

The OAI Object Re-Use & Exchange (ORE) Initiative

Herbert Van de Sompel ⁽¹⁾ & Carl Lagoze ⁽²⁾

⁽¹⁾ Research Library, Los Alamos National Laboratory

⁽²⁾ Information Science, Cornell University



ORE is supported by the Andrew W. Mellon Foundation
with additional support of the National Science Foundation



General information about OAI-ORE



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze



OAI Object Re-Use and Exchange

- OAI-ORE is a new effort conducted under the umbrella of the OAI
- Supported by the Andrew W. Mellon Foundation; additional support from the National Science Foundation
- International effort; October 2006 - September 2008
- <http://www.openarchives.org/ore/>



OAI Object Re-Use and Exchange

- Overall OAI-ORE objectives (elevator pitch):
 - Specify the next level of cross-repository interoperability:
 - Move interoperability from the metadata level to the resource level.
 - Improve the manner in which **machines** can deal with resources.
 - Aim for more optimal and consistent manners:
 - to facilitate discovery of resources,
 - to reference (link to) a resource,
 - to obtain a variety of representations of a resource,
 - to aggregate and disaggregate resources,
 - to re-use (parts of) a resource beyond the boundaries of the holding repository.
 - Establish the technical basis for repositories to become fundamental building blocks of the digital scholarly communication system.



OAI Object Re-Use and Exchange

- OAI-ORE project organization:
 - Coordinators: Carl Lagoze & Herbert Van de Sompel
 - ORE Advisory Committee
 - ORE Technical Committee
 - ORE Liaison Group



OAI Object Re-Use and Exchange

- ORE Advisory Committee:
 - Strategic guidance and outreach
 - Communication via oai-ac listserv and conference calls



ORE Advisory Committee

- Sayeed Choudhury - Johns Hopkins University
- Gregory Crane - Tufts University
- Lorcan Dempsey - OCLC
- Mark Doyle - The American Physical Society
- John Erickson - Hewlett-Packard Laboratories
- Steve Griffin - National Science Foundation
- Robert Hanisch - Space Telescope Science Institute
- Jane Hunter - The University of Queensland (Australia)
- Clifford Lynch - Coalition for Networked Information
- Liz Lyon - UKOLN (UK)
- Peter Murray Rust - University of Cambridge (UK)
- Jim Ostell - National Center for Biotechnology Information
- Sandy Payette - Cornell University
- Robby Robson - Eduworks
- MacKenzie Smith - MIT
- Leo Waaijers - SURF Platform ICT and Research (Netherlands)



OAI Object Re-Use and Exchange

- ORE Technical Committee:
 - Problem statement, scoping, identification of existing technologies, specification, experimentation, etc. (cf OAI-PMH)
 - Communication via in-person meetings, oai-tc listserv, conference calls



ORE Technical Committee

- Les Carr - University of Southampton (UK)
- Leigh Dodds - Ingenta (UK)
- Tim DiLauro - Johns Hopkins University
- Dave Fulker - University Corporation for Atmospheric Research
- Tony Hammond - Nature Publishing Group (UK)
- Richard Jones - Imperial College (UK)
- Peter Murray - OhioLINK
- Michael Nelson - Old Dominion University
- Ray Plante - National Center for Supercomputing Applications
- Andy Powell - Eduserv Foundation (UK)
- Rob Sanderson - University of Liverpool (UK)
- Simeon Warner - Cornell University
- Jeff Young - OCLC



OAI Object Re-Use and Exchange

- ORE Liaison Group:
 - Communication bridge with projects that share ORE objectives
 - Communication via oai-tc listserv, possibly invitations to in-person meetings



ORE Liaison Group

- Tim Cole - UUIC ; for DLF Aquifer
- Rachel Heery - UKOLN ; for the JISC Digital Repository support effort
- Thomas Place - University of Tilburg ; for DARE (soon to be renamed SurfShare)
- Rob Tansley - Google ; for Google and DSpace

- We also have extended invitations to:
 - the EC DRIVER project
 - Microsoft
- Looking into W3C connection.
- Additional liaisons can be added when deemed necessary and/or constructive



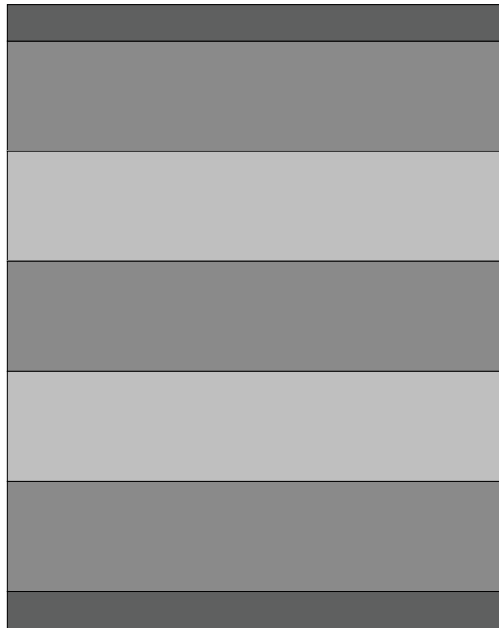
Context



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze



The Repository model

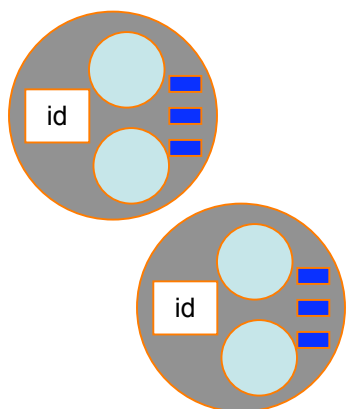


Repository

A scholarly environment
consisting of a variety of
Repositories:

- Institutional repositories
- Discipline-oriented repositories
- Publisher's repositories
- Dataset repositories
- Cultural heritage repositories
- Educational repositories
- ...

Compound Digital Objects



Digital Objects

A scholarly environment in which the resources (units of scholarly communication) are **compound**, consisting of multiple datastreams with a variety of:

- Media types
- Content types
 - Papers,
 - Datasets,
 - simulations,
 - software,
 - dynamic knowledge representations,
 - machine readable chemical structures,
 - Bibliographic metadata, ...

Digital Object use and re-use

- Leverage the value of the resources that become available in those distributed Repositories.
- Make it more straightforward for (parts of) these resources to be used beyond the borders of the hosting Repository:
 - In a variety of services: discovery services, citation managers, blogs, collaborative environments, ...
 - In a variety of scholarly workflows: authoring, citation, consecutive steps in processing a dataset, ...



A sample of perceived problems



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze



Exposing resources to robots

“Are repositories successfully exposing the full-text of articles (the PDF file or whatever) to Google rather than (or as well as) the abstract page?”

(from Andy Powell's [eFoundations](#) blog)



Referencing resources

"Are we consistent in the way we create hypertext links between research papers in repositories?"

(from Andy Powell's eFoundations blog)



Qualifying representations of resources

"Metadata records harvested using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) are often characterized by scarce, inconsistent and ambiguous resource URLs"

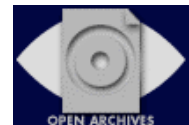
(Robert Chavez et al., from the D-Lib paper about the DLF-Aquifer Asset Actions Experiment)



Re-using (representations of) resources (1)

For example, if I have a page of pdf's of papers and the citations that go with them. I'd like Zotero users to be able to grab what they want ..."

(ts on the Zotero Forum on the topic "How to make Zotero friendly websites?")



Re-using (representations of) resources (2)

So I'd recommend use of Slideshare by anyone involved in developing institutional repositories - if you are going to develop similar services in-house, you'll need to be able to compete with such services, otherwise you may find your users have no interest in using your service.

(from Brian Kelly's UK Web Focus blog)



Richer scholarly workflow involving repositories

"In this infrastructure, repositories are not static components in a scholarly communication system that merely archive digital objects deposited by scholars. Rather, they are the building blocks of a global scholarly communication federation in which each individual digital object can be the starting point for value chains."

(Herbert Van de Sompel, Carl Lagoze et al. in the Pathways D-Lib paper)



Analyzing the problems: back to basics



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze



W3C Web Architecture

URI

```
http://weather.example.com/oaxaca
```



Resource



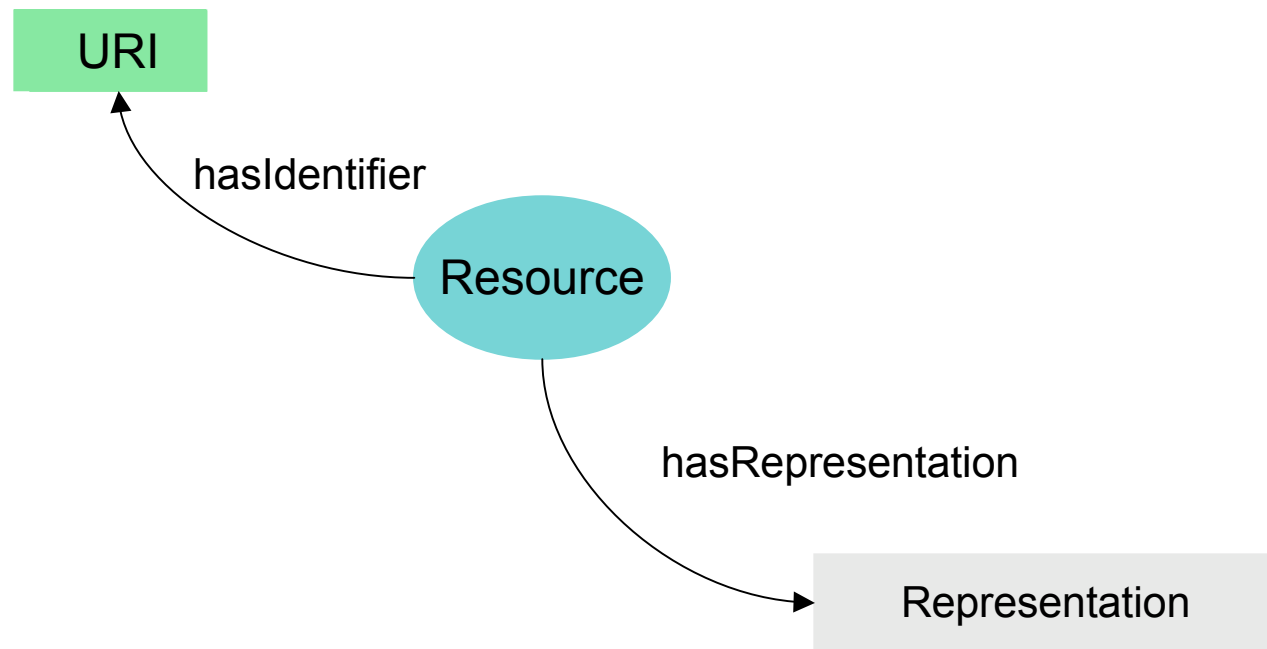
Representation

```
Metadata:  
Content-type:  
application/xhtml+xml  
-----  
Data:  
<html>  
<head>  
<title>5 Day Forecast for  
Oaxaca</title>...  
</html>
```

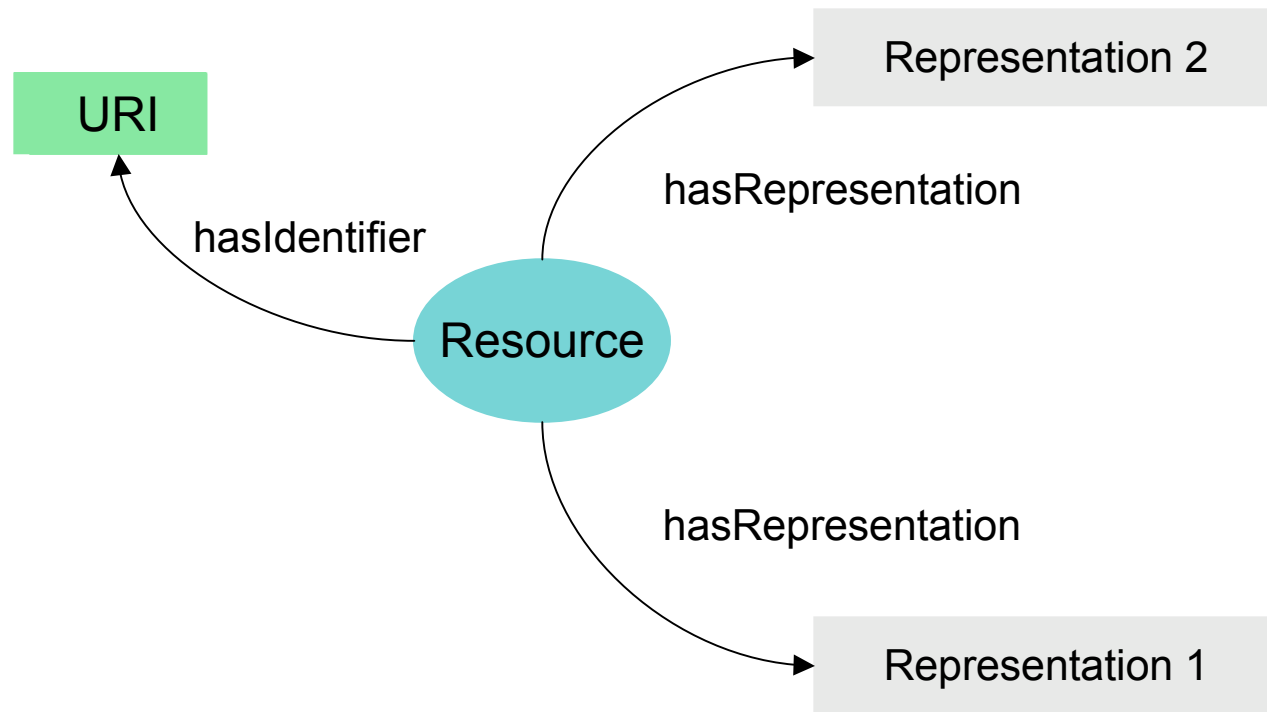
See <http://www.w3.org/TR/webarch/>



W3C Web Architecture



W3C Web Architecture



However ...



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze



However (1)

The Web architecture picture is great, but the notion of Resources and Representations is not fully implemented in the Web.

- Multiple Representations of a single Resource do not share a HTTP URI in the Web
- This is the result of the combination of:
 - Using HTTP URIs as identifiers (because they also locate, which is great)
 - Poor and/or not implemented HTTP Content Negotiation capabilities (an HTTP URI resolves to a very limited set of Representations)



http://static.flickr.com/107/304313960_223e87f4d7_m.jpg

Thumbnail

Resource 2



<http://www.flickr.com/photos/keoshi/304313960/>

Splash page

Resource 1

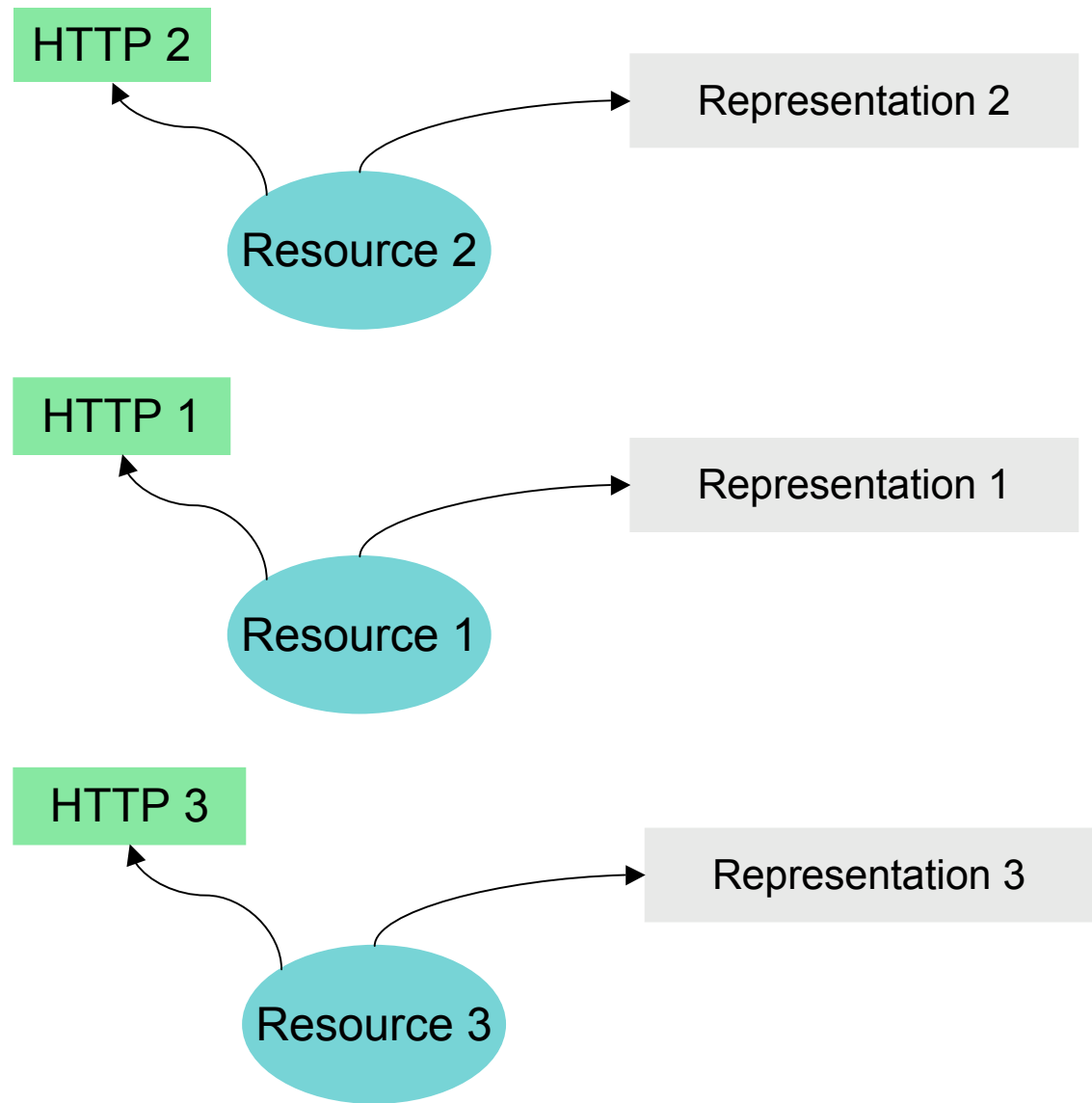


http://static.flickr.com/107/304313960_223e87f4d7.jpg?v=0

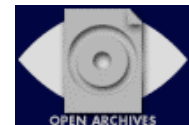
High quality

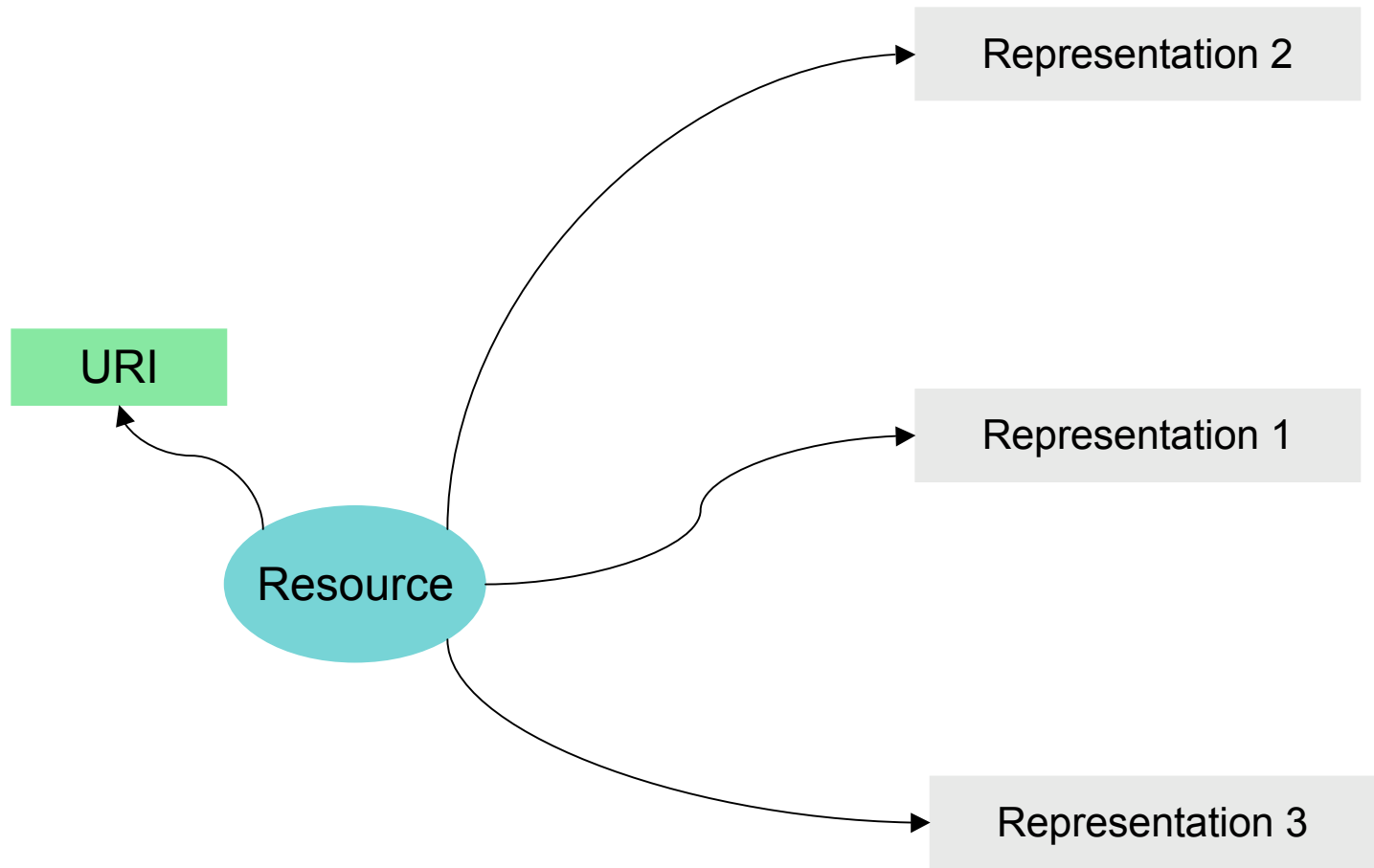
Resource 3





Web with HTTP URIs reality





URI Web architecture



However (2)

The Web architecture picture is great, but it does not natively support the compound objects (multiple datastreams, multiple content types) that become the norm in scholarly communication.

- The Web architecture does not natively support the notion of a Resource that is the aggregation of other Resources (each of which is an aggregation of Representations)
- Note: Even a simple scholarly Resource (e.g. an eprint) is compound in the HTTP Web as it needs to be modeled as one Resource per *view* (metadata record, "full-content", splash-page, ...) of the Resource.



<http://arxiv.org/pdf/astro-ph/0611775>

Resource 2

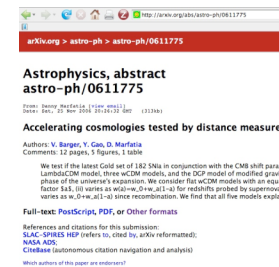
Article in PDF



<http://arxiv.org/abs/astro-ph/0611775>

Resource 1

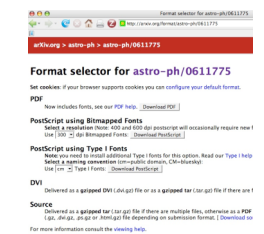
Splash page



<http://arxiv.org/other/astro-ph/0611775>

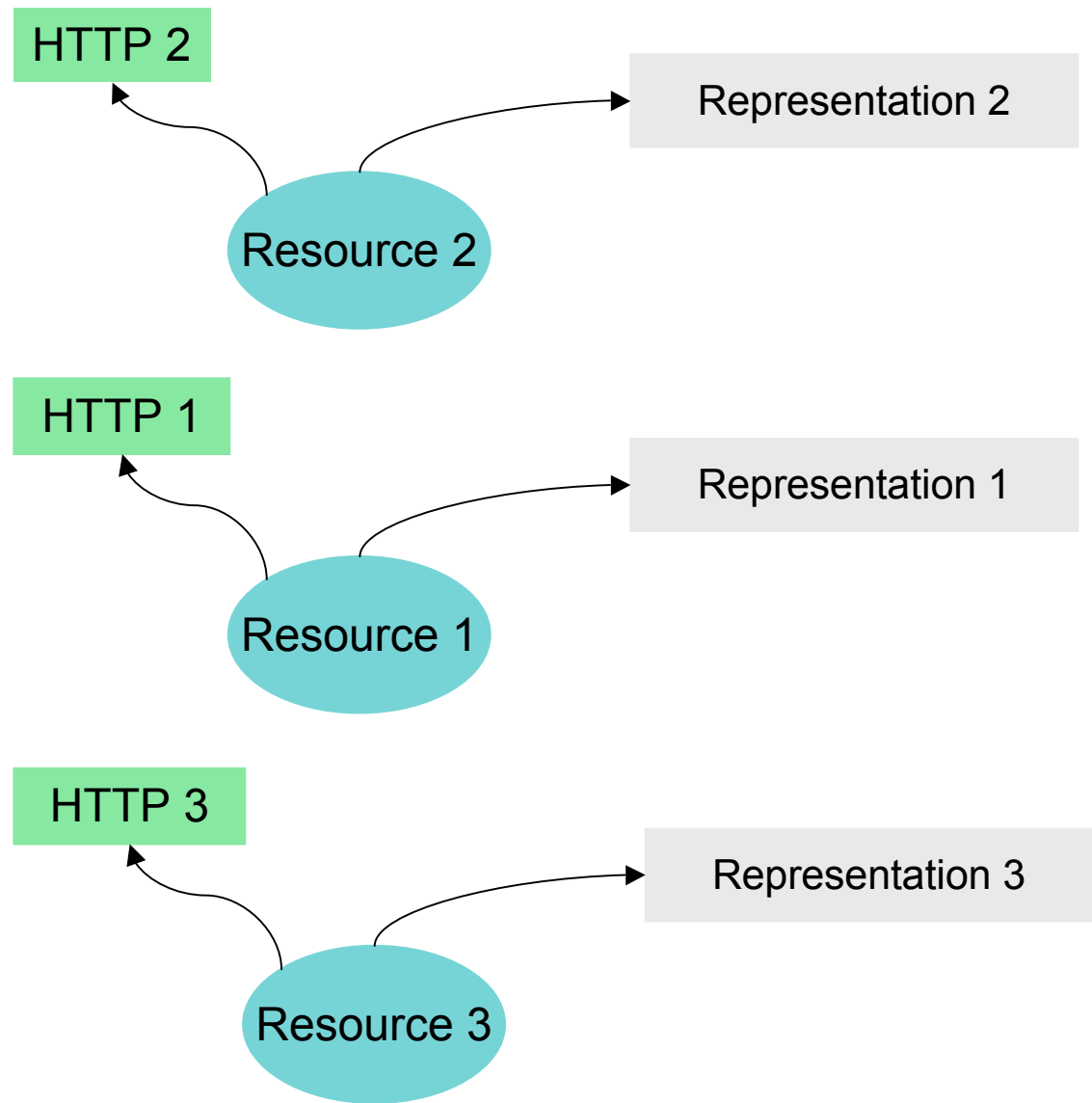
Resource 3

“Other formats” page



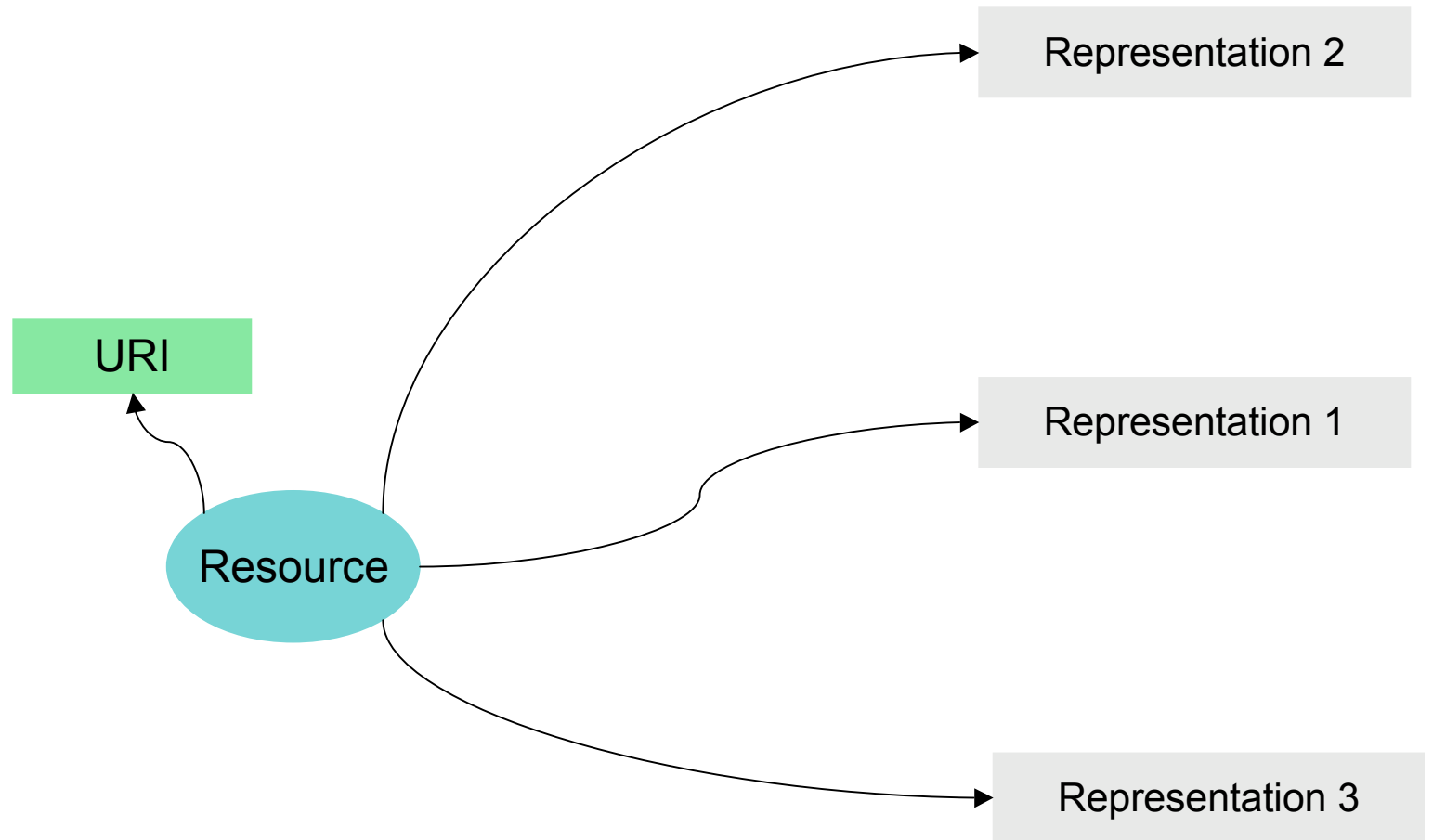
The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze





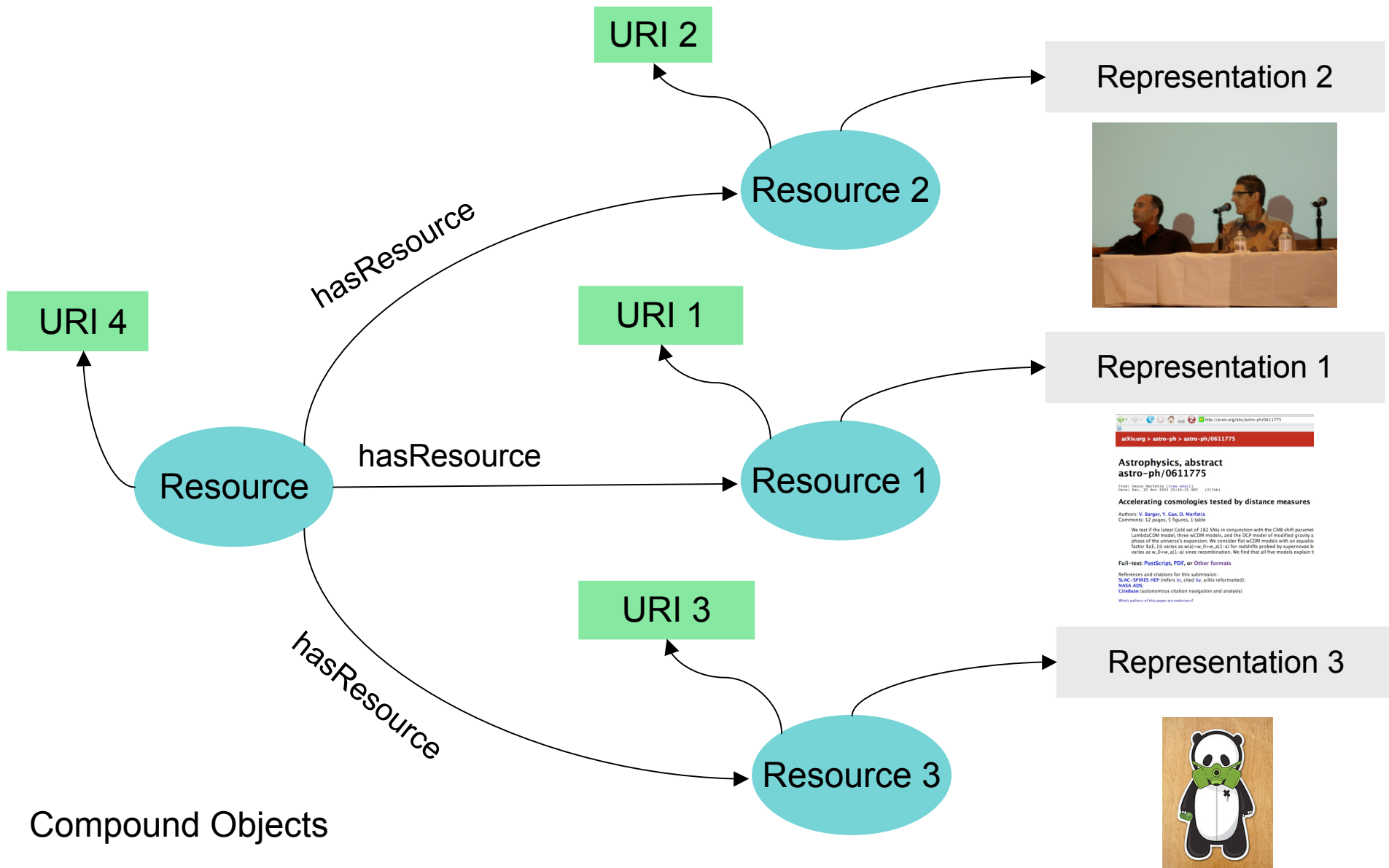
Web with HTTP URIs reality





URI Web architecture



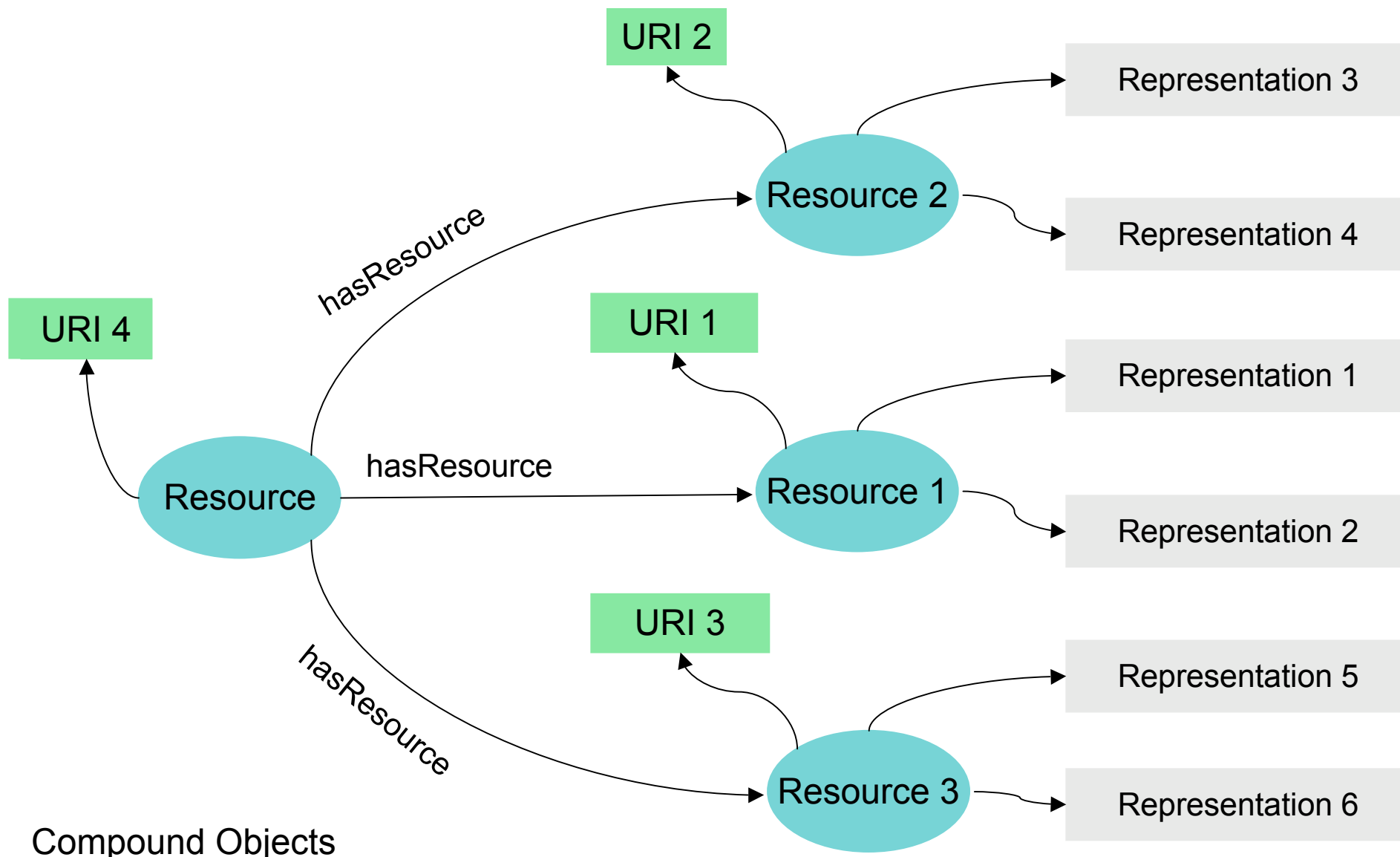


Compound Objects



The OAI Object Re-Use and Exchange Initiative
 CNI Task Force Meeting, Washington, DC, December 4th 2006
 Herbert Van de Sompel & Carl Lagoze





Compound Objects



Back to the perceived problems



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze



Exposing resources to robots

Are repositories successfully exposing the full-text of articles (the PDF file or whatever) to Google rather than (or as well as) the abstract page?

(from Andy Powell's eFoundations blog)

- Facilitate Resource discovery
 - Harvesting a batch of Representations, one per Resource
- Multiple Representation issue
 - Which Representation of the Resource to expose for harvesting?
 - Does this Representation reference (link to) other available Representations, and if so how?



Referencing resources

Are we consistent in the way we create hypertext links between research papers in repositories?

(from Andy Powell's eFoundations blog)

- Multiple Representation issue
- Referencing (linking to) a Resource
 - Which Representation of a Resource to link to?
- Obtaining (other) Representations of the Resource
 - Given the linked-to Representation, how do we obtain others?
 - Both machines and humans follow the reference



Qualifying representations of resources

Metadata records harvested using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) are often characterized by scarce, inconsistent and ambiguous resource URLs

(Robert Chavez et al., from the D-Lib paper about the DLF-Aquifer Asset Actions Experiment)

- Multiple Representation issue:
 - Listing all available Representations of a Resource;
 - Qualifying the nature of each Representation (beyond MIME type)
- Obtaining a variety of Representations of the Resource
- Harvest of a batch of Representations, one per Resource



Re-using (representations of) resources (1)

For example, if I have a page of pdf's of papers and the citations that go with them. I'd like Zotero users to be able to grab what they want ..."

(ts on the Zotero Forum on the topic "How to make Zotero friendly websites?")

- Multiple Representation issue:
 - Listing of all available Representations of a Resource;
 - Qualifying the nature of each Representation (beyond MIME type)
- Service discovery: discovering how to Obtain the listing of all available Representations



Re-using (representations of) resources (2)

So I'd recommend use of Slideshare by anyone involved in developing institutional repositories - if you are going to develop similar services in-house, you'll need to be able to compete with such services, otherwise you may find your users have no interest in using your service.

(from Brian Kelly's UK Web Focus blog)

- Putting (a Representation of) a Resource from repository A to repository B
- Which Representation of the Resource to put?



Richer scholarly workflow involving repositories

"In this infrastructure, repositories are not static components in a scholarly communication system that merely archive digital objects deposited by scholars. Rather, they are the building blocks of a global scholarly communication federation in which each individual digital object can be the starting point for value chains."

(Herbert Van de Sompel, Carl Lagoze et al. in the Pathways D-Lib paper)

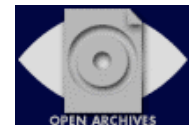
- Augmented cross-repository interoperability:
 - shared compound object model, shared representation of compound objects
 - shared repository interfaces for core functionality, i.e. harvest, reference, obtain, put
- Tracking/expressing relationships, i.e lineage, in workflows



Thoughts towards a solution



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze



Thought (0)

- Whatever we do needs to be embedded in the Web; we are not creating a parallel universe.
- Wherever possible and appropriate repurpose existing technologies, possibly with qualifications/extensions/modifications if required.
- Keep it simple; as simple as possible.



Thought (1)

- Need a foundation that allows us to consistently do network *transactions* for (parts of) Resources that are:
 - Aggregations of Representations
 - Aggregations of Resources



Thought (2)

- Transactions typically reside under the following categories:
 - To facilitate discovery: Harvest of a batch of Representations, one per Resource
 - To facilitate citation: Referencing (a Representation of) a Resource
 - To facilitate access: Obtain Representations of a Resource
 - To facilitate re-use: Put (a Representation of) a Resource



Thought (3)

- It would help to have a shared model/format to represent Resources that are:
 - Aggregations of Representations
 - Aggregations of Resources
- The Canonical Representation Format (CaRF): Format to express a manifest of all available Representations (and Resources) for a Resource
- Fleshing out the CaRF is probably the core effort of OAI-ORE



Thought (3), continued

- A Representation of a Resource compliant with the CaRF is a Canonical Representation (CaR):
- Not yet another metadata format to describe the Resource, but a manifest of all Representations of the Resource (including the metadata Representations):
 - Typically machine generated
 - Lists access points to Representations
 - It must be possible to unambiguously reference this CaR
 - Should allow to qualify Representations re MIME type, content type, ...
 - Should allow to express relationships, e.g. hierarchy, lineage, ...
 - Should be possible to merge multiple CaRs



Thought (3), continued

- A CaR is obtainable for each ORE Resource
- Should be possible to ask a Repository for a CaR for a specified Resource
- Should be possible to find the CaR for a Resource on the basis of a found Representation of the Resource
- The CaR must be transportable via various protocols that can play a role in the realm of harvest, reference, obtain, put



Thought (3), continued

- The CaR/CaRF world is not unexplored:
 - ATOM
 - DIDL
 - DLF Aquifer Asset Action Package
 - METS
 - Pathways Core
 - ...



Thought (4)

- Think of the *transactions* Harvest, Obtain, Reference, Put (machine oriented) in terms of CaRS
- Per *transaction* category:
 - Select, evolve, define one or more technologies
 - Profile them in terms of CaRS
- These worlds are not unexplored:
 - Harvest: Google SiteMaps, OAI-PMH, RSS, ...
 - Obtain: OpenURL, unAPI, ...
 - Put: ATOM publishing protocol, HTTP PUT, request for upload, WebDAV, ...



Harvest

RSS
OAI PMH
Google
SHERP

Reference
obtain

OpenURL
unAPI

Put

ATOM
push.
request
for Put

CaRF

ATOM, DIDC, METS, Aquifer, --



What's Next?



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze



What's Next?

- First meeting of OAI-ORE Technical Committee, Columbia University, January 11th and 12th 2007
- Goals:
 - Reach shared problem statement
 - Reach shared scoping agreement
 - Identify relevant technologies
 - Identify work items for January-June 2007
 - Discuss public communication approach



Questions



The OAI Object Re-Use and Exchange Initiative
CNI Task Force Meeting, Washington, DC, December 4th 2006
Herbert Van de Sompel & Carl Lagoze

