

Institutional Repositories and the OAI-PMH: beyond Dublin Core

Henry Jerez, Jeroen Bekaert, and Herbert Van de Sompel
Los Alamos National Laboratory, Research Library
Digital Library Research & Prototyping Team

Outline

(1) Motivation

(2) OAI-PMH for content

(3) Example 1 : LANL Repository

(4) Example 2 : mod_oai

(5) Example 3 : DSpace plug-in prototype

(6) Federations of IRs and OAI-PMH

(7) Conclusion

Motivation

- Digital Libraries, Institutional Repositories, Archives
 - Growing interest in exposing/harvesting content, not only metadata
 - cf. DARE, DINI, JISC FAIR, DSpace
 - Growing interest from Web search engines to harvest quality content from these repositories.
 - Well-established adoption of the OAI-PMH. Tools available. It makes sense to use OAI-PMH to expose/harvest content.
 - ***But can content be exposed/harvested through OAI-PMH? See later.***
- The Web
 - Web crawling solutions not utterly efficient.
 - No efficient change control mechanism on the Web.
 - OAI-PMH can provide optimizations.
 - ***But can general Web content be harvested through OAI-PMH? See later.***

Outline

(1) Motivation

(2) OAI-PMH for content

(3) Example 1 : LANL Repository

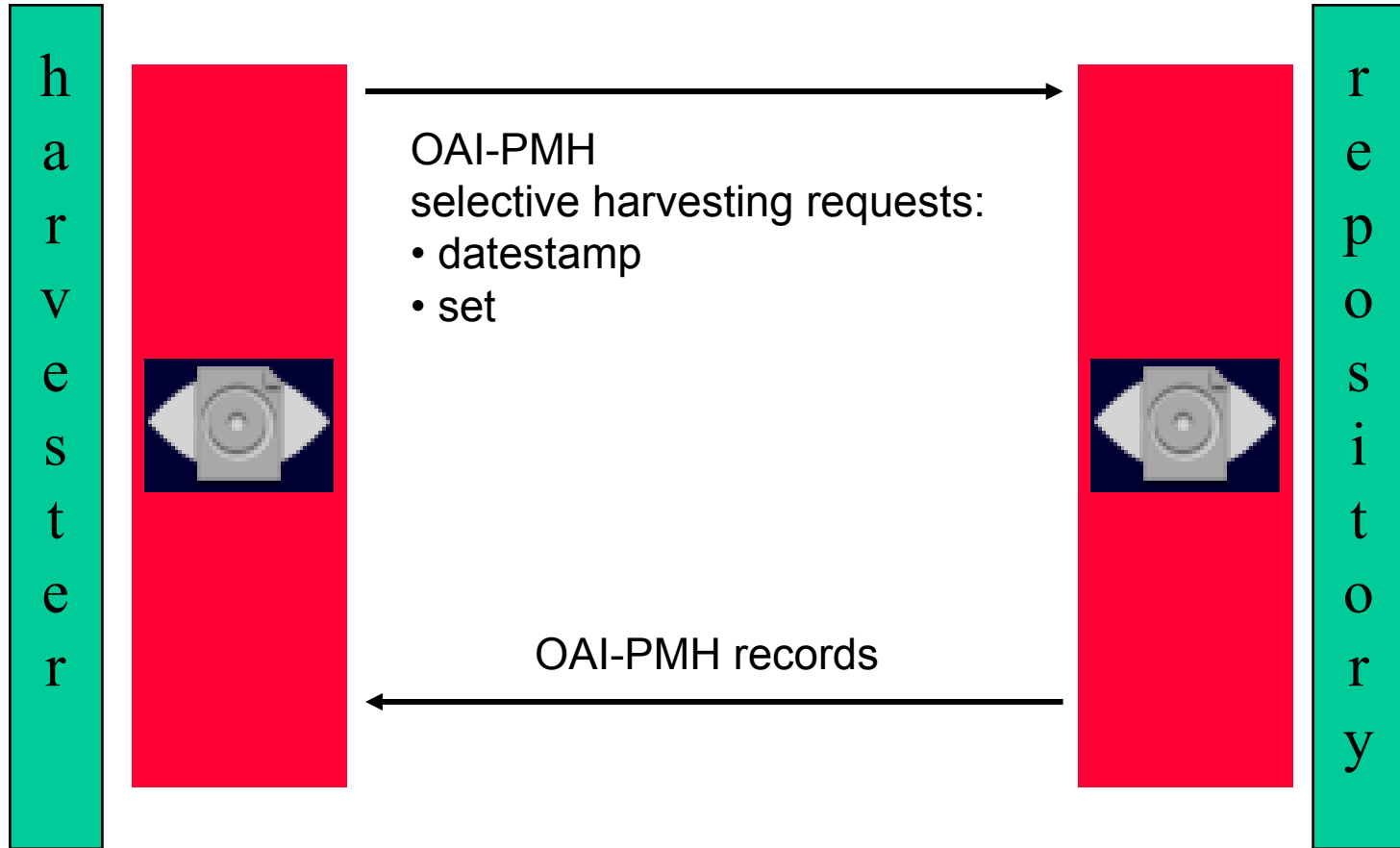
(4) Example 2 : mod_oai

(5) Example 3 : DSpace plug-in prototype

(6) Federations of IRs and OAI-PMH

(7) Conclusion

OAI-PMH



provides services using harvested metadata

exposes metadata pertaining to resources

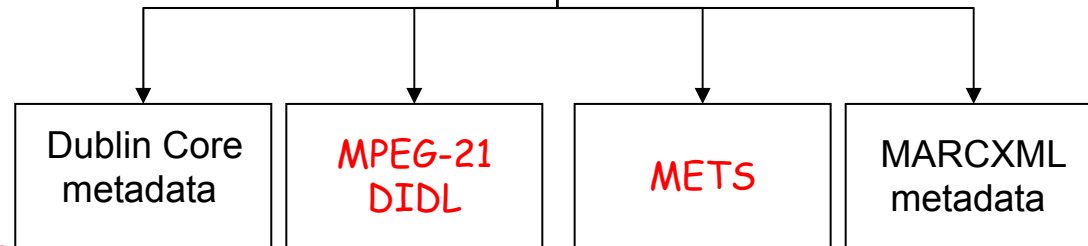
OAI-PMH data model



← resource

OAI-PMH identifier
= entry point to all records pertaining to the resource

← item



← records

• metadata pertaining to the resource

• XML data pertaining to the resource

• modeled representation of the resource

simple model

complex model

complex model

simple model

OAI-PMH and complex models

- OAI-PMH record == modeled representation of the resource
- Can be selectively harvested via OAI-PMH ~ datestamp, set
- Resource can be:
 - simple object (1 file)
 - compound object (multiple files)
- OAI-PMH records can contain:
 - Typical metadata
 - A variety of secondary information: rights, relationships, format information, ...
 - Actual resource(**s**)
 - By-Value – base64 encoded
 - By-Reference – http address of resource
 - both
 - Identifiers of metadata and resource(**s**), unambiguously mapped to the identified data

OAI-PMH and complex models: data/id mapping

- o Example: a compound object consisting of:
 - metadata
(id = info:lanl-repo/opac/LANLb10012271)
 - technical report
 - 1 file: pdf
(id = info:lanl-repo/tr/LA-9870)
 - 1 file: tiff
(id = info:lanl-repo/tr/LA-9871)

OAI-PMH and complex models: data/id mapping

complex model

meta - id: info:lanl-repo/opac/LANLb10012271
ref: http://library.lanl.gov/md/foo.xml
ds1 - id: info:lanl-repo/tr/LA-9870
ref: http://library.lanl.gov/tr/foo.pdf
ds2 - id: info:lanl-repo/tr/LA-9871
ref: http://library.lanl.gov/tr/foo.tiff

simple model : DC

dc:identifier: info:lanl-repo/tr/LA-9870
dc:identifier: info:lanl-repo/tr/LA-9871
dc:identifier: http://library.lanl.gov/tr/foo.pdf
dc:identifier: http://library.lanl.gov/tr/foo.tiff

- No distinction between identifiers & locators
- Unclear relation between identifiers & locators
- Where does the identifier of the metadata go?

OAI-PMH & complex models : related papers

- Using the OAI-PMH ... Differently.
<http://www.dlib.org/dlib/july03/young/07young.html>
- Using MPEG-21 DIDL to Represent Complex Digital Objects in LANL
<http://www.dlib.org/dlib/november03/bekaert/11bekaert.html>
- Using MPEG-21 DIP and NISO OpenURL for the Dynamic Dissemination of Complex Digital Objects in LANL
<http://www.dlib.org/dlib/february04/bekaert/02bekaert.html>
- The multi-faceted use of the OAI-PMH in the LANL Repository
<http://lib-www.lanl.gov/~herbertv/papers/jcdl2004-submitted-draft.pdf>

Outline

(1) Motivation

(2) OAI-PMH for content

(3) Example 1 : LANL Repository

(4) Example 2 : mod_oai

(5) Example 3 : DSpace plug-in prototype

(6) Federations of IRs and OAI-PMH

(7) Conclusion

Example 1 : LANL Repository

- Local storage of Terrabytes of scholarly assets
- Upon ingestion, assets are turned into MPEG-21 DIDL documents that contain:
 - Metadata pertaining to assets
 - Assets and/or pointers to assets
 - Identifiers of metadata, assets, DIDL documents
 - A variety of secondary information
- Stored MPEG-21 DIDL documents made accessible to – multiple – downstream applications via the OAI-PMH
- OAI-PMH as a Repository Access Protocol to access metadata and content.

Outline

(1) Motivation

(2) OAI-PMH for content

(3) Example 1 : LANL Repository

(4) Example 2 : mod_oai

(5) Example 3 : DSpace plug-in prototype

(6) Federations of IRs and OAI-PMH

(7) Conclusion

Example 2 : Old Dominion University & LANL mod_oai project

- Funded by Andrew W. Mellon Foundation
- Implement OAI-PMH plug-in for – Apache - Web servers
- Will allow selective & incremental OAI-PMH harvesting of content hosted by Web servers
 - OAI-PMH identifiers == URLs
 - datestamp
 - sets ~ MIME type
 - initially static Web content
- Two operating modes for crawlers:
 - General crawler: ListIdentifiers => URLs of Web content
 - Advanced crawler: ListRecords ~ Dublin Core and one or more complex object formats
- OAI-PMH as a tool to make harvesting of Web content more efficient

Outline

(1) Motivation

(2) OAI-PMH for content

(3) Example 1 : LANL Repository

(4) Example 2 : mod_oai

(5) Example 3 : DSpace plug-in prototype

(6) Federations of IRs and OAI-PMH

(7) Conclusion

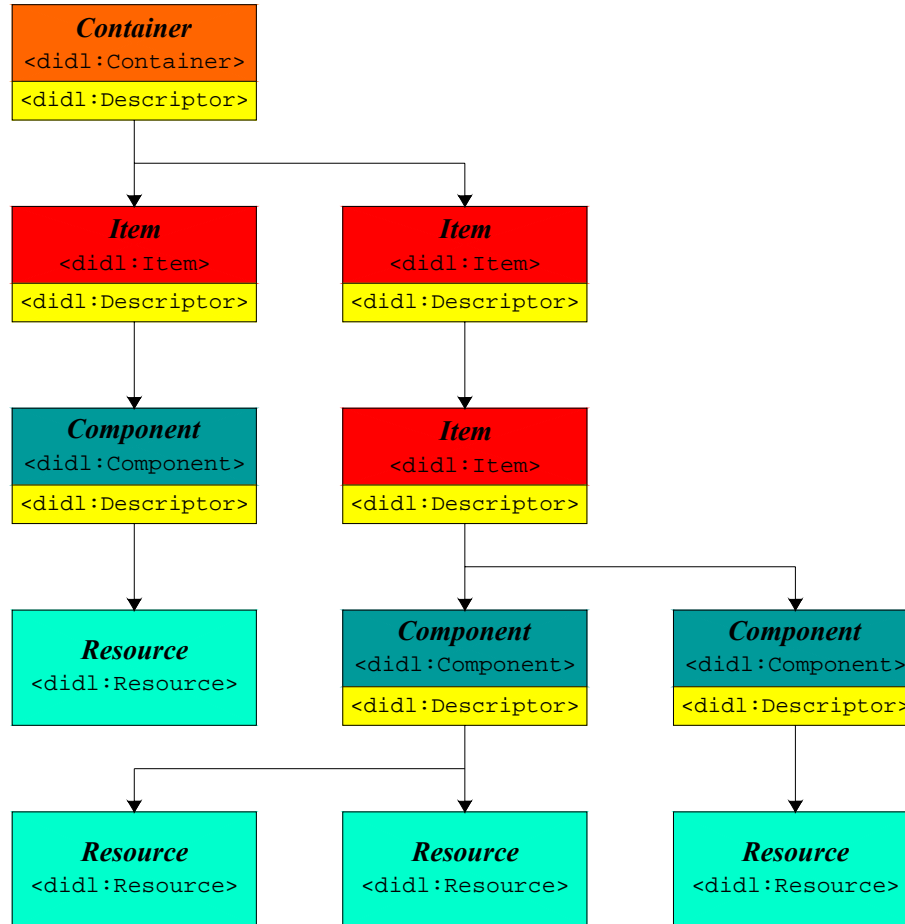
Example 3 : LANL DSpace plug-in prototype

- Introduced at recent DSpace Federation meeting
- Maps DSpace data model
[item – bundle – component]
to MPEG-21 DIDL data model
[Container – Item – Resource]
- Exposes MPEG-21 DIDL documents through built-in DSpace OAI-PMH infrastructure
- Metadata (Dublin Core) and Content (MPEG-21 DIDL) harvestable via the OAI-PMH

MPEG-21 DIDL : Data Model

- Abstract Definitions + W3C XML Schema
- Entities
 - a **Container** `didl:Container`
 - an **Item** `didl:Item`
 - a **Component** `didl:Component`
 - a **Resource** `didl:Resource`
 - a **Descriptor** `didl:Descriptor`
 - ...
- Remark
 - a DIDL compliant document == a DID

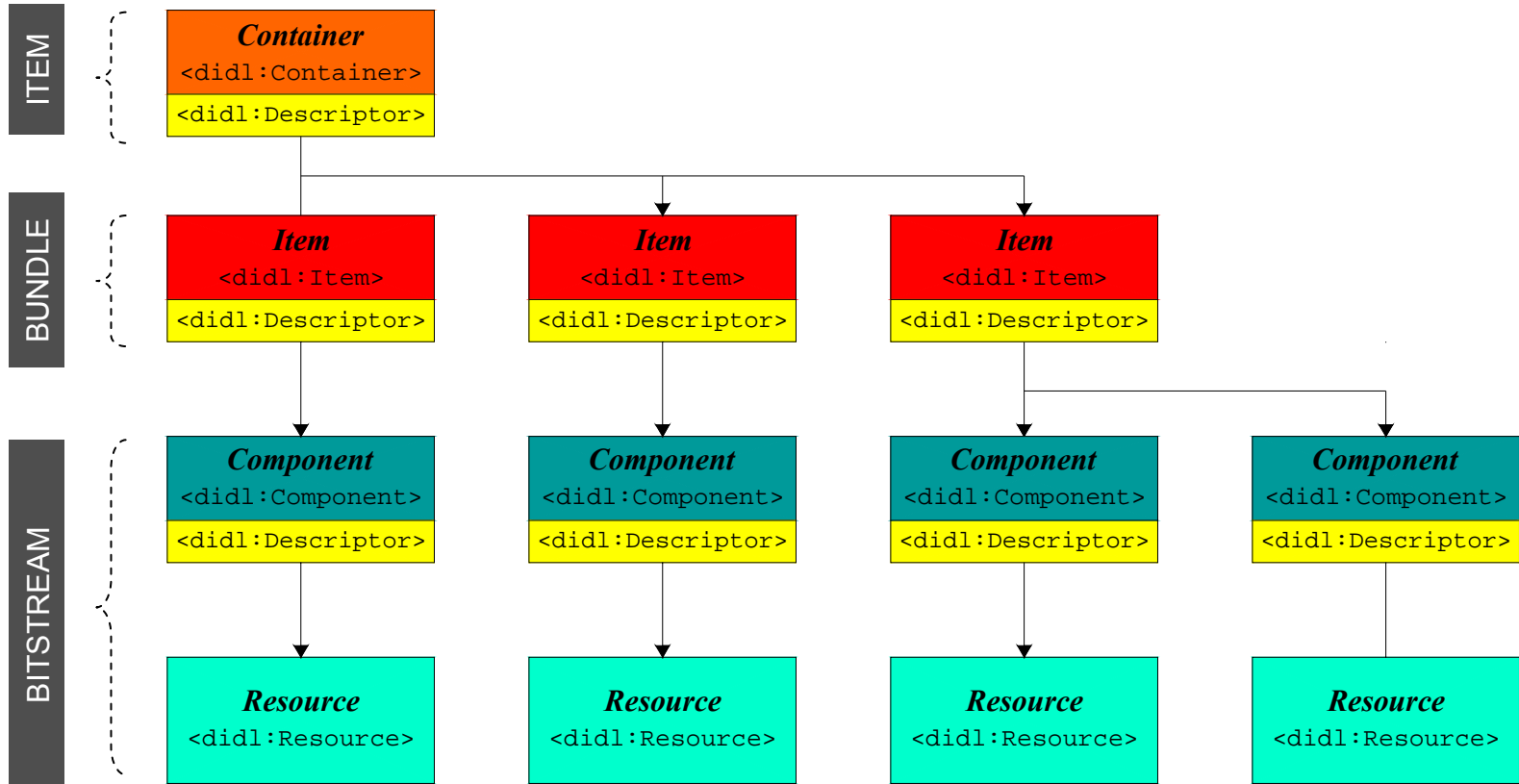
MPEG-21 DIDL : Data Model



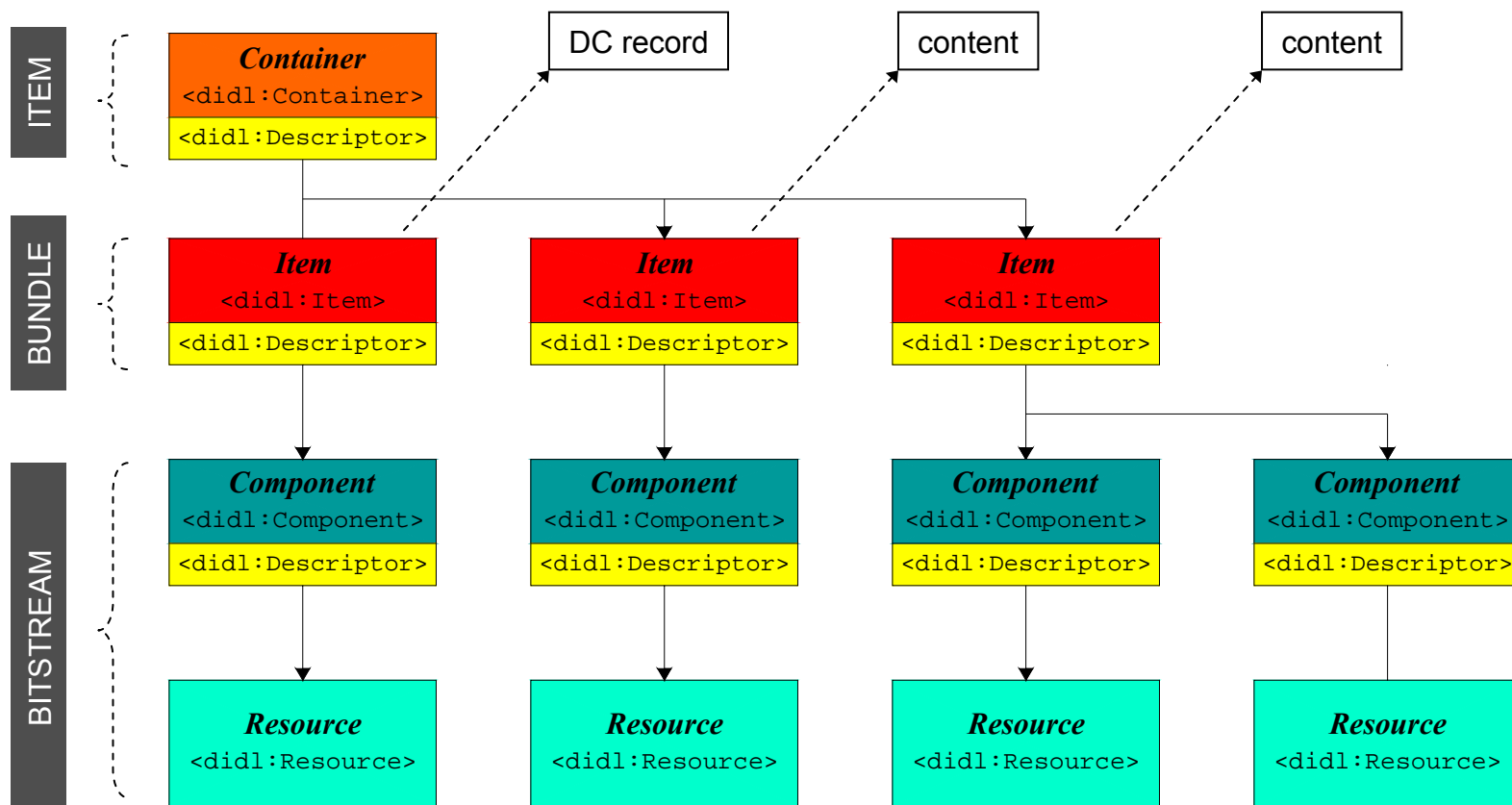
MPEG-21 DIDL : Descriptors

- Secondary information pertaining to Entities
 - MPEG-21 defined uses
 - identification information – MPEG-21 Part 3 : DII
 - rights information – MPEG-21 Part 5 : REL / Part 4 : IPMP
 - processing information – MPEG-21 Part 10 : DIP
 - community/application specific uses
 - e.g.: LANL use, DSpace use, ...

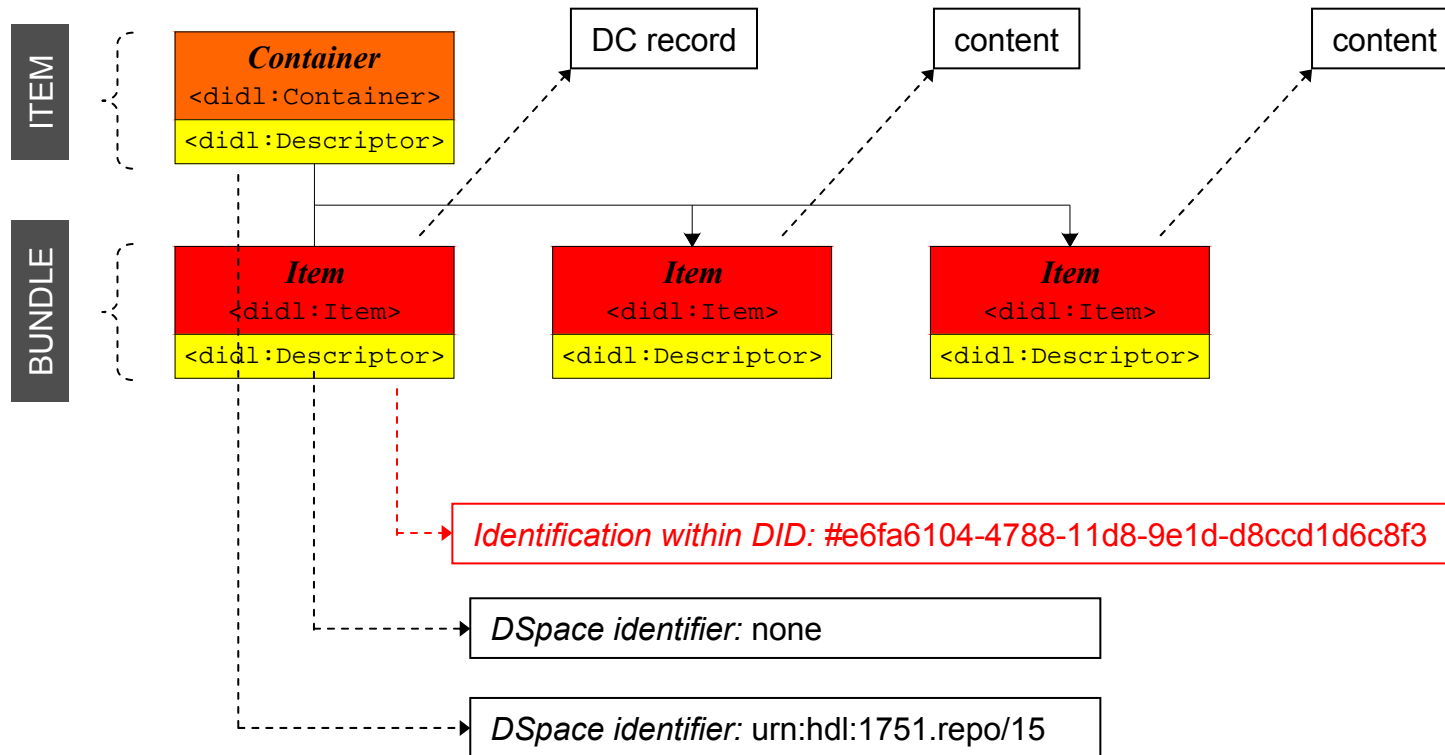
DSpace DID: general structure



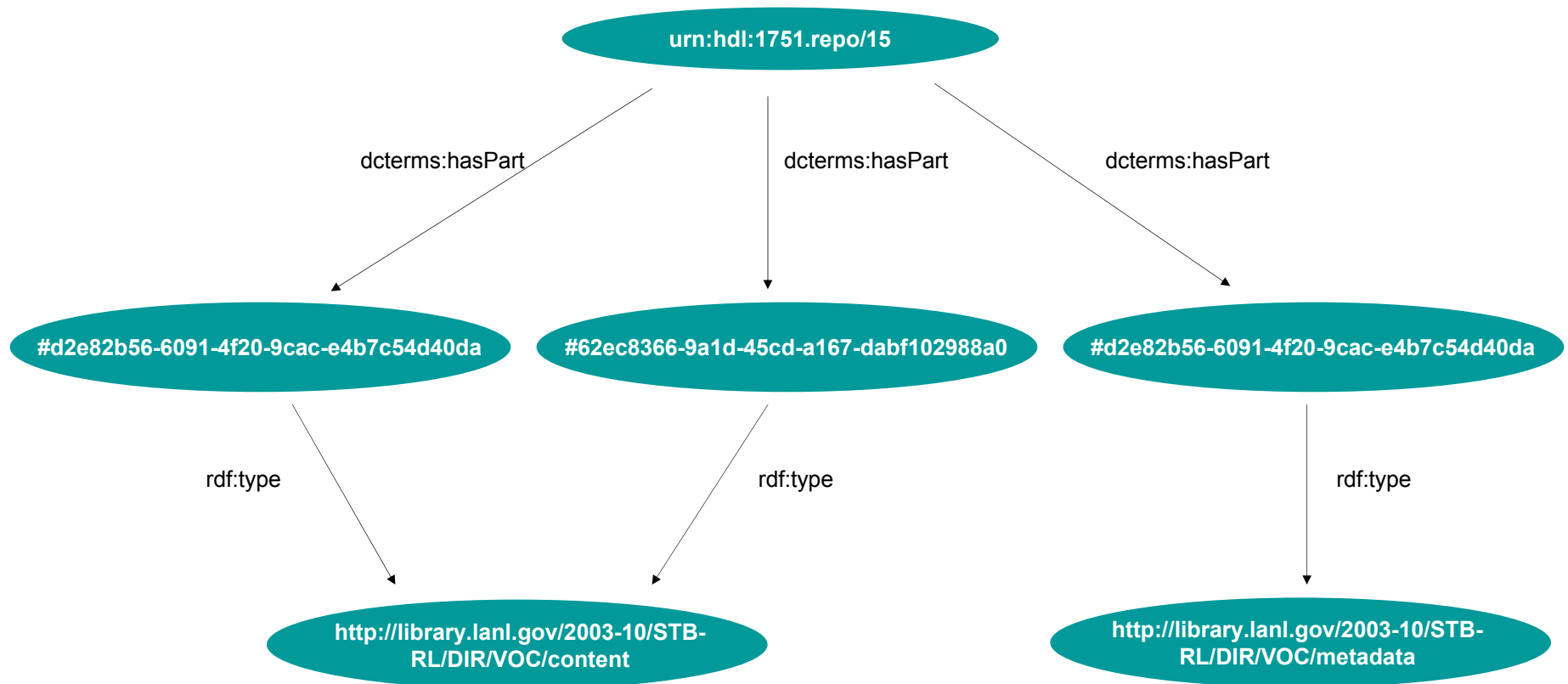
DSpace DID: mapping descriptive metadata & content



DSpace DID Descriptors : identifier



DSpace DID Descriptors : RDF relationships



DSpace to DID : mapping overview

MPEG-21 DIDL	DSpace
--------------	--------

Container	dii (MPEG-21)	Identifier	handle	Item
	diadm	dcterms:created	date_issued	
	dir	rdf		
	dipr	created		

Item	dii (MPEG-21)	Identifier		Bundle
	diadm	dcterms:created		

Component	dii (MPEG-21)	Identifier		Bitstream	
	diadm	dcterms:created			
	dii (MPEG-21)	@mimetype			mimetype
	diadm	digestValue			checksum
	diadm	digestMethod			checksum_algorithm

DSpace to DID – mapping considerations

- DSpace:
 - Lack of identifiers at Bundle and Bitstream level
 - Unknown mimeType
 - Unequal treatment of descriptive metadata and content. cf. MD5 digest.
 - Unclear use of rights and licenses
- DIDL:
 - Digest ~ W3C XML Signature
 - Community defined Namespaces for Descriptors required. For example: RDF.

LANL DSpace plug-in : DIDs via OAI-PMH

- DSpace DIDs contain:
 - identifiers
 - descriptive metadata
 - content
 - secondary information
- Harvestable through OAI-PMH
 - OCLC OAICat
 - Crosswalks
 - OAIDCCrosswalk.java
 - Components of LANL DSpace Plugin:
 - crosswalk: DIDLCrosswalk.java
 - Additional procedures:
 - XML ID creation UUID
 - RDF creation
 - metadata digest creation
 - full content base64 encoding

DIDLCrosswalk

- DSpace API procedures for complex objects
 - Item.java:DSpace:Item = DID.Item {DC}
 - Bundle.java:DSpace:Bundle = DID.Item
 - Bitstream.java:DSpace.Bitstreams = DID.Component
 - BistreamFormat.java to obtain secondary information
 - BitstreamStorageManager.java DSP:bitstream = DID.Resource
- Additional procedures
 - XML ID creation UUID
 - RDF creation
 - metadata digest creation
 - full content base64 encoding

LANL DSpace plug-in: further considerations

- DSpace DIDL plugin tested at LANL and Ghent University
- Issues encountered:
 - Lastmodified and OAI-PMH datestamp issues
 - Memory issues and the MAX_RECORDS
 - DSpace plugin implementation framework

Outline

(1) Motivation

(2) OAI-PMH for content

(3) Example 1 : LANL Repository

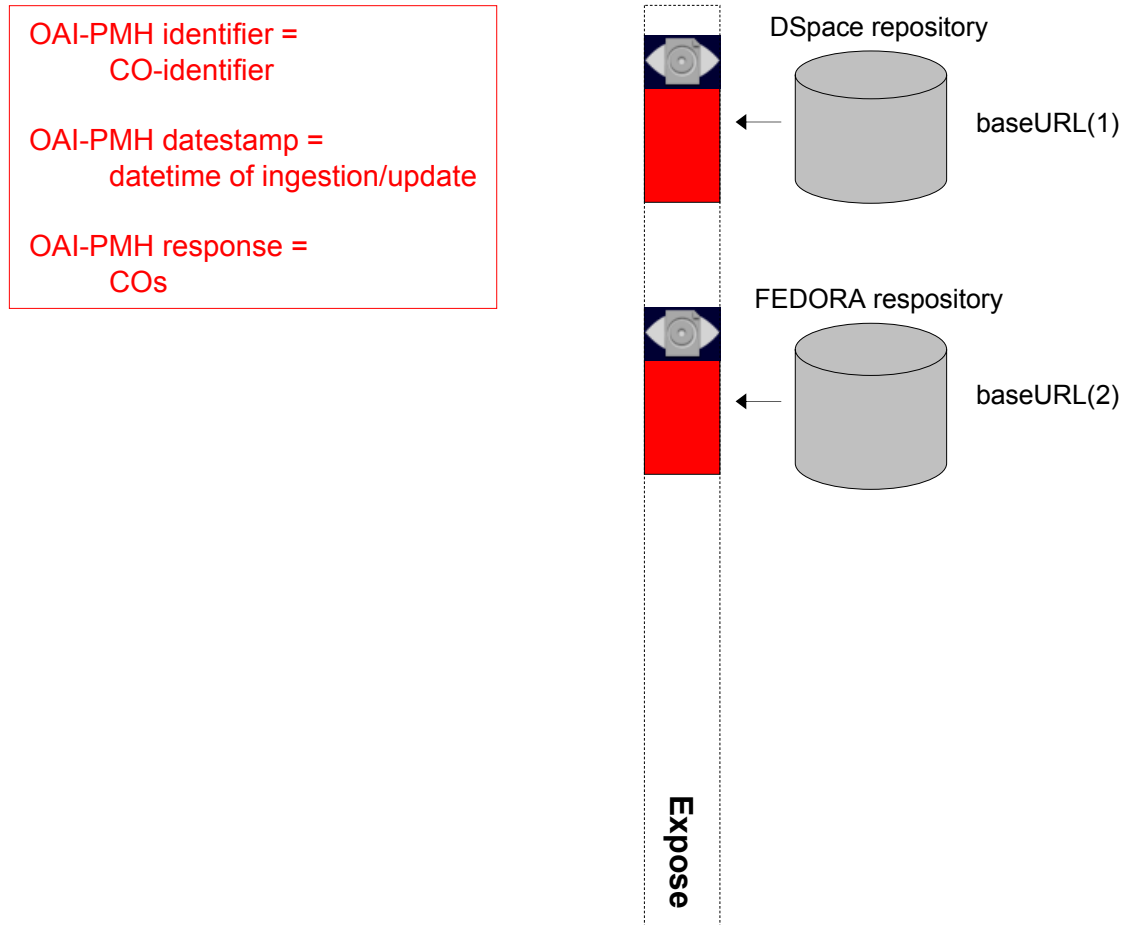
(4) Example 2 : mod_oai

(5) Example 3 : DSpace plug-in prototype

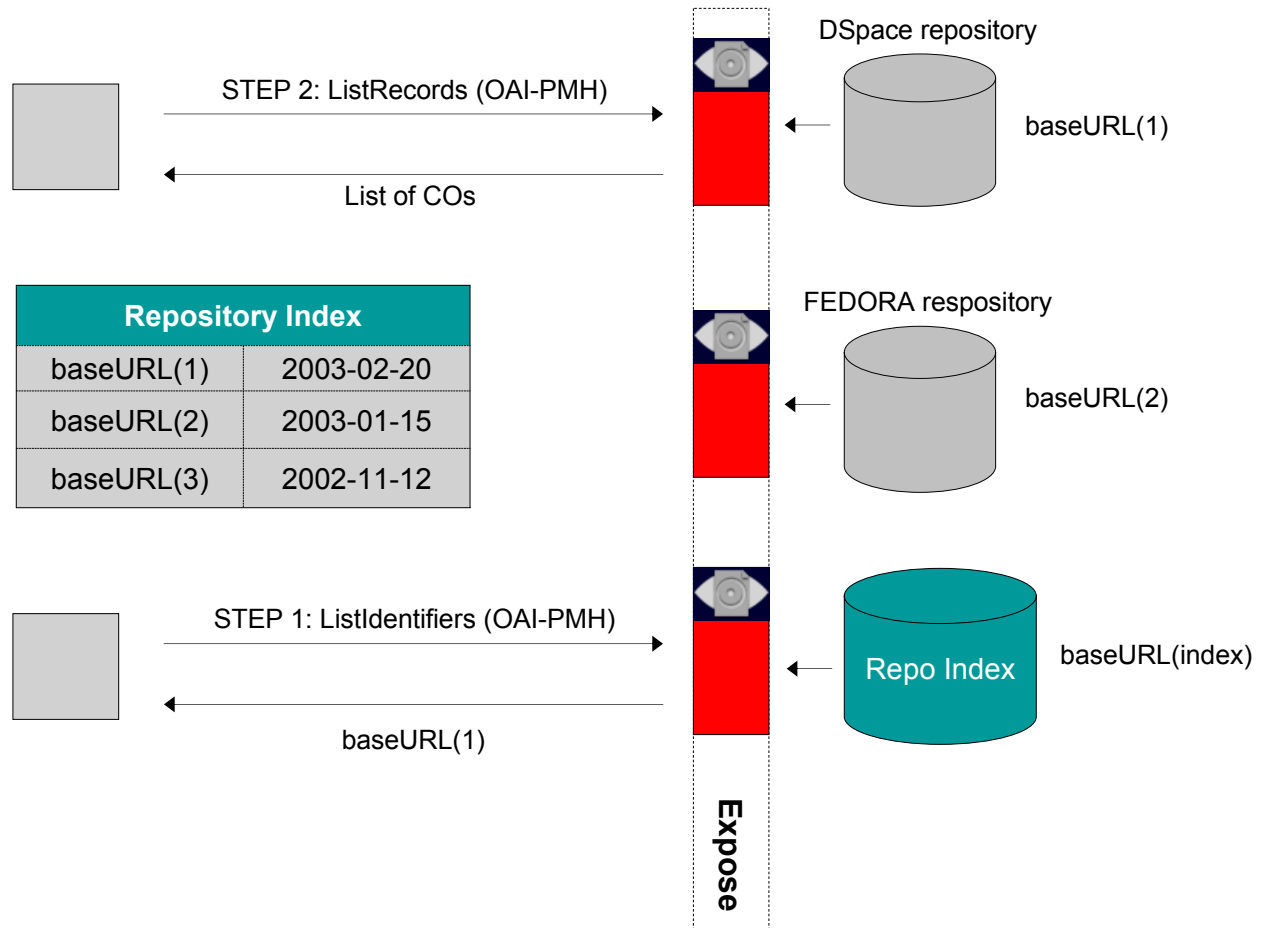
(6) Federations of IRs and OAI-PMH

(7) Conclusion

Harvesting COs from OAI-PMH repositories

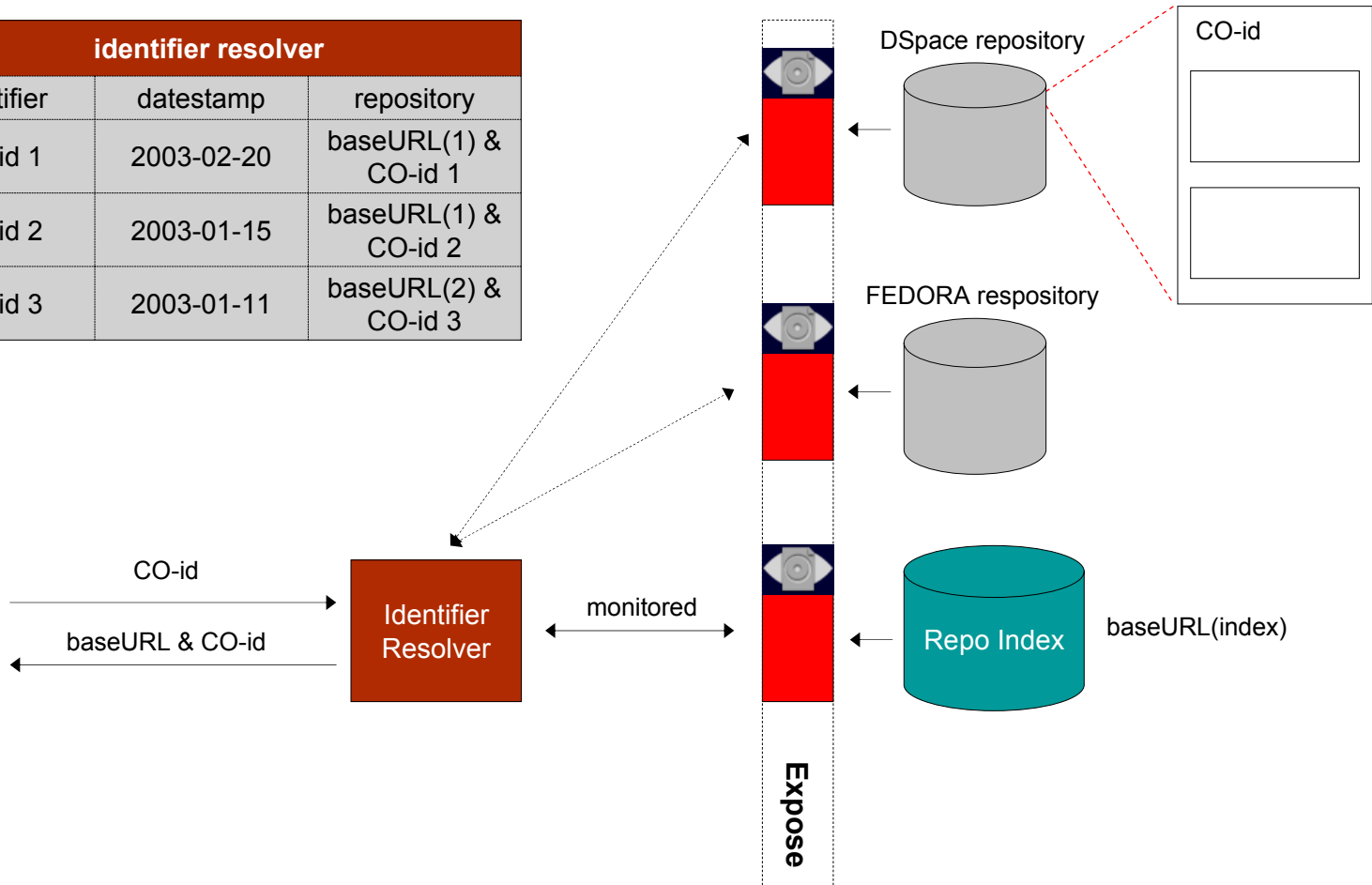


Repository Index: listing OAI-PMH repositories of a federation

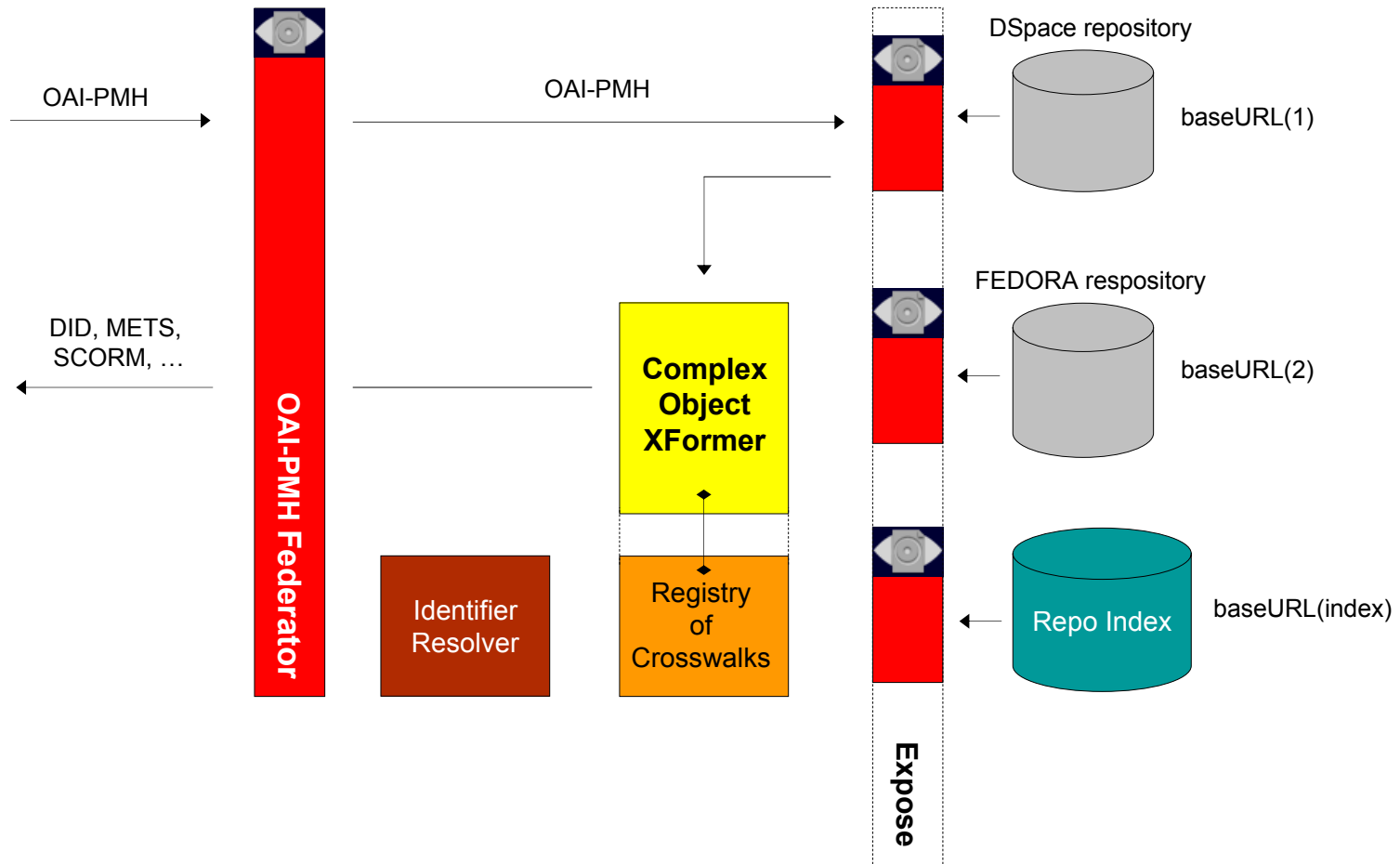


Identifier Resolver: locating COs in the OAI-PMH federation

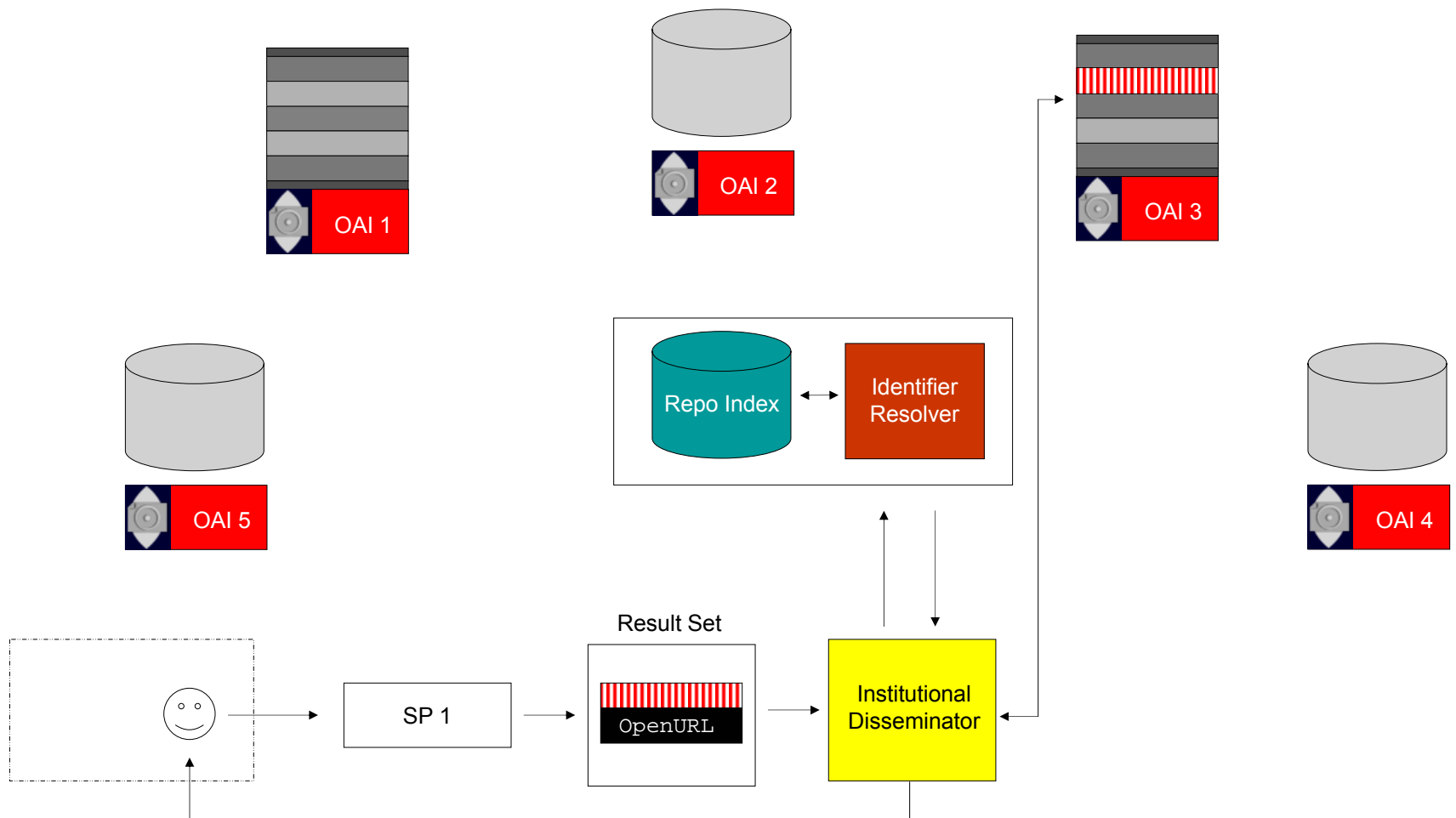
identifier resolver		
identifier	datestamp	repository
CO-id 1	2003-02-20	baseURL(1) & CO-id 1
CO-id 2	2003-01-15	baseURL(1) & CO-id 2
CO-id 3	2003-01-11	baseURL(2) & CO-id 3



Single point of OAI-PMH access to COs in the federation



OpenURL gateway in a distributed architecture



Outline

(1) Motivation

(2) OAI-PMH for content

(3) Example 1 : LANL Repository

(4) Example 2 : mod_oai

(5) Example 3 : DSpace plug-in prototype

(6) Federations of IRs and OAI-PMH

(7) Conclusion

Conclusion: OAI-PMH can be used to harvest content!

- OAI-PMH Advantages:
 - Simple yet powerful protocol.
 - Efficiency through selective & incremental harvesting.
 - Active community. Tools available.
 - Well-established adoption in Digital Libraries, Institutional Repositories, Archives
 - OAI can help (and is very willing to do so):
 - oai-rights – ongoing - how to convey rights in the OAI-PMH framework
 - Could help define - profile(s) of - complex object models that meet the needs
- Complex model advantages:
 - Unambiguous mapping between identifiers and metadata/resources
 - By-reference pointers to resources can be ‘real’ URLs, not hdl, doi, purl
 - Complex models can have simple profiles