From Heterogeneous Information Spaces to Virtual Documents

L. Candela D. Castelli P. Pagano, M. Simi

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - CNR
Via G. Moruzzi, 1 - 56124 PISA - Italy
{candela|castelli|pagano|simi}@isti.cnr.it

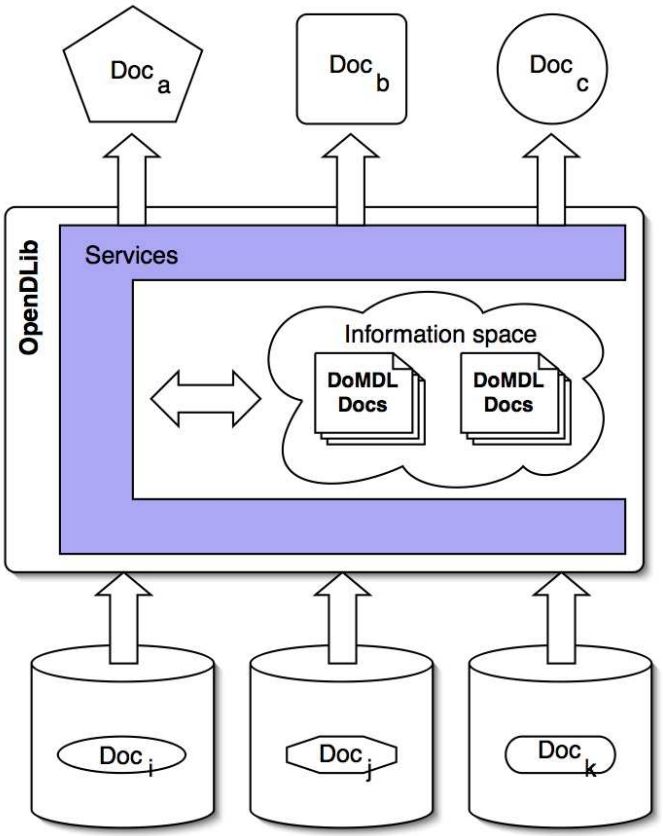
The 8th International Conference on Asian Digital Libraries, ICADL 2005

Outline

DLs and DLMSs Nowadays

Definition

Digital Library Management Systems are complex systems whose main role is to *mediate* between *content providers* and *content consumers* in order to fulfill information and functionality needs of the DL users.



Heterogeneous Information Sources

DLMSs must support storage and management of documents collected from heterogeneous information sources which differ for

- structure, format, media, physical representation of documents
- metadata formats
- access policies

Community Specific Views

DLMS must supports end-user functions for search, retrieve, access and manipulation on community specific documents

- e.g. a journal which contains articles, a text and a set of images

community specific documents do not necessarily correspond to those submitted to the DL but are virtual documents created by reusing and/or processing real documents or part of them

DoMDL Overview

Document Model for Digital Library (DoMDL) is the document model designed at ISTI to represent

- multi-edition, structured, multimedia documents
- multiple manifestation formats
- multiple metadata descriptions in different formats
- linking relationships with other documents and parts of them
- information that services exchanges and operates on

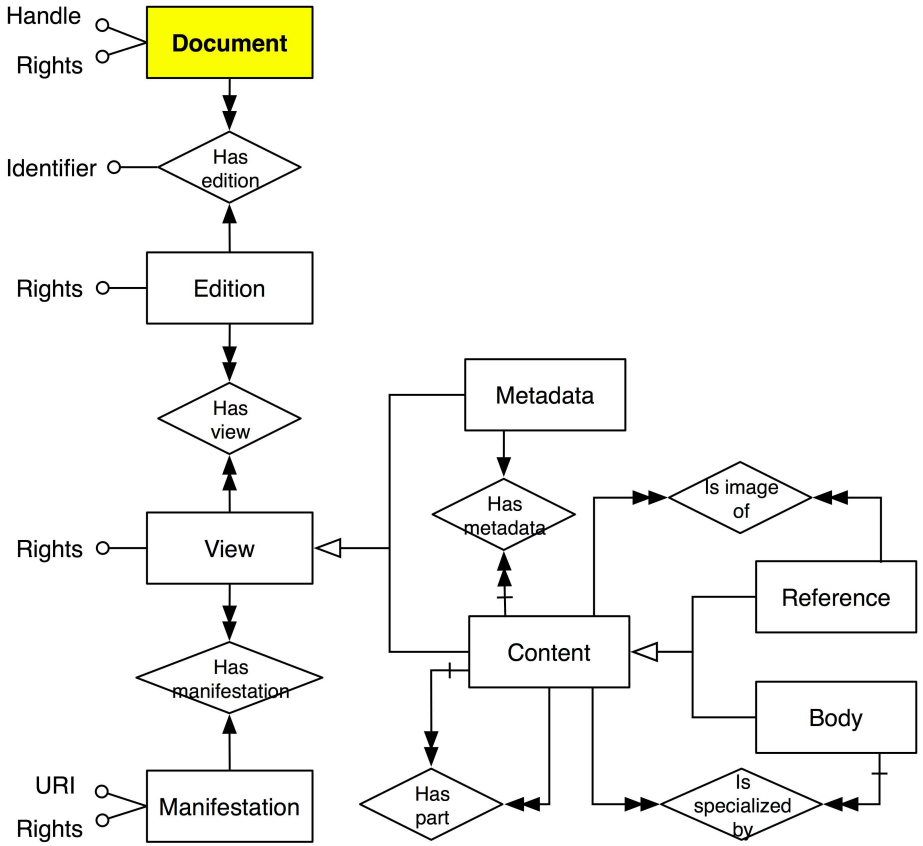
The Document Entity

Definition

A Document entity represents a distinct intellectual creation.

Example

The book *Digital Libraries and Electronic Publishing* by W. Arms, the lecture *Introduction to Mixed Media Digital Libraries* by C. Lagoze.



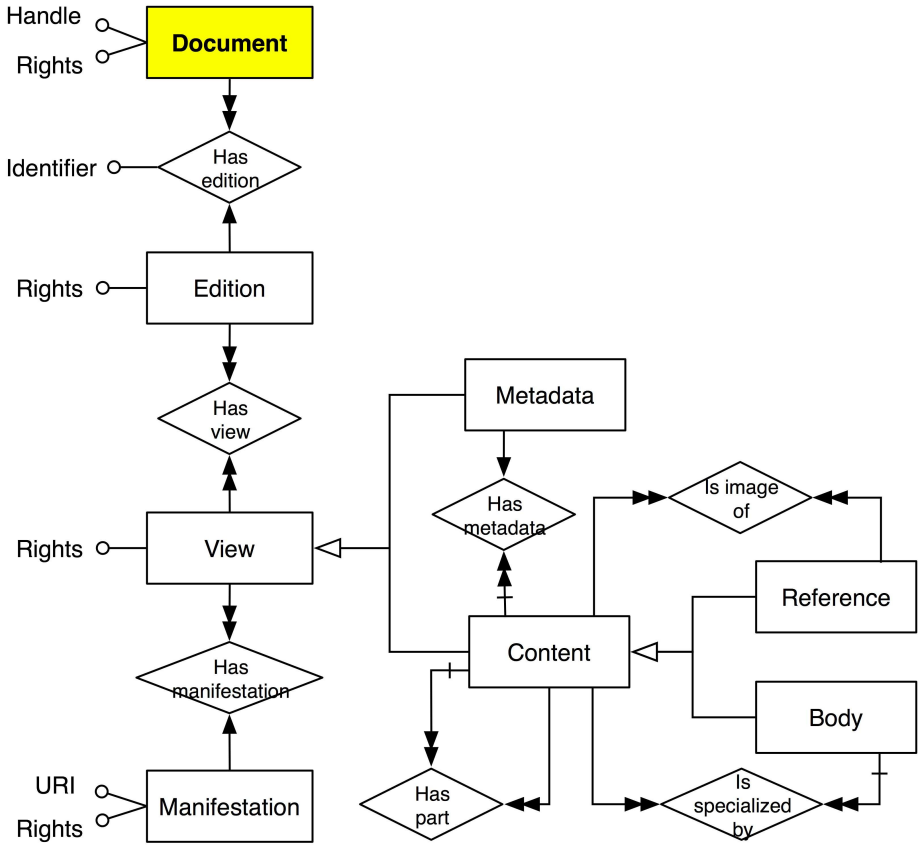
The Document Entity

Definition

A Document entity represents a distinct intellectual creation.

Example

The book *Digital Libraries and Electronic Publishing* by W. Arms, the lecture *Introduction to Mixed Media Digital Libraries* by C. Lagoze.



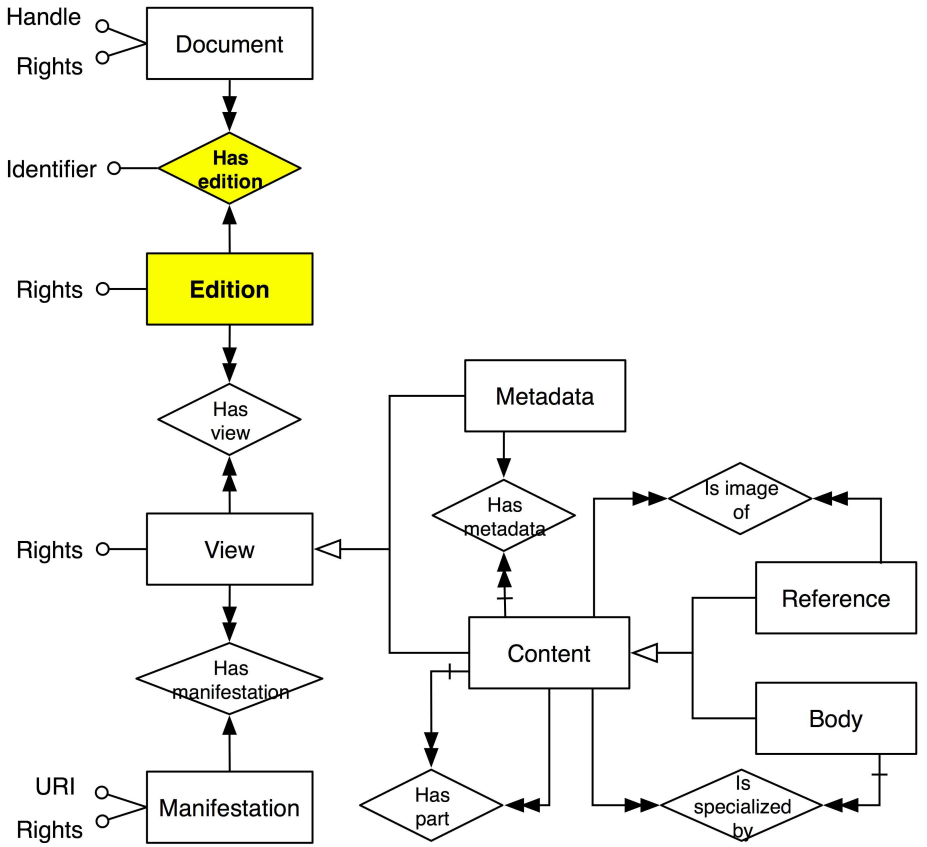
The Edition Entity

Definition

An Edition entity represents an expression of the Document along the time dimension.

Example

The preliminary version of the paper, the version submitted to the conference, the version published into the conference proceedings.



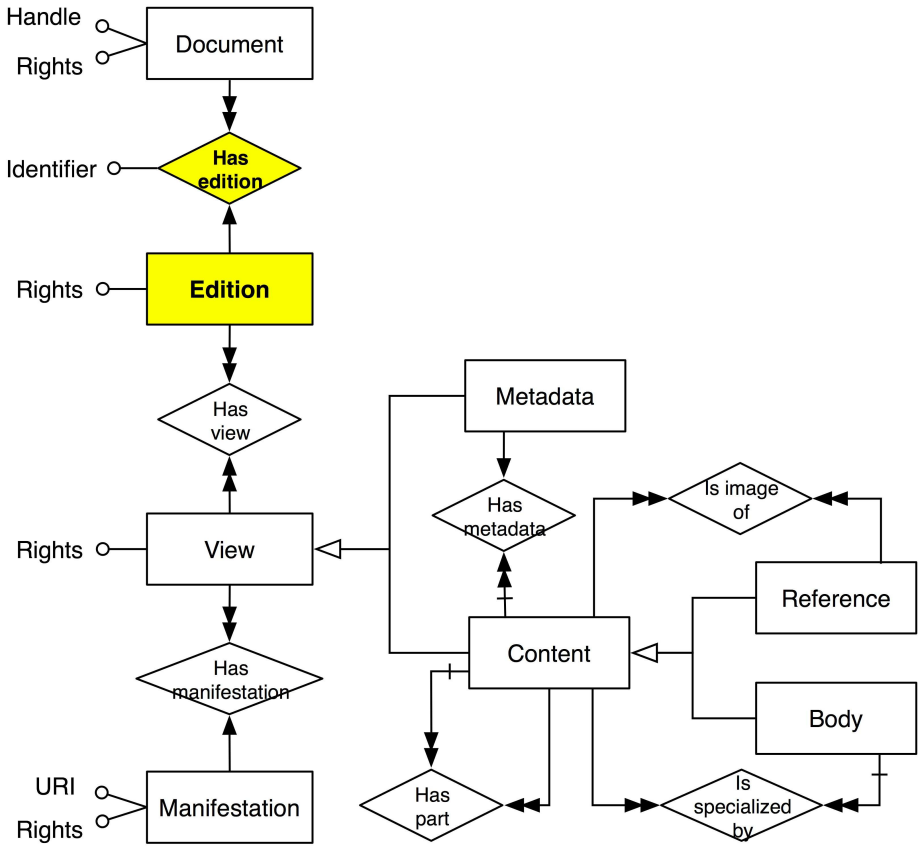
The Edition Entity

Definition

An Edition entity represents an expression of the Document along the time dimension.

Example

The preliminary version of the paper, the version submitted to the conference, the version published into the conference proceedings.



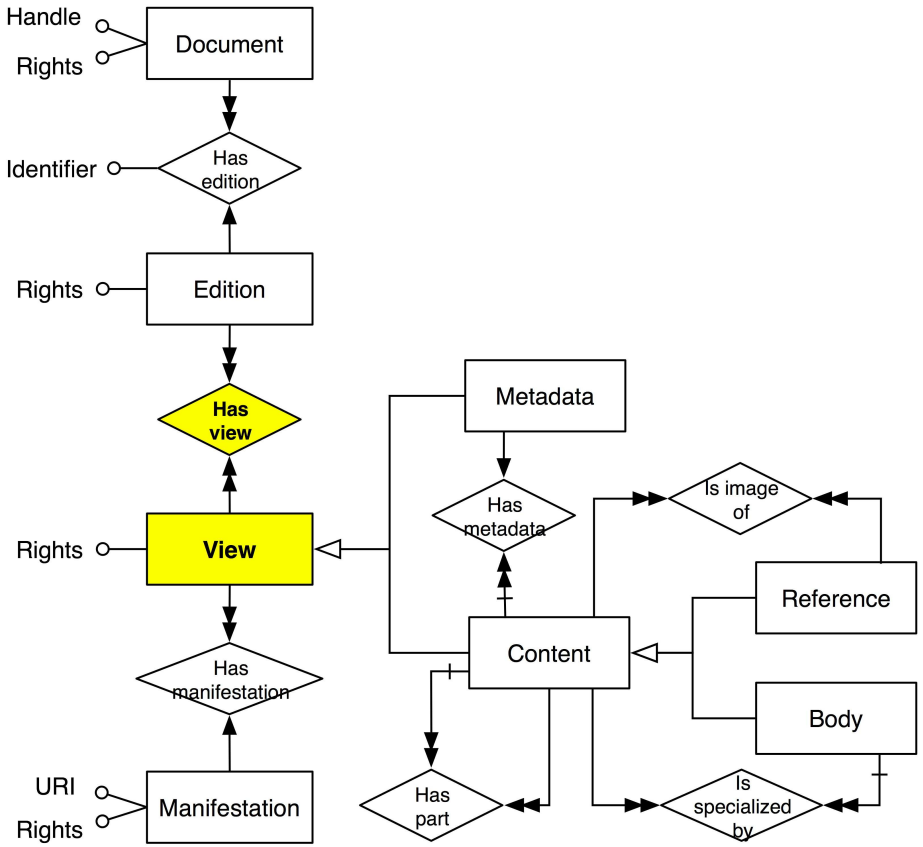
The View Entity

Definition

A View entity represents the way through which an edition is perceived.

Example

Workshop views: a *structured view* (the preface and the papers), a *presentation view* (the slides and their abstract), and a *metadata view* (structured description of the proceedings).



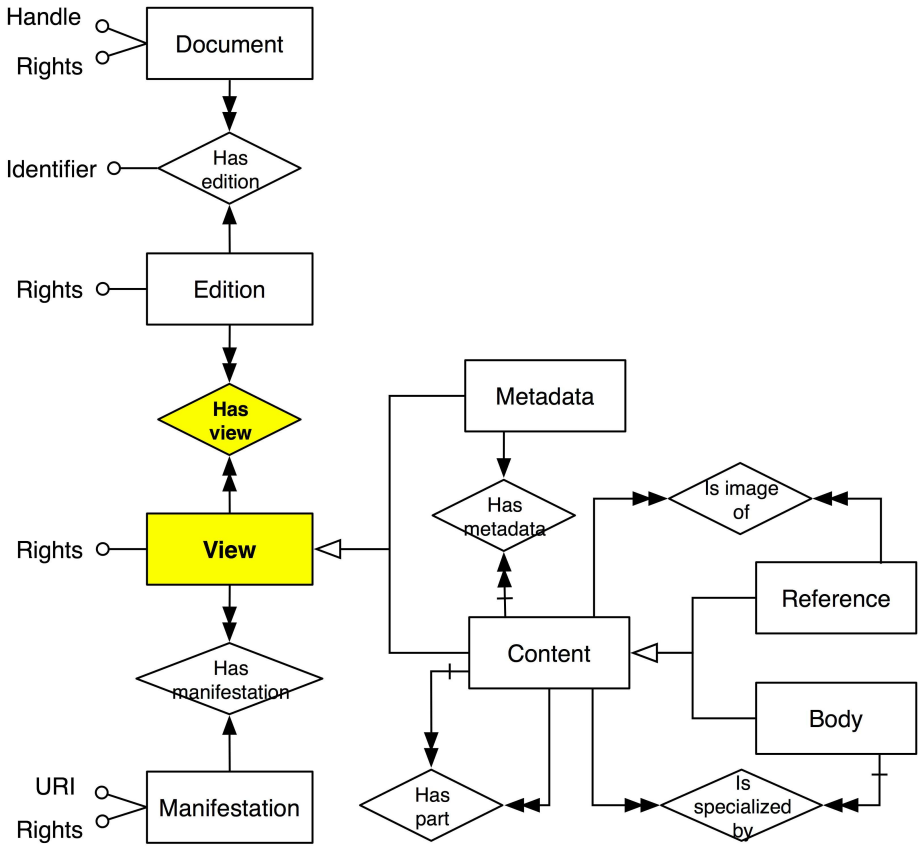
The View Entity

Definition

A View entity represents the way through which an edition is perceived.

Example

Workshop views: a *structured view* (the preface and the papers), a *presentation view* (the slides and their abstract), and a *metadata view* (structured description of the proceedings).



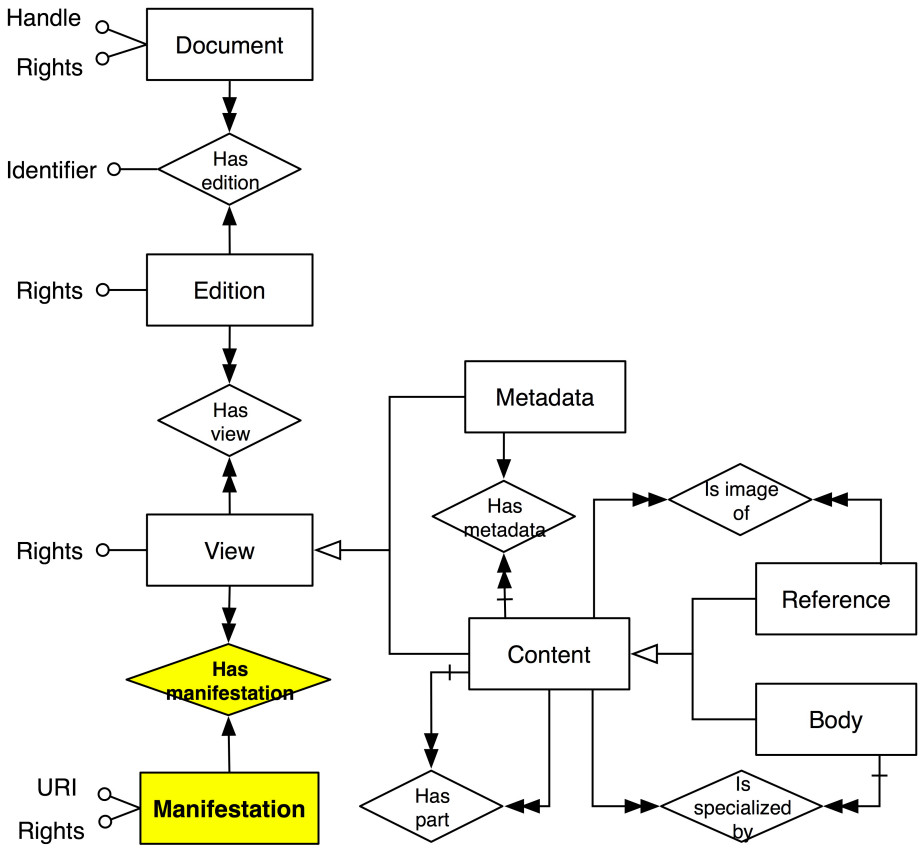
The Manifestation Entity

Definition

The Manifestation entity represents the physical format by which a document is disseminated.

Example

The MPEG with the video of a lecture, the AVI of the same video.



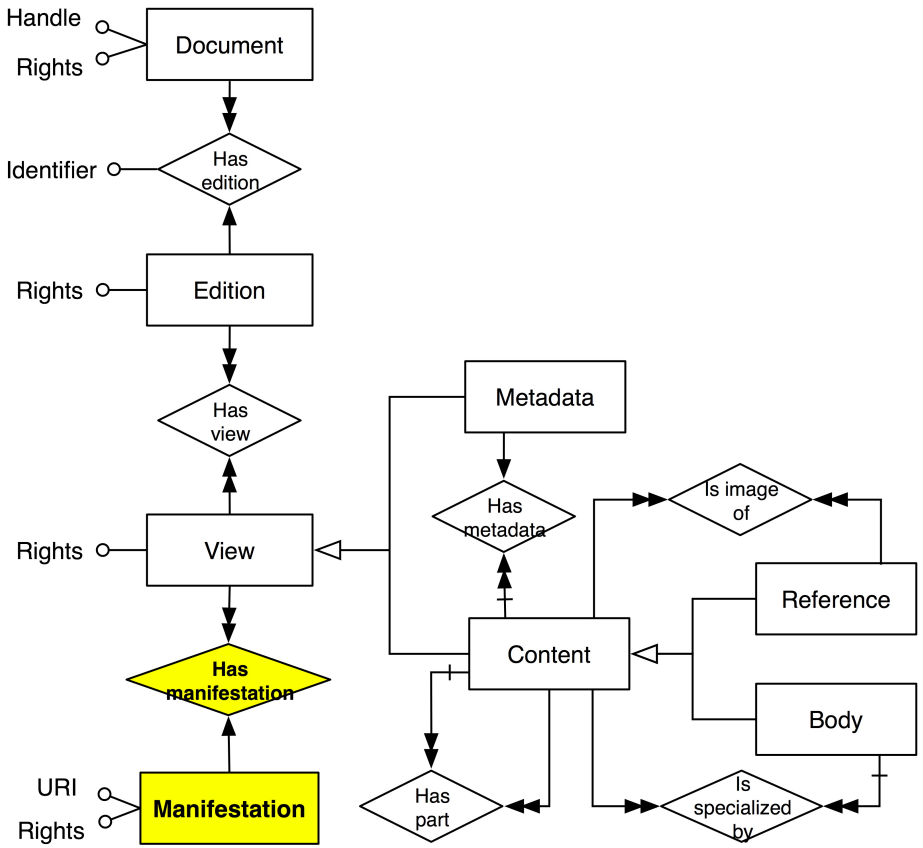
The Manifestation Entity

Definition

The Manifestation entity represents the physical format by which a document is disseminated.

Example

The MPEG with the video of a lecture, the AVI of the same video.



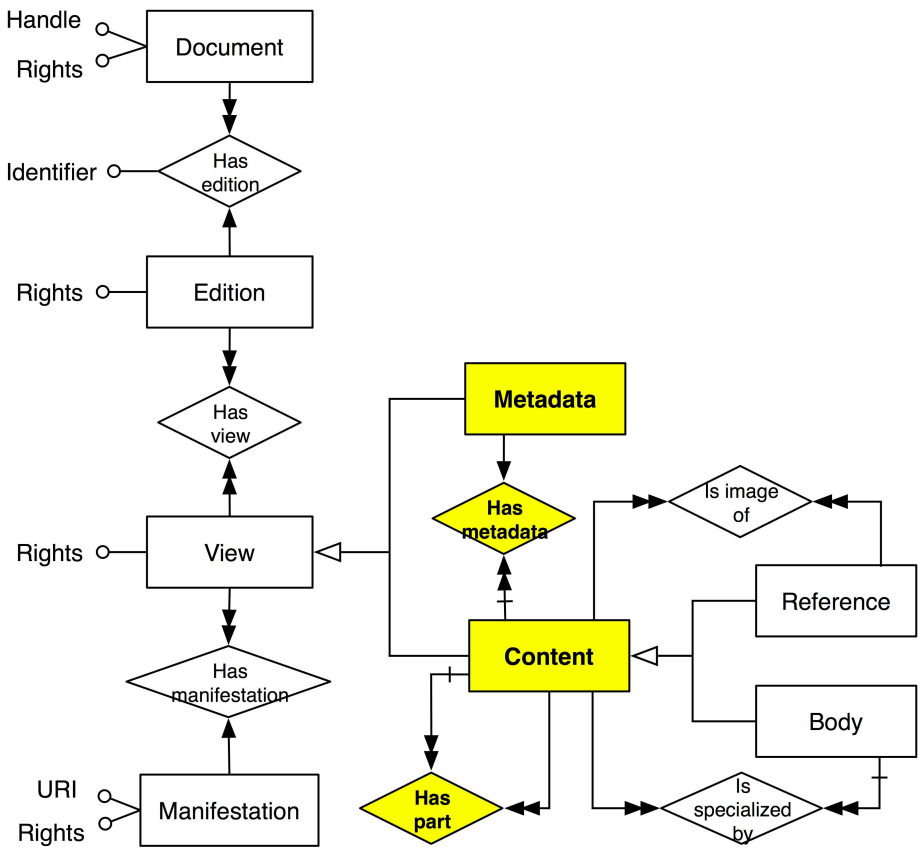
Metadata and Content Views

Definition

A Metadata View entity models the metadata representation through which a document edition is perceived. Typically used to support index and browse.

Definition

A Content View entity models the content through which a document edition is perceived.



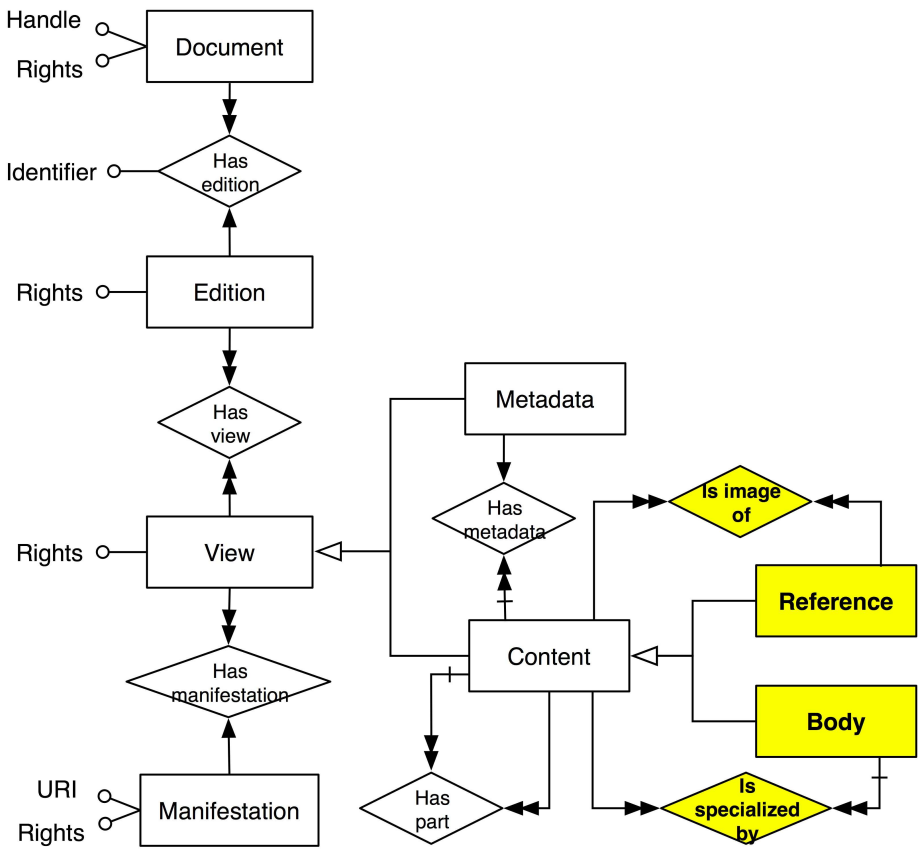
Body and Reference Contents

Definition

A Body View entity represents the document content either as a whole or as an aggregation of other views.

Definition

A Reference View entity represents the document content without having a proper manifestation but just holding a link to another document.



OpenDLib

A DLMS, i.e. a system for creating and managing DLs, developed at ISTI-CNR[a].
It is a software toolkit that appropriately configured and instantiated allows to set up DLs capable to meet diverse requirements. It provides:

- a pool of *core* customizable DL services
- a customizable DL architecture
- a powerful and customizable document model by implementing DoMDL

[a] `http://www.opendlib.com`

DoMDL Representation

In OpenDLib a DoMDL document is logically composed by two parts:

- the structure file, i.e. the description of the structure the document is organized in and thus the relationships among the part files
  - modeled as XML document validated against the DoMDL XML Schema
- the part files, i.e. the real data constituting the document
  - single files managed separately

DoMDL Representation

In OpenDLib a DoMDL document is logically composed by two parts:

- the structure file, i.e. the description of the structure the document is organized in and thus the relationships among the part files
  - modeled as XML document validated against the DoMDL XML Schema
- the part files, i.e. the real data constituting the document
  - single files managed separately

DoMDL Storage

The functionality of permanent holding *documents*

- the physical storage is up to the underlying technology, i.e. [distributed] file system
- constraints and opportunities
  - multiplicity of metadata and manifestations enables the possibility to use transformers, e.g. for more convenient dissemination, for preservation purposes
  - manifestations identified by URI enable different storage strategies
  - reference views reduce data duplication

DoMDL Access

The functionality of using a *document*, namely its content

- Possible approaches:
  1. expose the document structure data
  2. provide an access API (hide the structure)
- OpenDLib implements both since services may:
  1. need to have access to the document structure
  2. retrieve the parts they are interested in

DoMDL Discovery

The functionality enabling users to identify *documents* they are interested in

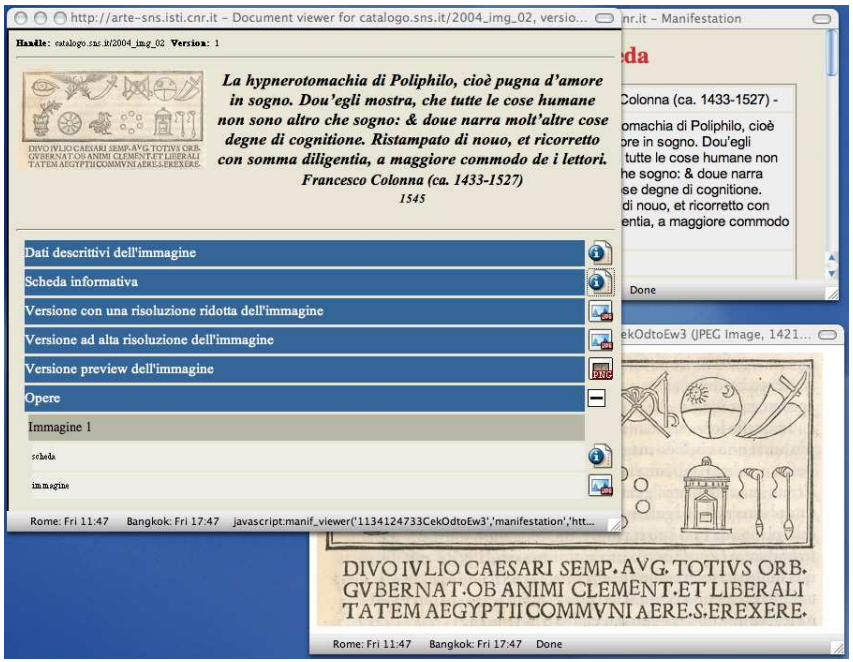
- usually provided through index and search services
- build over metadata
- DoMDL impacts on search and index design
  - indexes highly configurable, e.g. fields to index, result set format
  - search acts as a query mediator over indexes
  - transformers enable to have full text search

DoMDL Visualization

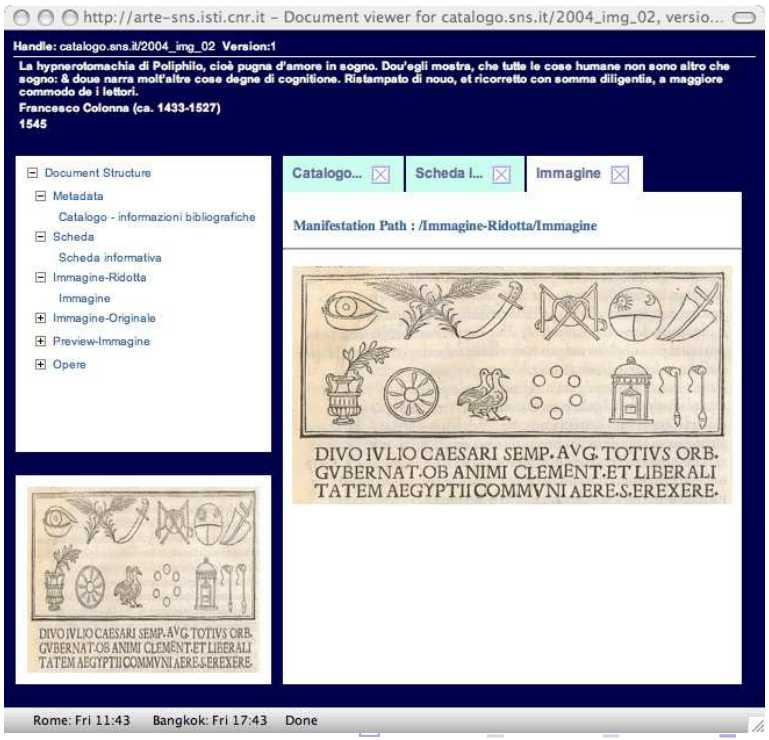
The functionality through which users perceive DL *documents*

- highly configurable due to the fine grained document access facilities

window based

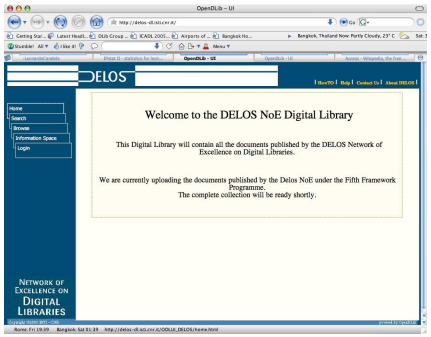


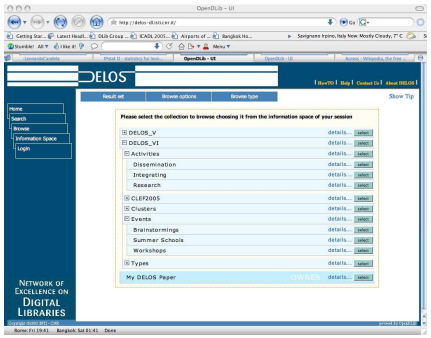
tab based

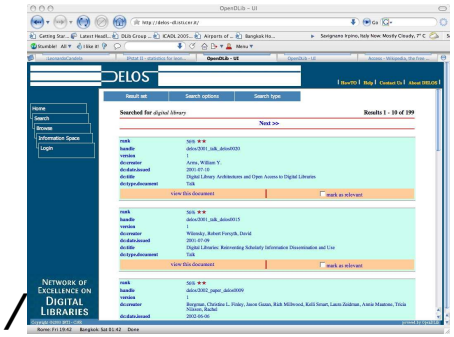


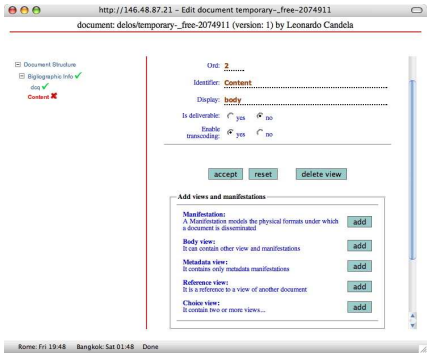
The DELOS DL Experience

- DELOS DL[a] supports the DELOS NoE[b] activities
- contains material of thematic workshops, brainstormings, summer schools, etc.
- perceived as a tool to promote cooperation and collaboration



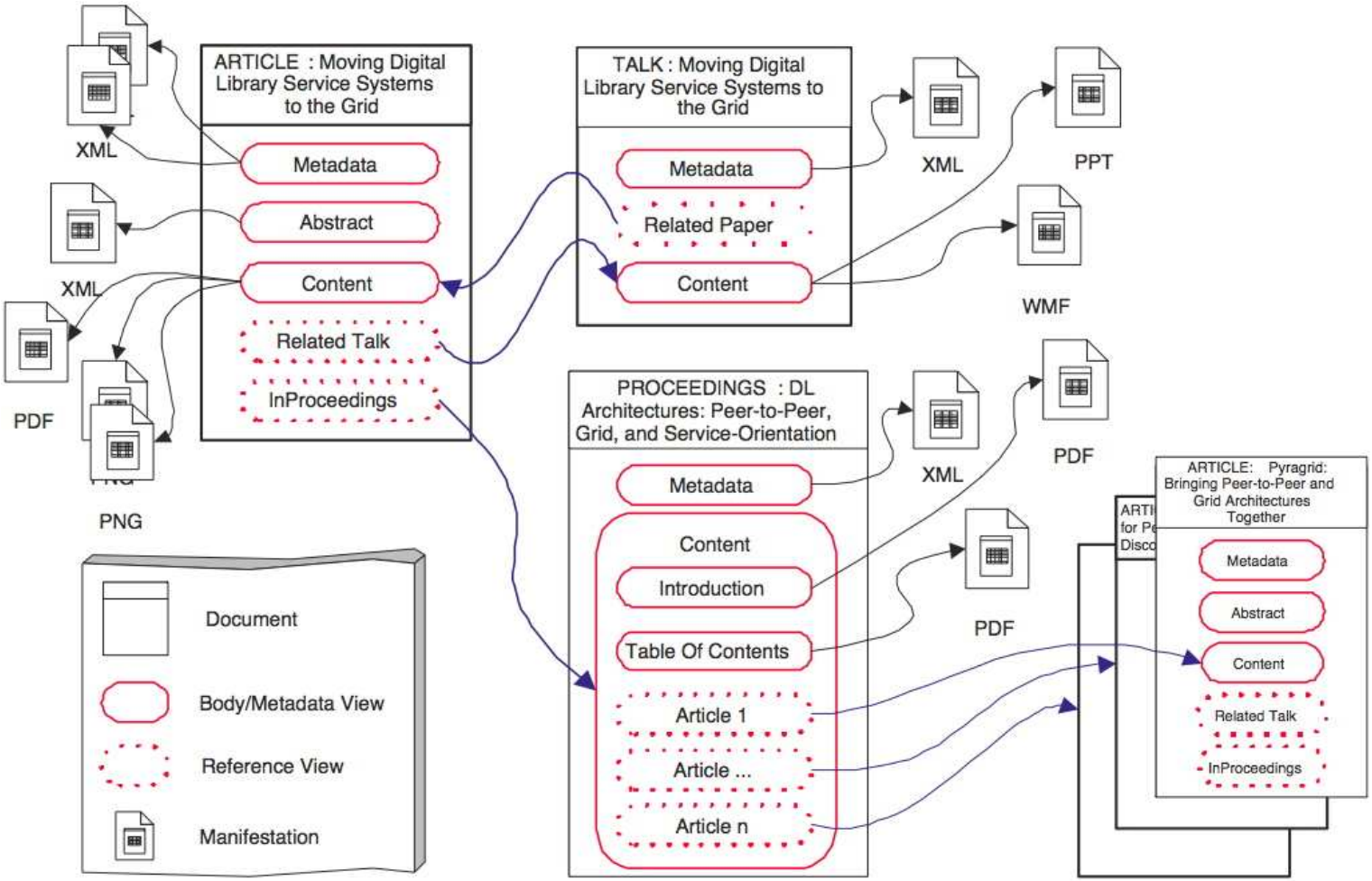






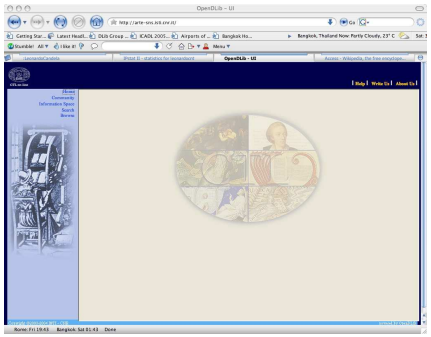
[a] http://delos-dl.isti.cnr.it/
[b] http://www.delos.info

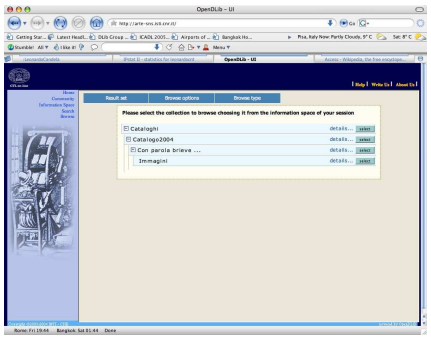
The DELOS DL Experience (cont.)

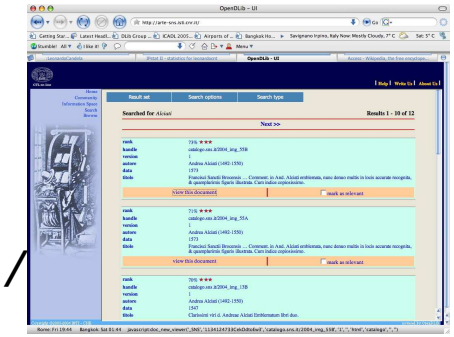


The ARTE DL Experience

- ARTE DL [a] supports the ARTE project
- contains the digitized versions on ancient books and their images
- documents enriched with a set of semantic links

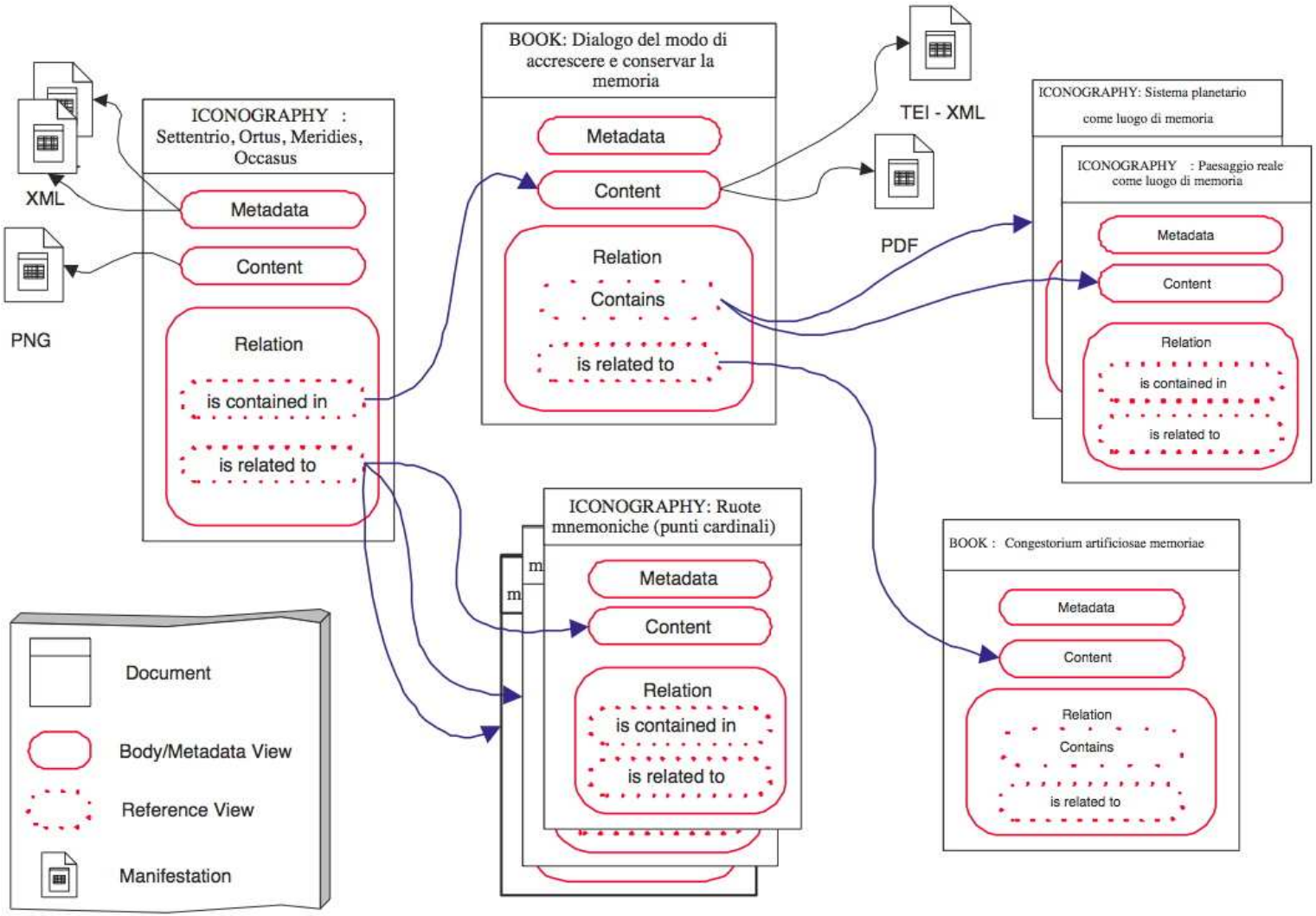






[a] http://arte-sns.isti.cnr.it/

The ARTE DL Experience (cont.)



Summary

- DoMDL is a powerful and flexible document model capable to represent multi-edition, structured, multimedia documents that can be disseminated in multiple formats
- OpenDLib implements DoMDL
- The model has been validated by communities belonging to different application domains
- Next step: living documents, i.e. using Grid technologies to dynamically generate parts of documents

http://www.opendlib.com

Summary

- DoMDL is a powerful and flexible document model capable to represent multi-edition, structured, multimedia documents that can be disseminated in multiple formats
- OpenDLib implements DoMDL
- The model has been validated by communities belonging to different application domains
- Next step: living documents, i.e. using Grid technologies to dynamically generate parts of documents

http://www.opendlib.com

Summary

- DoMDL is a powerful and flexible document model capable to represent multi-edition, structured, multimedia documents that can be disseminated in multiple formats
- OpenDLib implements DoMDL
- The model has been validated by communities belonging to different application domains
- Next step: living documents, i.e. using Grid technologies to dynamically generate parts of documents

http://www.opendlib.com

Summary

- DoMDL is a powerful and flexible document model capable to represent multi-edition, structured, multimedia documents that can be disseminated in multiple formats
- OpenDLib implements DoMDL
- The model has been validated by communities belonging to different application domains
- Next step: living documents, i.e. using Grid technologies to dynamically generate parts of documents

http://www.opendlib.com

Summary

- DoMDL is a powerful and flexible document model capable to represent multi-edition, structured, multimedia documents that can be disseminated in multiple formats
- OpenDLib implements DoMDL
- The model has been validated by communities belonging to different application domains
- Next step: living documents, i.e. using Grid technologies to dynamically generate parts of documents

`http://www.opendlib.com`

Additional slides

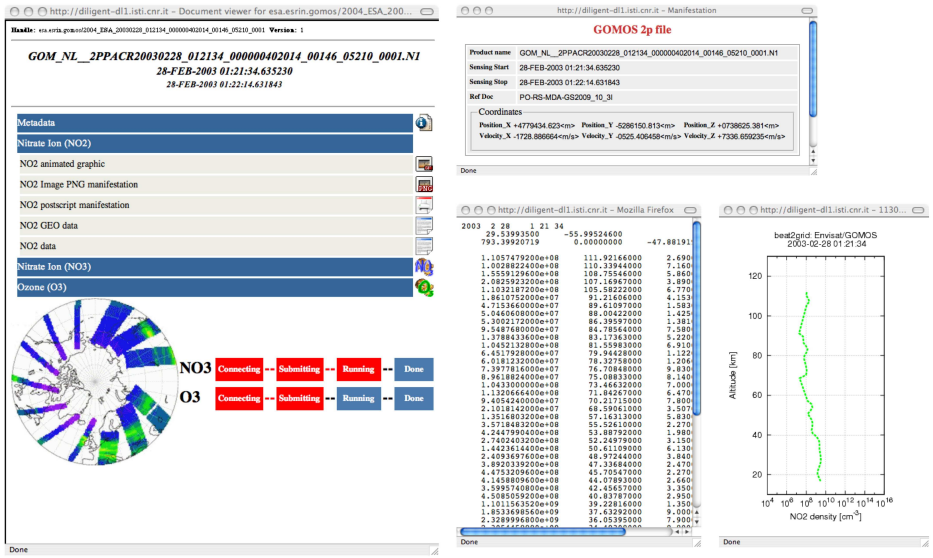
Living Documents: The ESA Experience

Goal

- provide to ESA dynamic reports, i.e. a living documents whose ozone and nitrate maps are dynamically generated
- experiment Grid and DLs technologies

Outcome

- DoMDL is able to represent living documents
- Grid + DLs: many opportunities



Living Documents: The Architecture