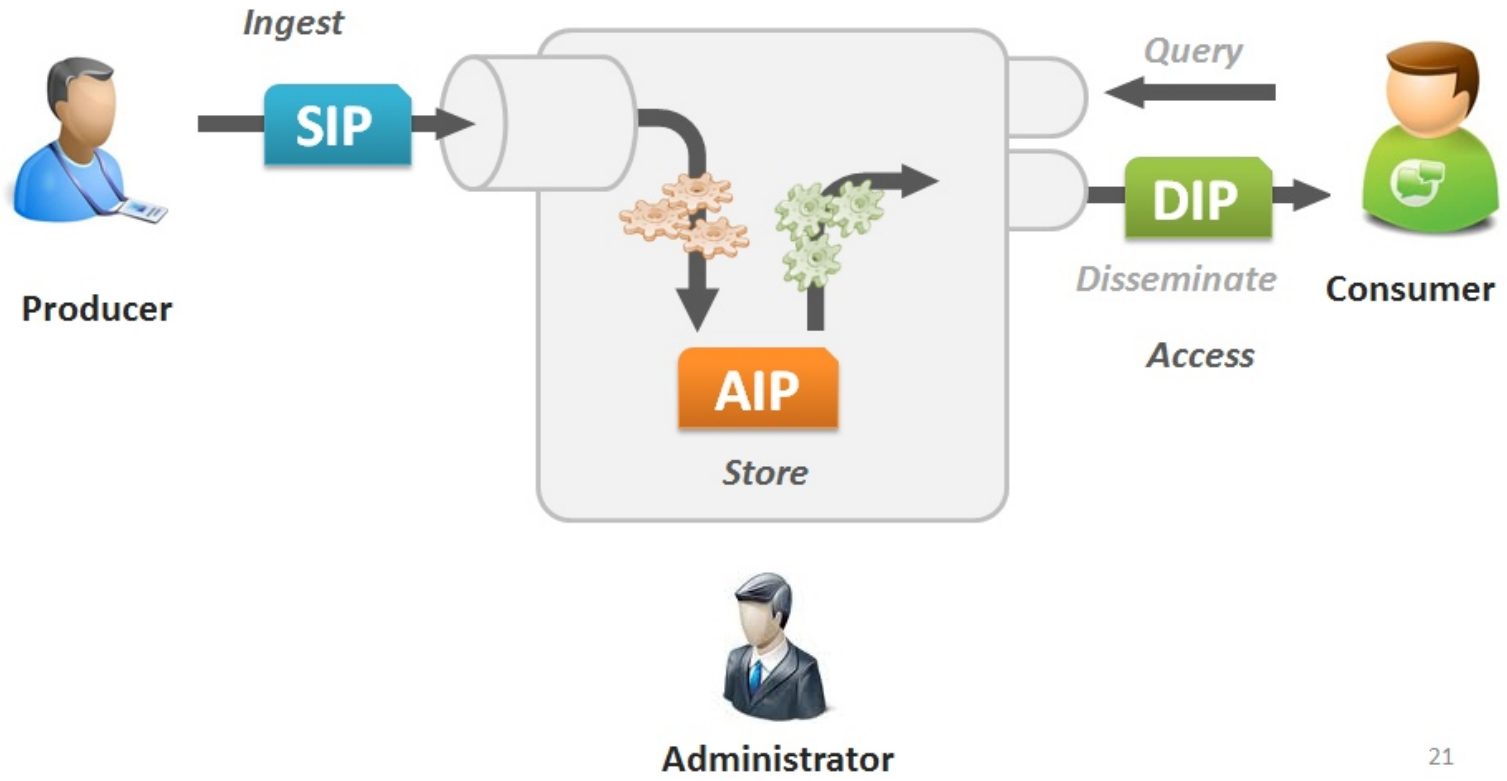
OAIS – Open Archival Information System Reference Model

Administrator (cont.) – Preservation Planning

- Monitor environment of OAIS and provide recommendations
- Monitor Designated Community
  - Allow consumer and producer to track change of available technologies
- Monitor Technology
  - Report change of software and hardware contributing to preservation process
- Develop Preservation Strategies and Standards
  - Develop and recommend strategies and standards for future change of technology
- Develop Packaging Designs and Migration Plans
  - Customize SIP and AIP template for migration goal

OAIS – Open Archival Information System Reference Model

3 functions, 3 information packages, and 3 roles



OAIS – Open Archival Information System Reference Model

Ingest

- Accept SIPs from Producers
- Verify SIPs that user submits
- Generate AIPs for archive storage

- Receive Submission
  - Upload SIP package from producer by electronic transfer such as FTP
- Quality Assurance
  - Validate transmission (e.g. checksum) error SIP package and log a result
- Generate AIP
  - Transform SIP to AIP and report result
- Coordinate Update
  - Provide a single access point (add, modify, remove, get) to storage area

OAIS – Open Archival Information System Reference Model

Store

- Receive Data
  - Receive AIP from Ingest to permanent storage
- Manage Storage Hierarchy
  - Provide administration functions for storage media
- Replace Media
  - Support functions of migration from a media to another media
- Error Checking
  - Check and notification error from data in storage area
- Disaster Recovery
  - Provide mechanism for replicating digital content to backup

OAIS – Open Archival Information System Reference Model

Access

- Coordinate Access Activities
  - Provide single user interface for features like browse, search and access
- Generate DIP
  - Generate DIP from AIP
- Generate DIP
  - Handle response from query and access and delivery to consumer
  - Report access activities to administrator

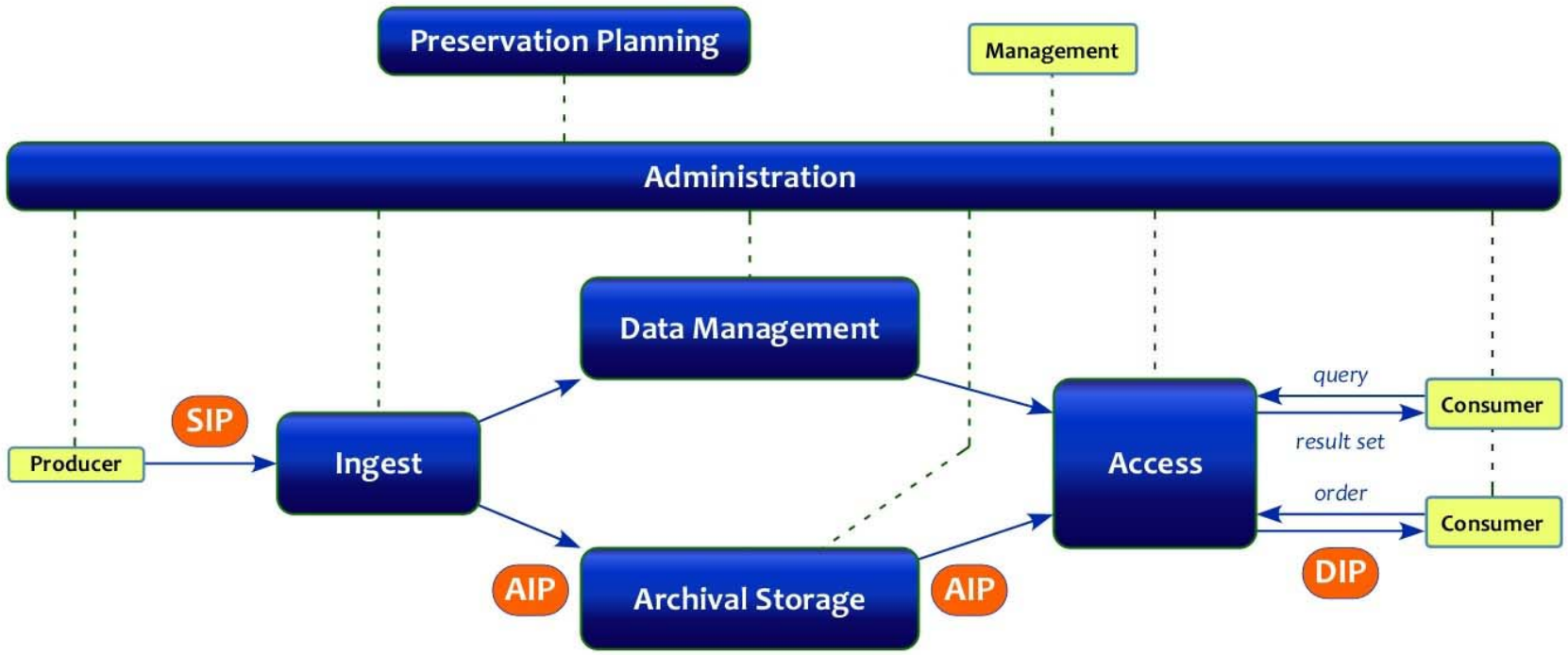
OAIS – Open Archival Information System Reference Model

Summary

- Producer
  - Ingest package to system
    - System stores AIP in archival storage
    - System stores descriptive metadata in data management
- Consumer
  - Queries data via Access
    - Query from descriptive metadata from data management
  - Retrieve data via access
    - Get data from archival storage
- Administrator
  - Manages and monitors every process in system

OAIS – Open Archival Information System Reference Model

Big picture of all functions and process of packages



OAIS Reference Model

OAIS – Open Archival Information System Reference Model

Useful Resources

- OAIS specification
  - http://public.ccsds.org/publications/archive/650x0m2.pdf
- OAIS primer
  - http://www.dpconline.org/docs/lavoie_OAIS.pdf

PREMIS – Preservation Metadata: Implementation Strategies

PREMIS is an Information Model

- Focus is on the preservation of digital objects
- The information a repository uses to support the digital preservation process
- Things that most working preservation repositories need to know to support digital preservation functions
- Data dictionary defines a set of semantic units

PREMIS – Preservation Metadata: Implementation Strategies

Different Types of Metadata

- Descriptive
  - Supports the identification and discovery of a resource
- Administrative
  - Supports the management and tracking of a resource
- Structural
  - Defines the arrangement and composition of a resource
- Preservation
  - Supports activities intended to ensure the long term usability of a resource

PREMIS – Preservation Metadata: Implementation Strategies

What is out of Scope?

- Descriptive Metadata
  - Many existing standards support this
- File Format specific metadata
  - Metadata that pertains to only one file format or class of formats
- Implementation metadata
  - Metadata that describes specific policies and practices of an individual repository
- Detailed media and hardware information
  - Left to other communities to define
  - Technical environment metadata is in scope

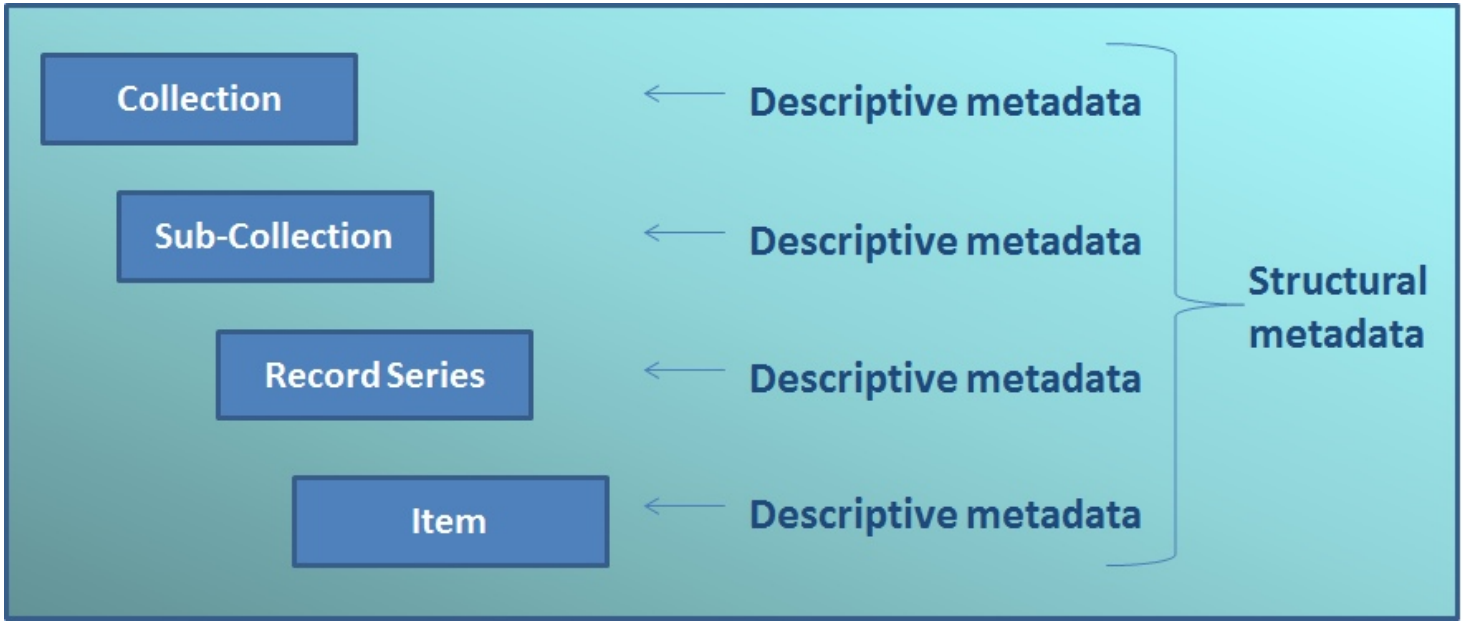
PREMIS – Preservation Metadata: Implementation Strategies

Usage

- Repository Design
  - Provides guidelines on what information should be obtained and maintained by a preservation repository
- Repository evaluation
  - Provides a checklist to determine effective preservation management of digital objects
- Exchange of objects between repositories
  - Provides a common set of data elements that can be understood by the provider and consumer repositories

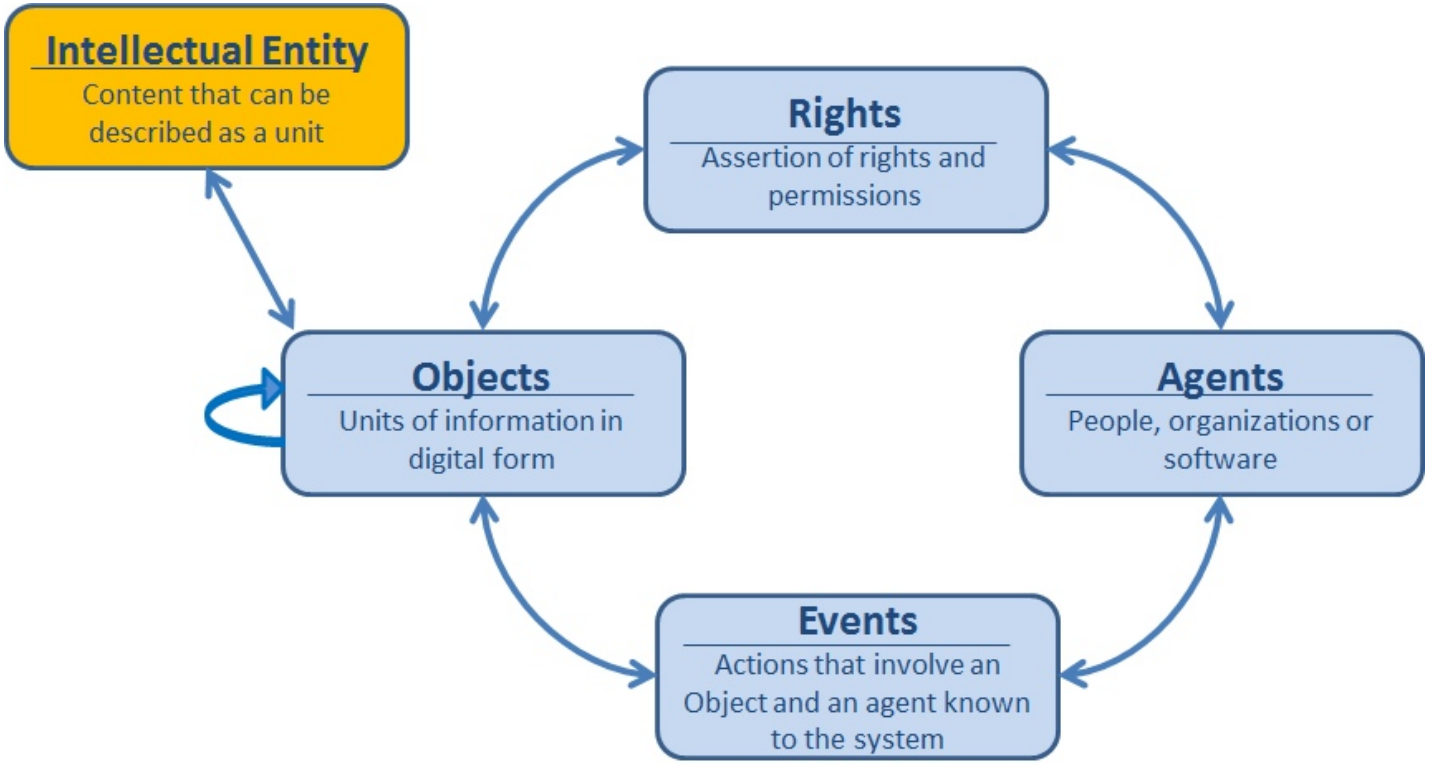
PREMIS – Preservation Metadata: Implementation Strategies

Always has intellectual entities



PREMIS – Preservation Metadata: Implementation Strategies

Data Model



PREMIS – Preservation Metadata: Implementation Strategies

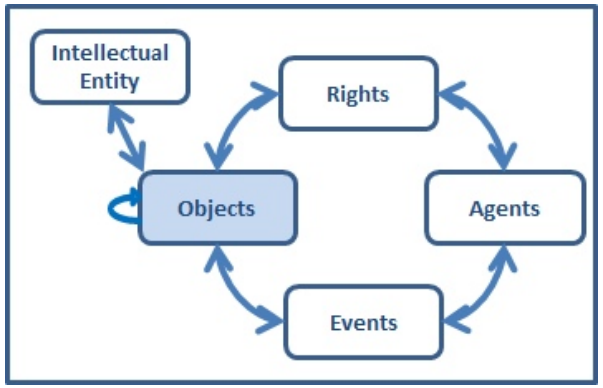
Semantic Units

- Semantic Units
  - Convey a piece of information / knowledge
  - Do not specify how they should be represented in a particular system (e.g. to metadata elements)
  - Should be exportable to other systems
  - May have a direct mapping to metadata elements in an XML schema
- Repository evaluation
  - Provides a checklist to determine effective preservation management of digital objects
- Containers and sub units
  - Some semantic units are defined as container
  - Facilitiates a hierarchical structure to the data dictionary

PREMIS – Preservation Metadata: Implementation Strategies

Objects

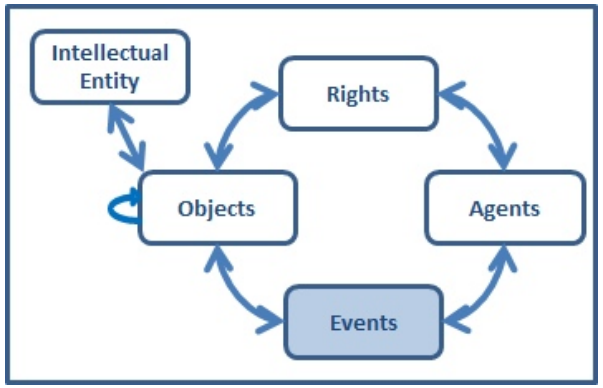
- Identifier
- Category (Representation, File, Bitsteam)
- Preservation level
- Characteristics: Fixity, Size (bytes), Format, Creating application
- Original name
- Environment: ...,Software, Hardware, ...
- Signature Information
- Linked events
- Linked intellectual entity
- Linked rights statement



PREMIS – Preservation Metadata: Implementation Strategies

Events

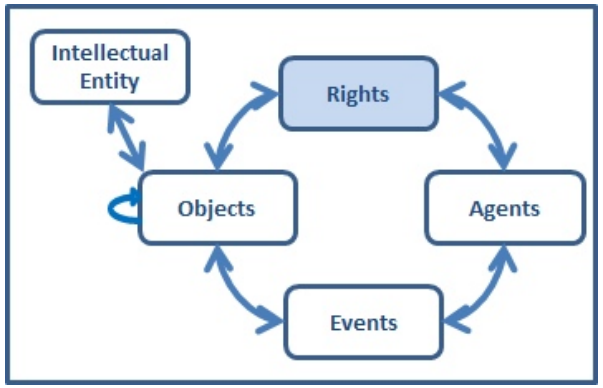
- Identifier
- Type
- Date Time
- Outcome Information
- Linking Agent Identifier
- Linking Object Identifier



PREMIS – Preservation Metadata: Implementation Strategies

Rights

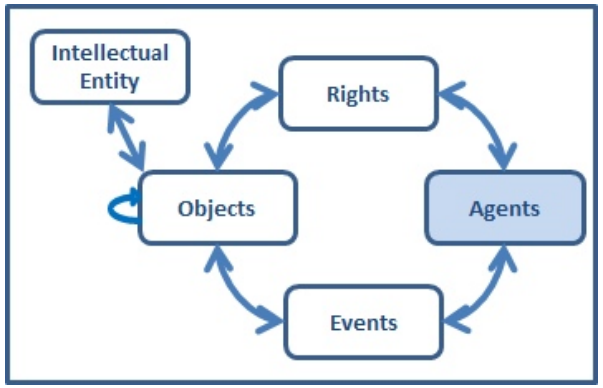
- Identifier
- Copyright Information
- Licence Information
- Rights granted
- Linking Agent Identifier
- Linking Object Identifier



PREMIS – Preservation Metadata: Implementation Strategies

Agents

- Identifier
- Name
- Type



PREMIS – Preservation Metadata: Implementation Strategies

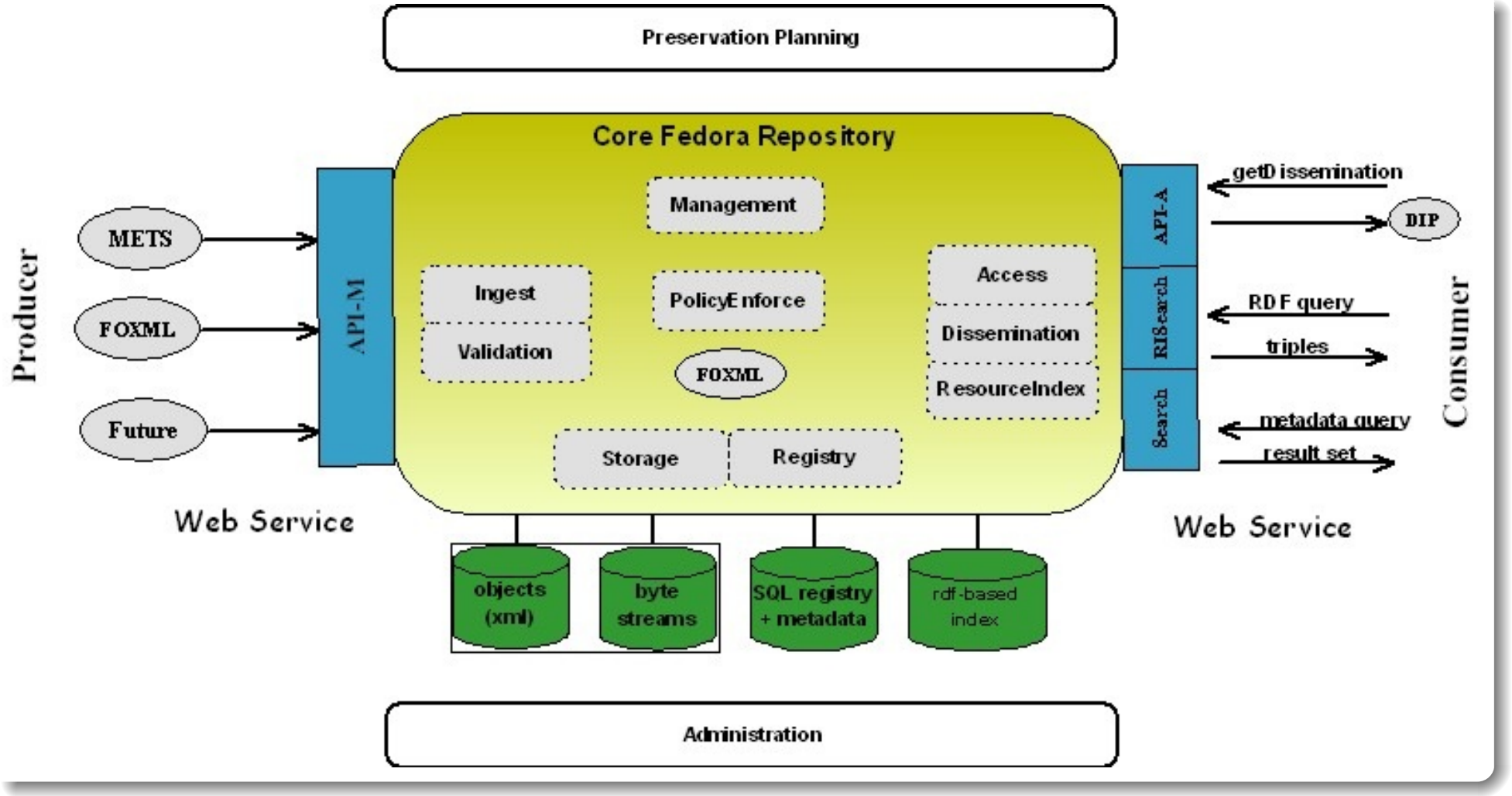
Useful Resources

- PREMIS specification
  - http://www.loc.gov/standards/premis/
- PREMIS primer
  - http://www.loc.gov/standards/premis/understanding-premis.pdf

FEDORA – An Example Implementation of OAIS

- Flexible Extensible Digital Object Repository Architecture
- A system that provides a digital content repository for a wide variety of users
  - E.g. institutional repository, digital archive, content management system, scholarly publishing enterprises, and digital library
- Fedora as OAIS
  - Is based on OAIS data model, function model and architecture models
  - End client (e.g. CIRILO) can access repository functions via web services

FEDORA – Flexible Extensible Digital Object Repository Architecture



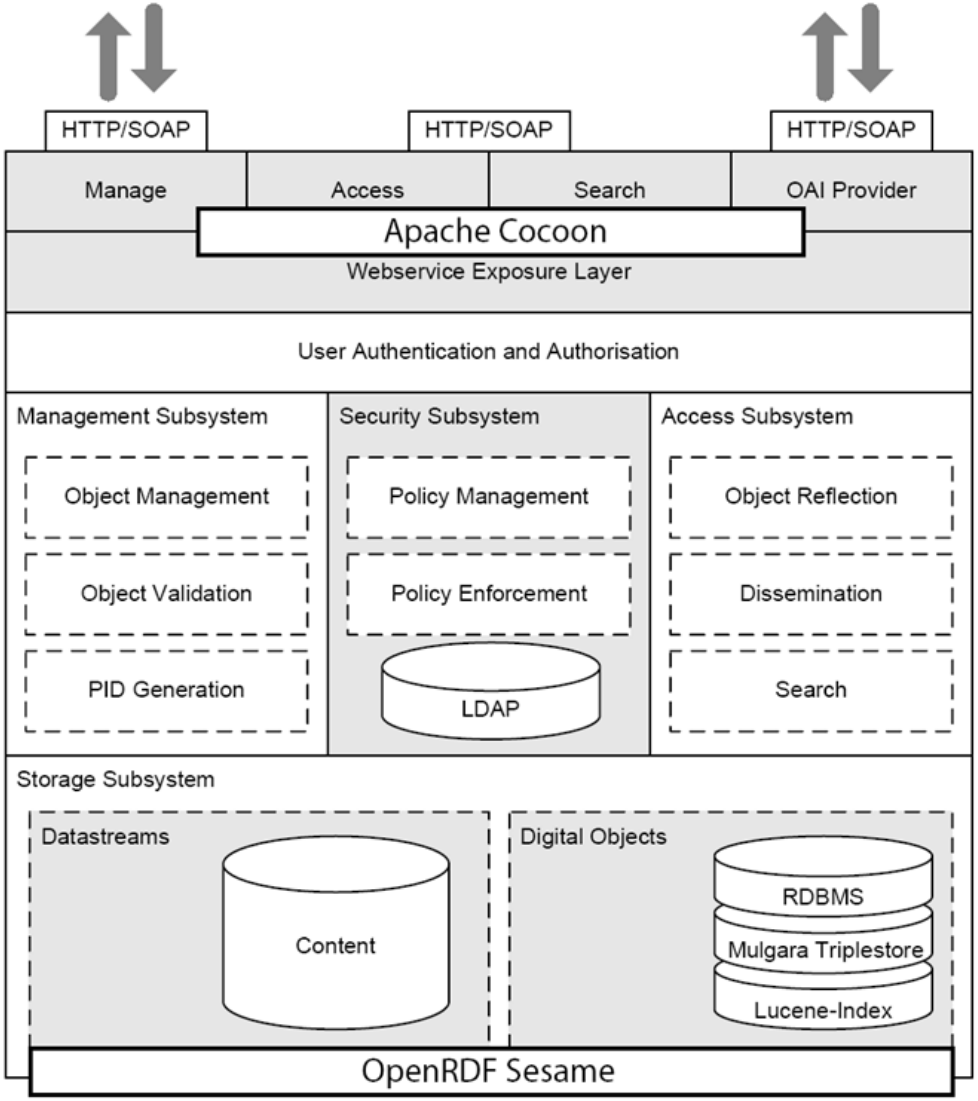
FEDORA – Engineering Figures

- Scalable, persistent storage infrastructure for content and metadata
- Webservice based (SOAP), platform independent, distributed system architecture
- Supports standardized protocols for data exchange, e.g. OAI-PMH etc.
- Extensible digital object models with associate services
- Provides dynamic transformation and multiple views of content and metadata
- Includes version management strategies for data streams
- Definition of access rights with eXtensible Access Control Markup Language

GAMS – Engineering Figures

- A FEDORA based repository with other embedded open source projects: OpenRDF Sesame and IIP Image Server
- A multitude of web services for content dissemination (e.g. Apache Cocoon)
- Handle System based resolution of persistent identifiers
- A Client named CIRILO for mass ingest and data curation supporting a set of content models
- Workflows and Tools for generating data
- CIRILO and an „archive-in-a-box“ solution as an Austrian contribution to DARIAH `https://github.com/acdh/cirilo`

FEDORA – Flexible Extensible Digital Object Repository Architecture



FEDORA – Content Models

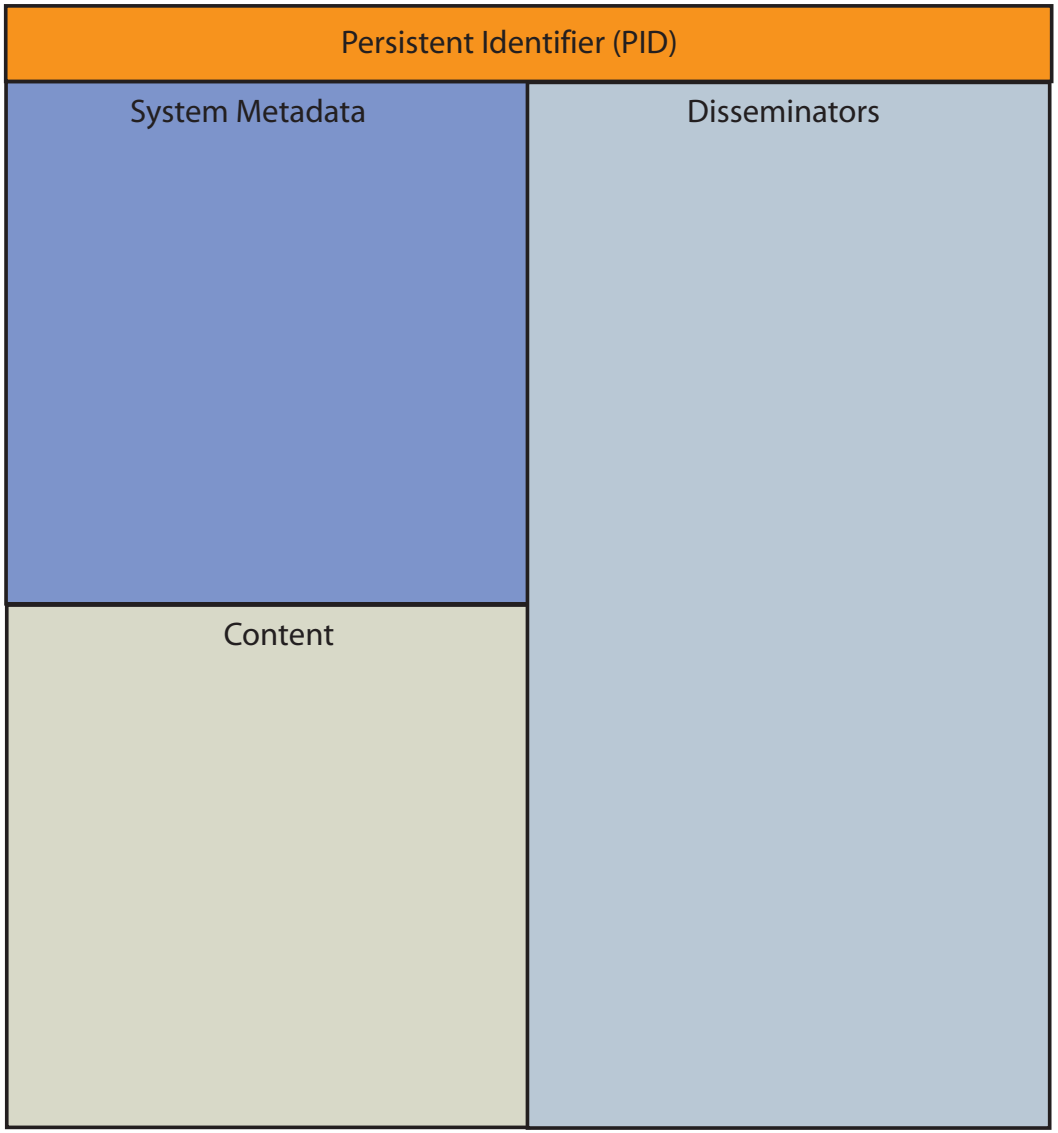
What is a Content Model?

- A structural definition for a „type“ of object (e.g. article, book, learning object, podcast, ontology etc.)
- A pattern of datastreams (number and type)
- A pattern of datastreams and their disseminators
- A set of rules for creating a digital object
- A set of constraints on a digital object

What use are Content Models?

- Object Typing
  - Group identity for different kinds of objects
  - Facilitates discovery via query/search
- Object Creation
  - Templates for user interfaces enabling object creation
  - Drive workflows/creation of „batches“ of similar objects
- Object Validation
  - At ingest, check that object conforms to a model
  - At modification, make sure changes dont't break conformance to model

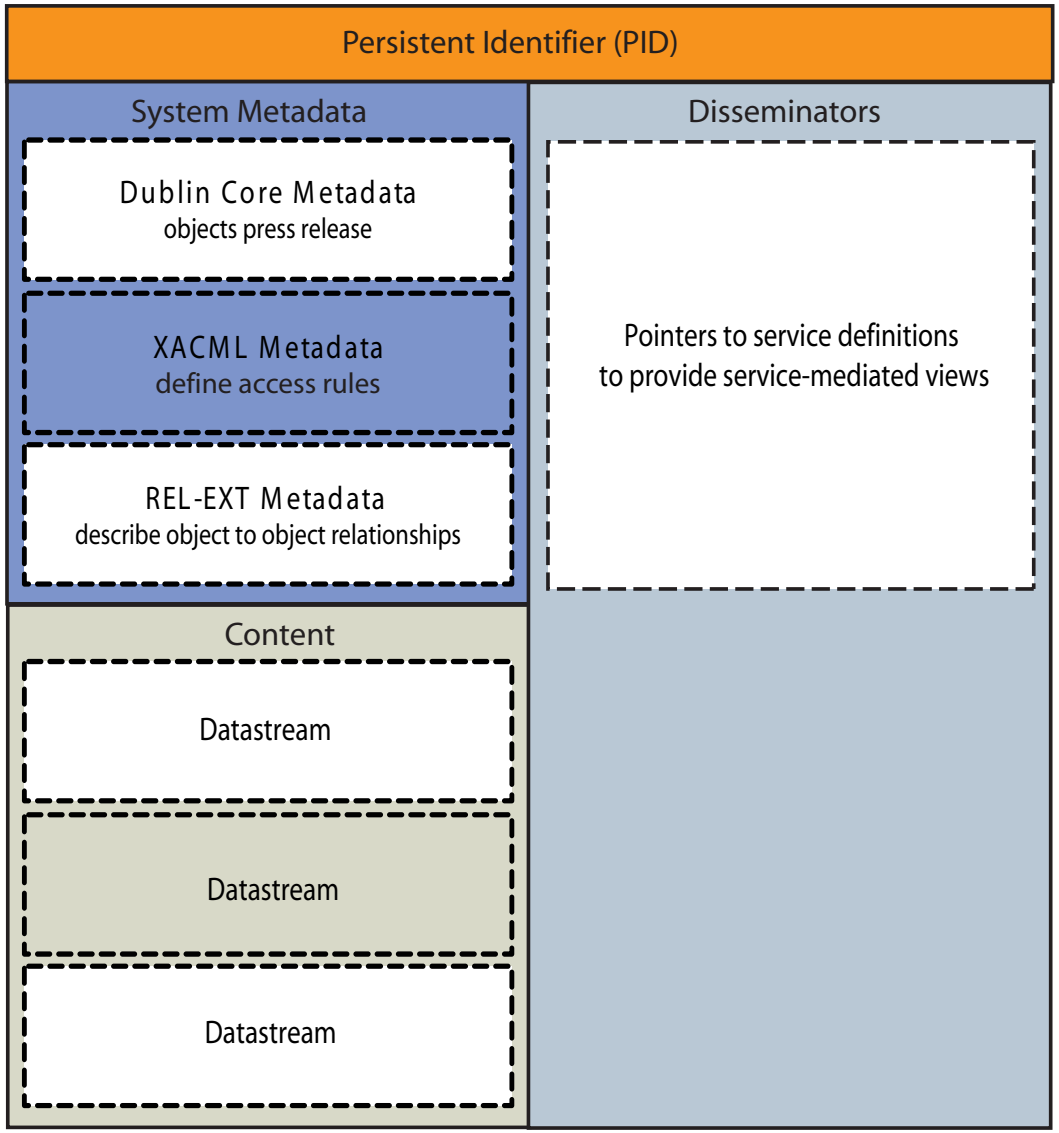
FEDORA – A Structural Model of a Content Model



FEDORA – A Structural Model of a Content Model

FEDORA – A Structural Model of a Content Model

FEDORA – A Structural Model of a Content Model



FEDORA – Content Models

TEI content model

- System Datastreams
  - DC – Object press release
  - STYLESHEET – XSLT Stylesheet to transform object content to HTML
  - FO_STYLESHEET – XSLT Stylesheet to transform object content to XSL:FO
  - RELS-EXT – Describe object to object relations
- Content Datastreams
  - TEI_SOURCE – TEI source
  - BIBTEX - BibTeX source
  - DC_MAPPING – TEI to DC mapping rules
  - TORDF – XSLT Stylesheet to create RDF-Triples from TEI_SOURCE
  - REPLACEMENT_RULESET – A set of regex replacement
  - THUMBNAIL - Thumbnail that repesents the object in views
  - QR - OR Code image
- Disseminators
  - sdef:TEI/get
  - sdef:TEI/getPDF
  - sdef:Object/get
  - sdef:Object/getDC
  - sdef:BibTeX/get
  - sdef:BibTeX/getRIS
  - sdef:BibTeX/getENDNOTE

FEDORA – Content Models

Workflows during the creation or updating of a TEI object

- Rule based extraction of Dublin Core metadata
- Rule based extraction of semantic constructs
- Resolution of ontology concepts
- Resolution of place names against geonames.org
- Execution of TEI customization
- Creation of context objects
- Uploading of images

GAMS – An Overview of Default Content Models

- cirilo:Context – Aggregate and display ordered lists of objects
- cirilo:TEI – Encapsulate a TEI file and its semantic representation
- cirilo:dfgMETS – Store and display DFG-viewer conform METS Files
- cirilo:Ontology – Navigate through hierarchies of concepts
- cirilo:Query – Do a multicategory search
- cirilo:BibTeX – Create a bibliography in a specific style
- ...

Ingest variants provided by CIRILO

Ingest ...

- from filesystem
- from eXist databases
- from EXCEL spreadsheets

FEDORA and GAMS

Useful Resources

- FEDORA
  - http://fedora-commons.org
- GAMS
  - http://gams.uni-graz.at
  - http://rdf4j.org
  - http://iipimage.sourceforge.net