Interoperability, Metadata, and Complex Objects

Carl Lagoze
Cornell Information Science

Herbert Van de Sompel
Los Alamos Research Library

March 16, 2007
CUL Metadata Working Group

The cast of collaborators and supporters:

- Pathways project (NSF IIS-0430906)
  - Cornell University (Carl Lagoze, Sandy Payette, Simeon Warner)
  - Los Alamos National Laboratory (Herbert Van de Sompel).
- Fedora Open Source Repository Project (Mellon)
  - Cornell University (Sandy Payette)
  - University of Virginia (Thorton Staples)
- OAI Object Reuse and Exchange OAI-ORE (Mellon)
  - Cornell University (Carl Lagoze)
  - Los Alamos National Laboratory (Herbert Van de Sompel)
  - and a larger community….

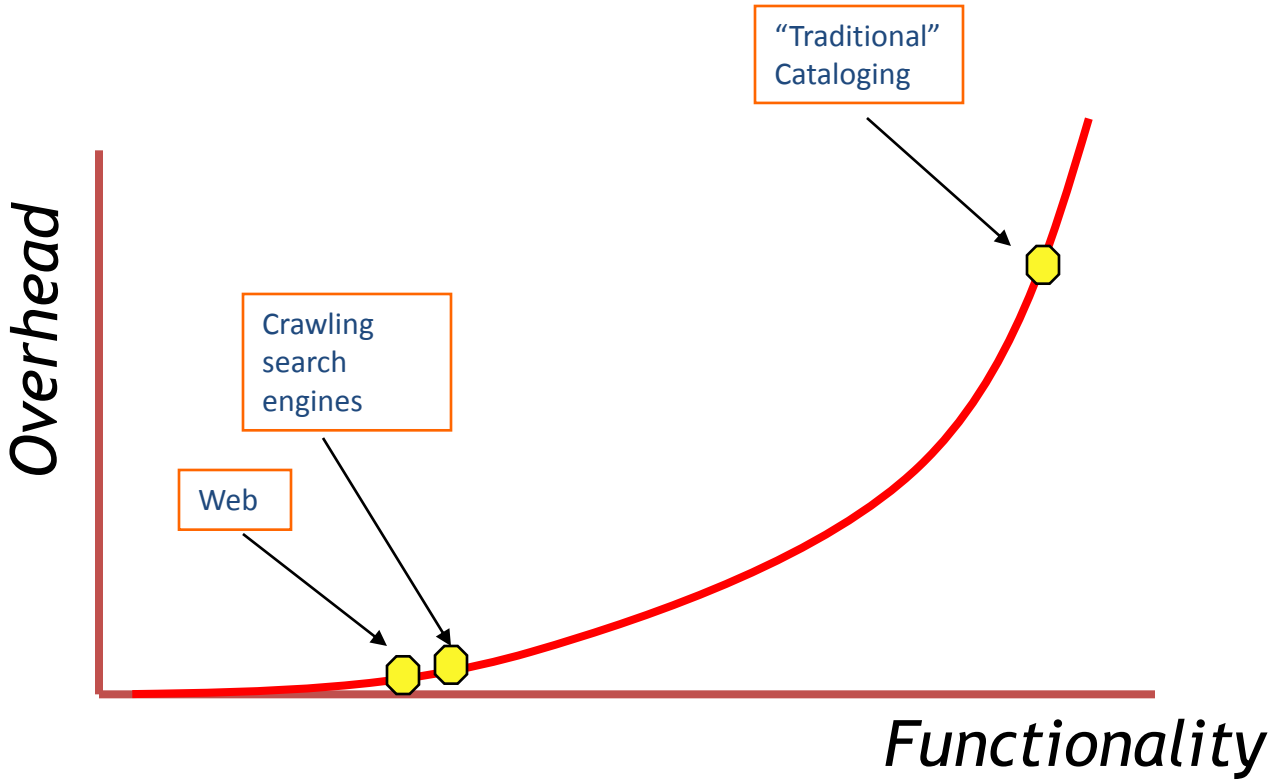
Interoperability

- The ability to create applications that provide seamless integration of distributed heterogeneous systems and resources.
- Interoperability has many dimensions – e.g., syntax, semantics – and approaches – e.g., shared protocols, shared models, mediation, etc.

Interoperability Research at Cornell Information Science

- Dienst
  - Meta-search
  - Metadata vocabulary
  - Document model
  - Protocol
- Dublin Core
  - Metadata vocabulary
- ABC
  - Ontology
- Fedora
  - Document model
  - API
  - Support for multiple metadata vocabularies
- OAI-PMH
  - Metadata-centric Protocol
  - Support for multiple metadata vocabularies
- OAI-ORE
  - Resource-centric
  - Document model and services

The Challenge:
Keeping it Simple and Affecting Functionality



Recommended Reading

D-Lib Magazine
January/February 2007

Volume 13 Number 1/2

ISSN 1082-9873

Resource Description and Access (RDA)

Cataloging Rules for the 20th Century

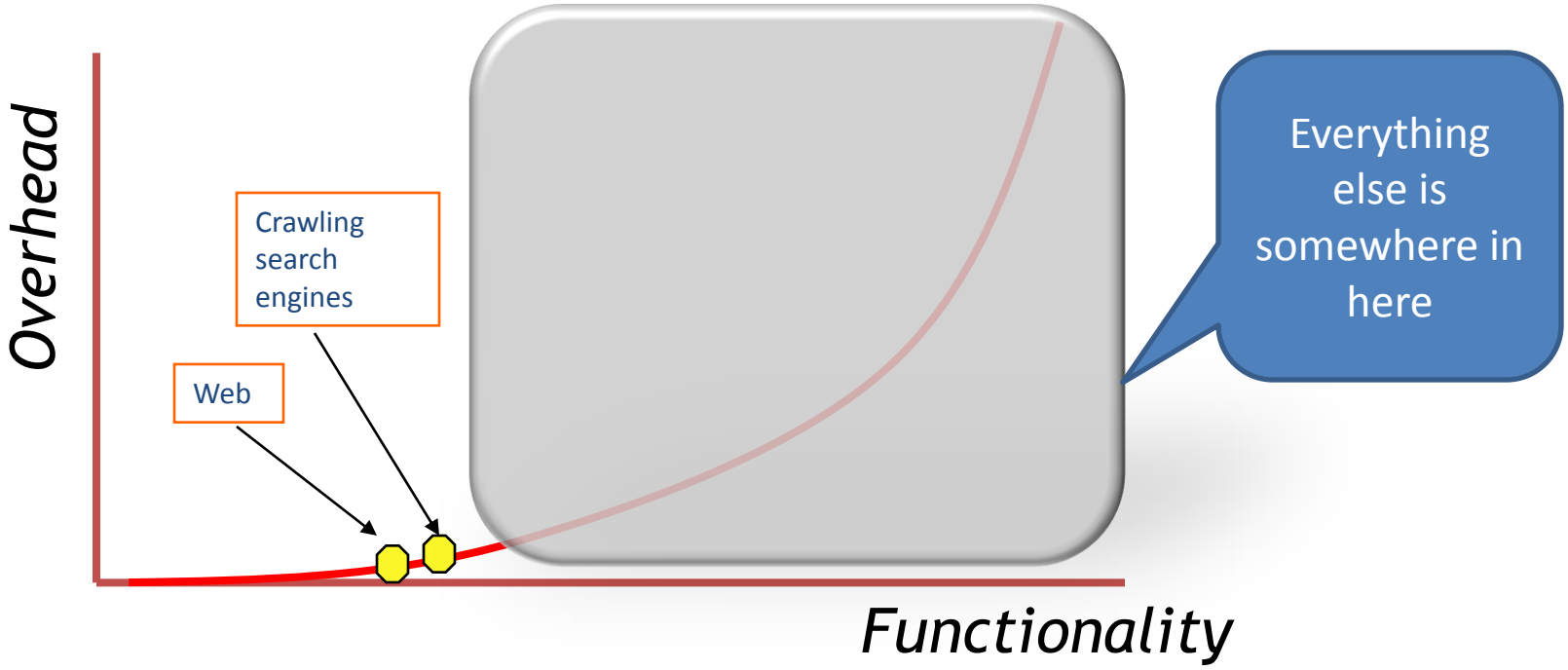
Karen Coyle
kcoyle.net
<kcoyle@kcoyle.net>

Diane Hillmann
Cornell University
<dih1@cornell.edu>

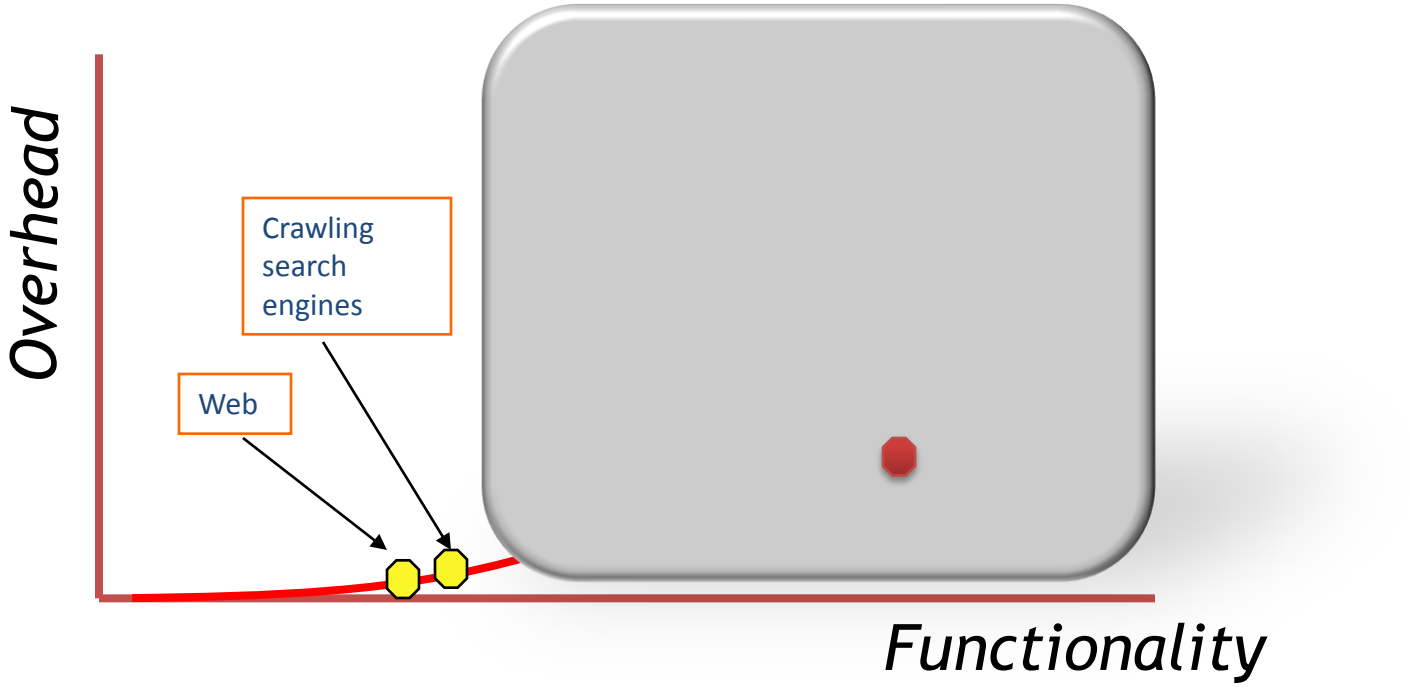
Coyle and Hillmann

Changes in the context in which libraries function have brought the library and its catalog to a crisis point. Today the development of computer technology and electronic document production presents a significantly different challenge than libraries had only fifty years ago, a time when information resources and the libraries that held them were still rooted in the era of books and periodicals, and the card catalog was the entry point to the library's physical holdings. The effect of computers and networks of information resources on the mission of libraries is still being debated, but the very existence of libraries in the future rides on their ability to respond to today's – and tomorrow's – technology.

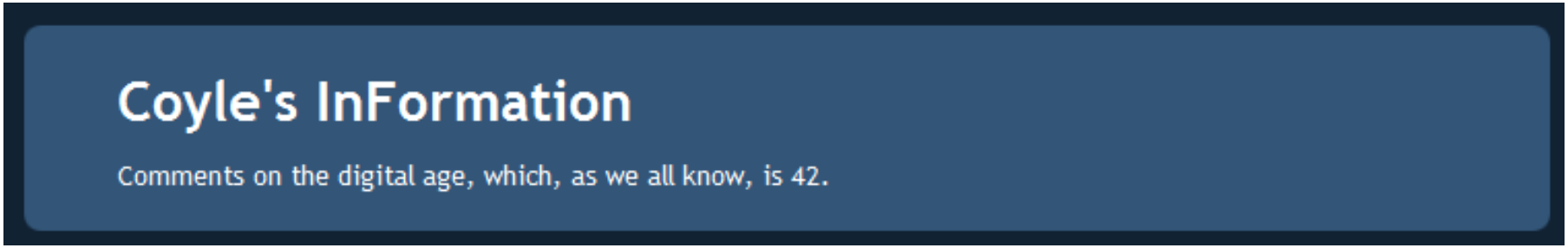
The Challenge:
Keeping it Simple and Affecting Functionality



Cost/Functionality of Meta-Searching



Recommended Reading



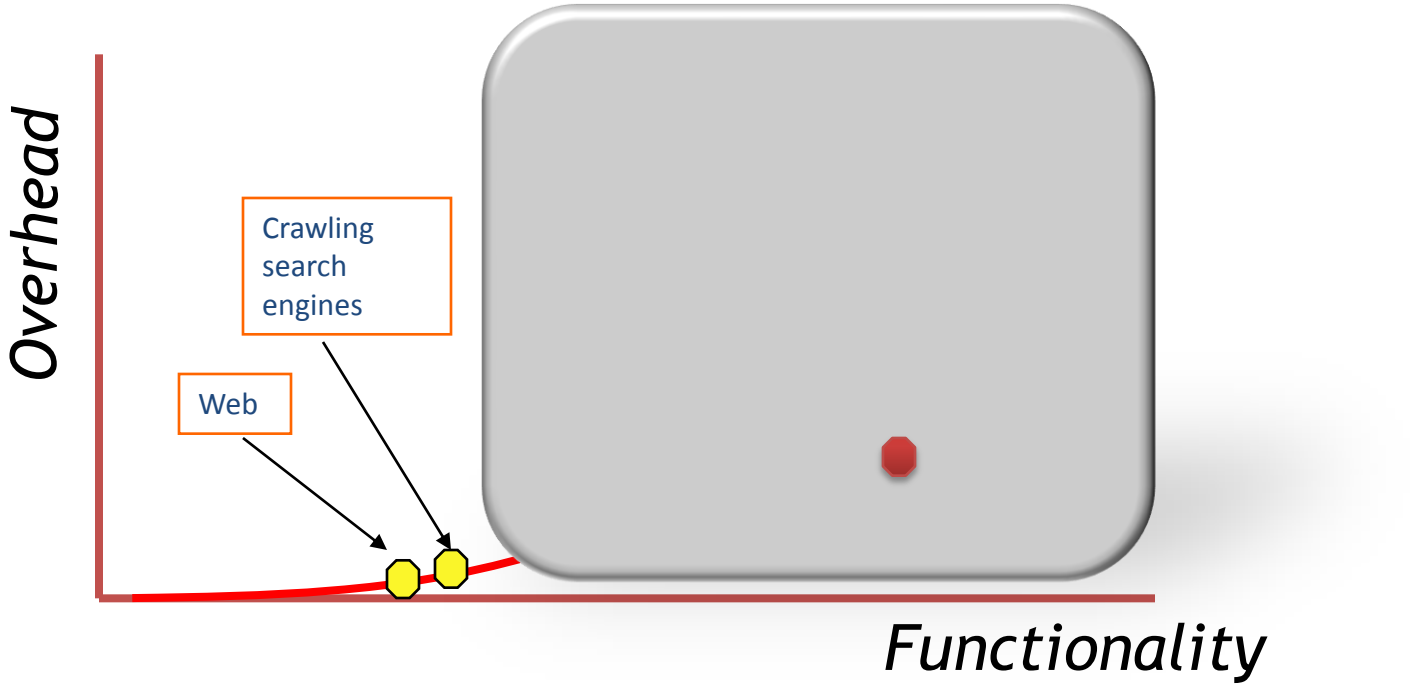
http://kcoyle.blogspot.com/

Dan Clancy from Google in Coyle Blog

Metasearch of a diverse sources is a dead-end.
- ranked lists are hard to merge, even when you know the ranking functions
- can't do whole-corpus analyses
- speed is defined by the slowest search

March 9, 2007

Cost/Functionality of “Simple” Metadata: e.g., Dublin Core



Recommended Reading

Journal of the American Society for Information Science and Technology

What is RSS?

Volume 58, Issue 5 , Pages 613 - 628

Published Online: 2 Feb 2007

Save Title to My Profile Set E-Mail Alert

Go to the homepage for this journal to access trials, sample copies, editorial and author information, news, and more.

Save Article to My Profile Download Citation < Previou

Abstract | References | Full Text: HTML, PDF (329k) | Related Articles | Citation Tracking

Research Article

Does topic metadata help with Web search?

David Hawking [1], Justin Zobel [2]

[1]CSIRO ICT Centre, Canberra ACT 2601, Australia

[2]School of Computer Science and Information Technology, RMIT, Melbourne, Australia

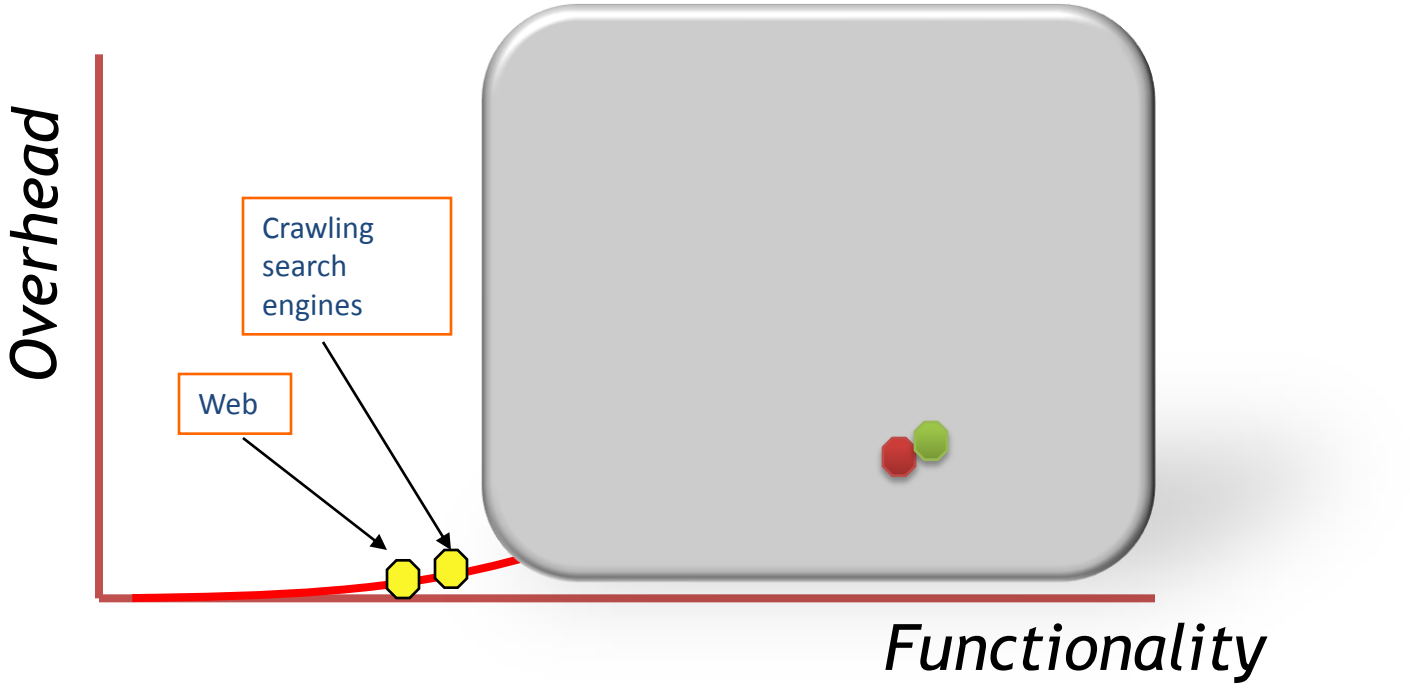
email: David Hawking (david.hawking@acm.org) Justin Zobel (jz@cs.rmit.edu.au)

Hawking & Zobel

Using a large institutional Web site we have explored the value of topic metadata in search. We found little evidence that metadata was of value for queries extracted from the query log for that site, even when the index was restricted to the central, well-managed site.

We found that topic metadata was of limited value for common queries, even when only pages with metadata were considered; there was mismatch between query and metadata vocabulary; and much of the metadata was inaccurate or misleading.

Cost/Functionality of Metadata Harvesting: e.g., OAI-PMH



“Recommended” Reading

Metadata aggregation and “automated digital libraries”: A retrospective on the NSDL experience

Carl Lagoze
Cornell Information Science
301 College Ave.
Ithaca, NY 14850
+1-607-255-6046
lagoze@cs.cornell.edu

Dean Krafft, Tim Cornwell,
Naomi Dushay, Dean Eckstrom,
Cornell Information Science
301 College Ave.
Ithaca, NY 14850
+1-607-255-5925
{dean,cornwell,naomi,eckstrom}
@cs.cornell.edu

John Saylor
Cornell University Library.
Ithaca, NY 14853
+1-607-255-4134
jms1@cornell.edu

JCDL 2006

Lagoze, et. al.

Over the last three years the NSDL CI team has learned that a seemingly modest architecture based on metadata harvesting is surprisingly difficult to manage in a large-scale implementation. The administrative difficulties result from a combination of provider difficulties with OAI-PMH and Dublin Core, the complexities in consistent handling of multiple metadata feeds over a large number of iterations, and the limitations of metadata quality remediation.

“Lagoze’s First Principle”

Be skeptical of false prophets of interoperability

Motivation: Scholarship is Changing

- Influenced by:
  - High performance computing and connectivity
  - Peta-scale data storage
  - Advanced data mining and data storage
  - Web services, Web 2.0
  - Open Access movement
- Evolution towards:
  - Highly collaborative
  - Network-based
  - Data-driven
- Visible in Science & Engineering but also in humanities and social sciences
- And, there are increasing links between these formerly separated fields (benefits of having “everything” online)

And changes in the way that scholarship is disseminated

- E-print repositories – arXiv, Cogprints
- Institutional repositories – DSpace, FEDORA, ePrints.org
- Publication repositories – PubMed Central
- Data Repositories – NVO, NCBI
- Interoperability architecture – DCMI, OAI-PMH
- Networked discovery services – Google Scholar, CiteSeer

- But many of these are changes in form rather than in nature
- Or, at best, not solutions that generalize across disciplines

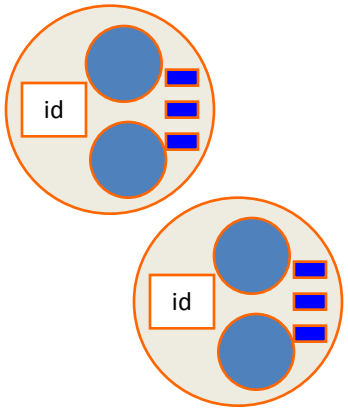
Setting More Ambitious Goals

- In many cases we've only created an electronic equivalent of the paper-based system.
- While 'open access' is important, it should not be our only focus.
- The networked environment provides opportunities for more radical changes.
  - Expose component products and process
  - Allow components to move across multiple workflows
  - Promote recombination, refactoring, and transformation of information – Mash-ups
  - Transform repositories/databases from passive storage to active building blocks for higher level services

Scenarios

- Specialized search engines – beyond just text scraping
- Overlay journals
- Evidential rather than bibliographic citation
- Dataset exposure and reuse

Support complex content



Digital Objects

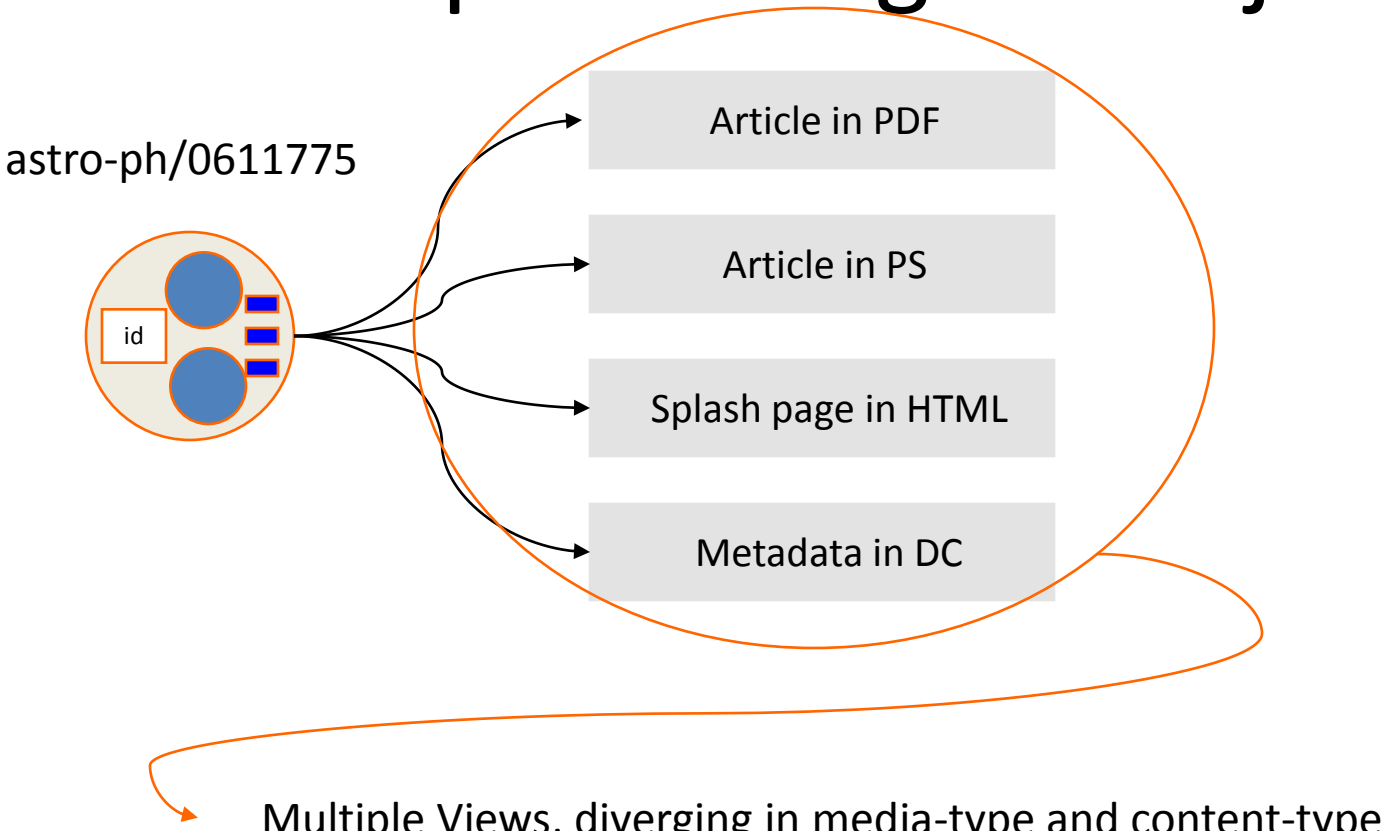
Digital content with multiple components varying on:

- **Content (semantic) types** including:
  - Text
  - Datasets
  - Simulations
  - Software
  - Dynamic knowledge representations
  - Machine readable chemical structures
  - Bibliographic and other types of metadata
- **Media types** including
  - IANA registered MIME types
  - Other type registries such as GDFR
- **Network locations** including content from:
  - Institutional repositories
  - Scientific data repositories
  - Social networking sites
  - General web
- **Relationships** including:
  - Lineage
  - Versions
  - Derivations

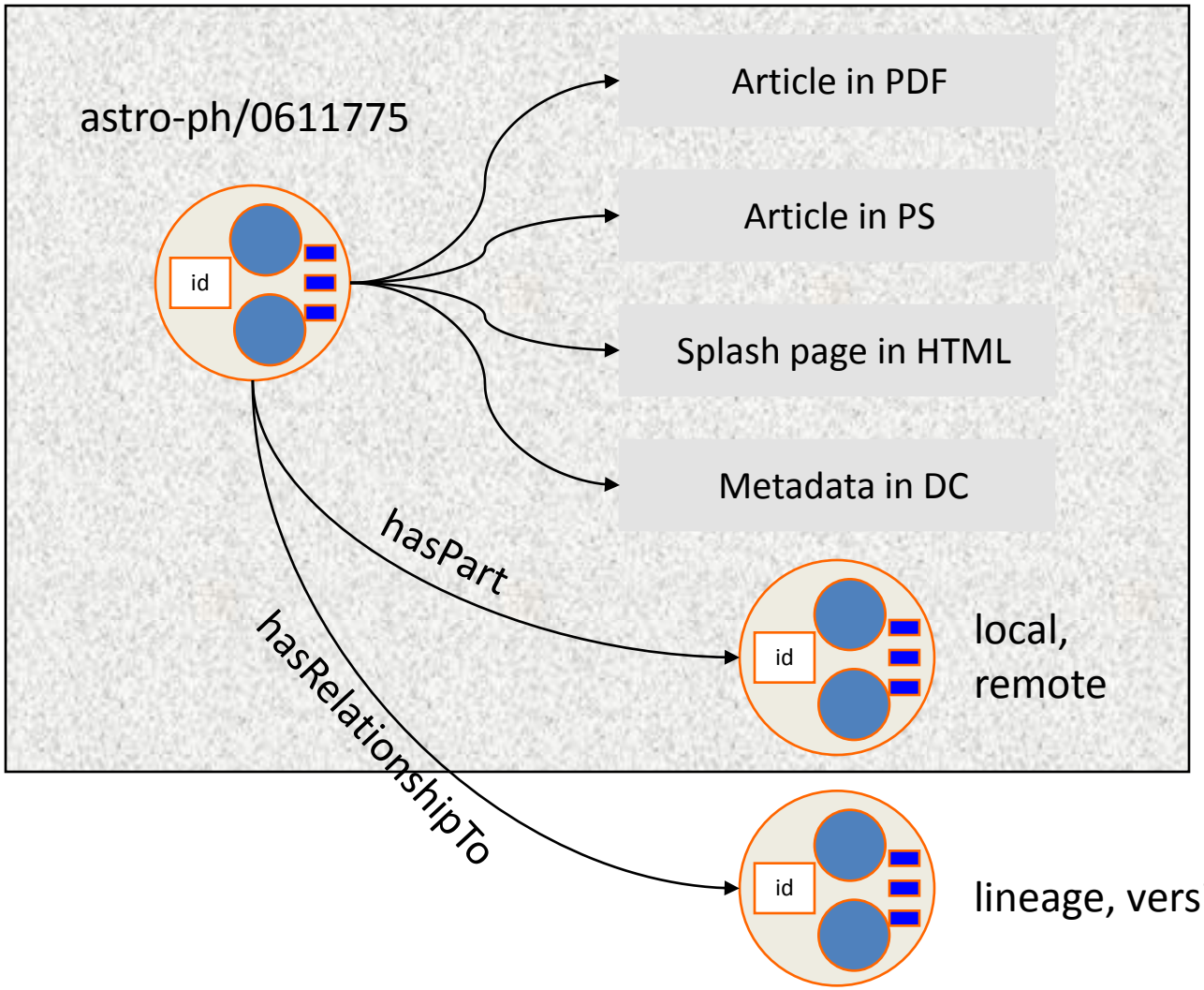
Examples of Compound Digital Objects

- arXiv paper with different disseminations
- An issue of an overlay journal built from distributed ePrints
- eScience publication combining text, data, simulations
- eHumanities resource combining primary and derived content

Compound Digital Object

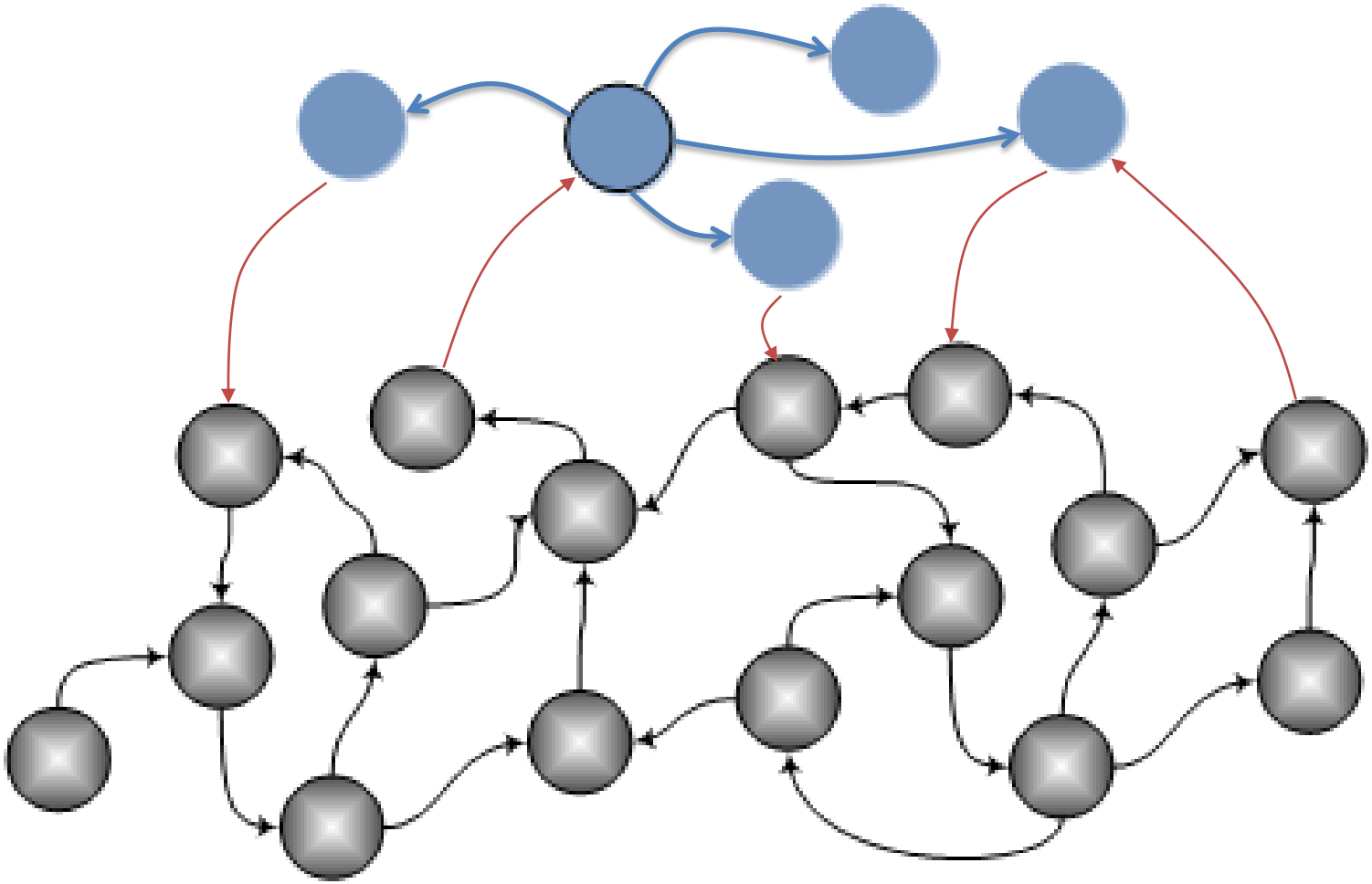


More complexity …



boundary, logical unit

Expose structure and relationships of these objects



http://www.openarchives.org/ore

Open Archives Initiative
Object Reuse and Exchange

Home | Projects | Specifications | Community | About OAI

Open Archives Initiative -> ORE

Exchanging Information about Digital Objects

ORE will develop specifications that allow distributed repositories to exchange information about their constituent digital objects. These specifications will include approaches for representing digital objects and repository services that facilitate access and ingest of these representations. The specifications will enable a new generation of cross-repository services that leverage the intrinsic value of digital objects beyond the borders of hosting repositories.

The OAI-ORE Community

- Executive Committee
- Advisory Committee
- Technical Committee
- Liaison Group

Contact us

- ore@openarchives.org

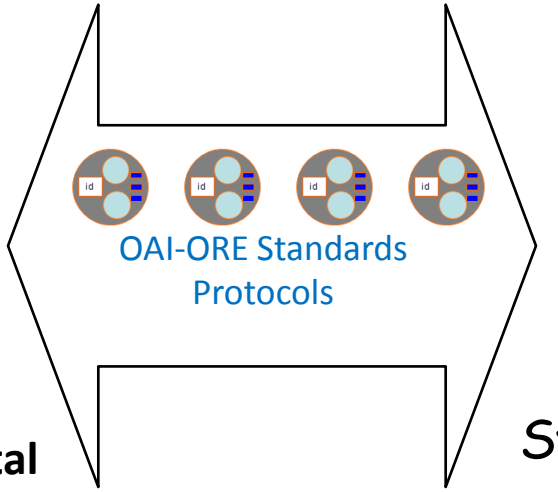
OAI-ORE Resources

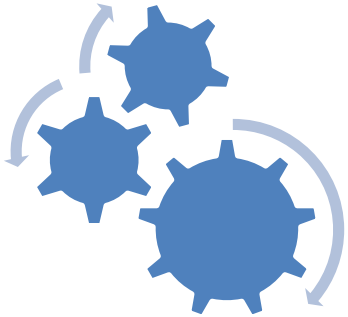
- Report of the January 2007 ORE-TC Meeting
  A detailed report of the results of the meeting of OAI-ORE Technical Committee describing features and requirements of the ORE model and its context in the Web Architecture.
- Open Repositories 2007
  A presentation describing OAI-ORE and progress based on the January 2007 ORE Technical Committee Meeting.
- CNI 2006 ORE Project Briefing
  A presentation describing motivations, requirements, and preliminary thinking on the OAI-ORE project.
- Proposal for funding to the Mellon Foundation
  Details the plan for work developing OAI-ORE specifications over the two-year period beginning October 2006
- Press Release
  Announces the Mellon grant and summarizes the two-year work plan
- Augmenting interoperability across scholarly repositories
  An April 20-21 2006 meeting sponsored by Microsoft, CNI, DLF, and JISC that laid he foundation for OAI-ORE
- Rethinking Scholarly Communication: Building the System that Scholars Deserve
  An opinion article in D-lib that describes a model of a scholarly communication system that interoperable repositories could provide

OAI Object Re-Use and Exchange

- Develop, identify, and profile extensible standards and protocols to allow *repositories, agents, and services to interoperate* in the context of *use and reuse of compound digital objects* beyond the boundaries of the holding repositories.

- Aim for more effective and consistent ways:
  - to facilitate discovery of these objects,
  - to reference (link to) these objects (and parts thereof),
  - to obtain a variety of disseminations of these objects,
  - to aggregate and disaggregate these objects,
  - Enable processing by automated agents





Systems that manage digital objects

- Institutional repositories
- Research-group and managed personal (ePortfolio) repositories
- Discipline-oriented repositories
- Publisher repositories
- Dataset repositories
- Cultural heritage repositories
- Learning object repositories
- Digitized book and manuscript collections

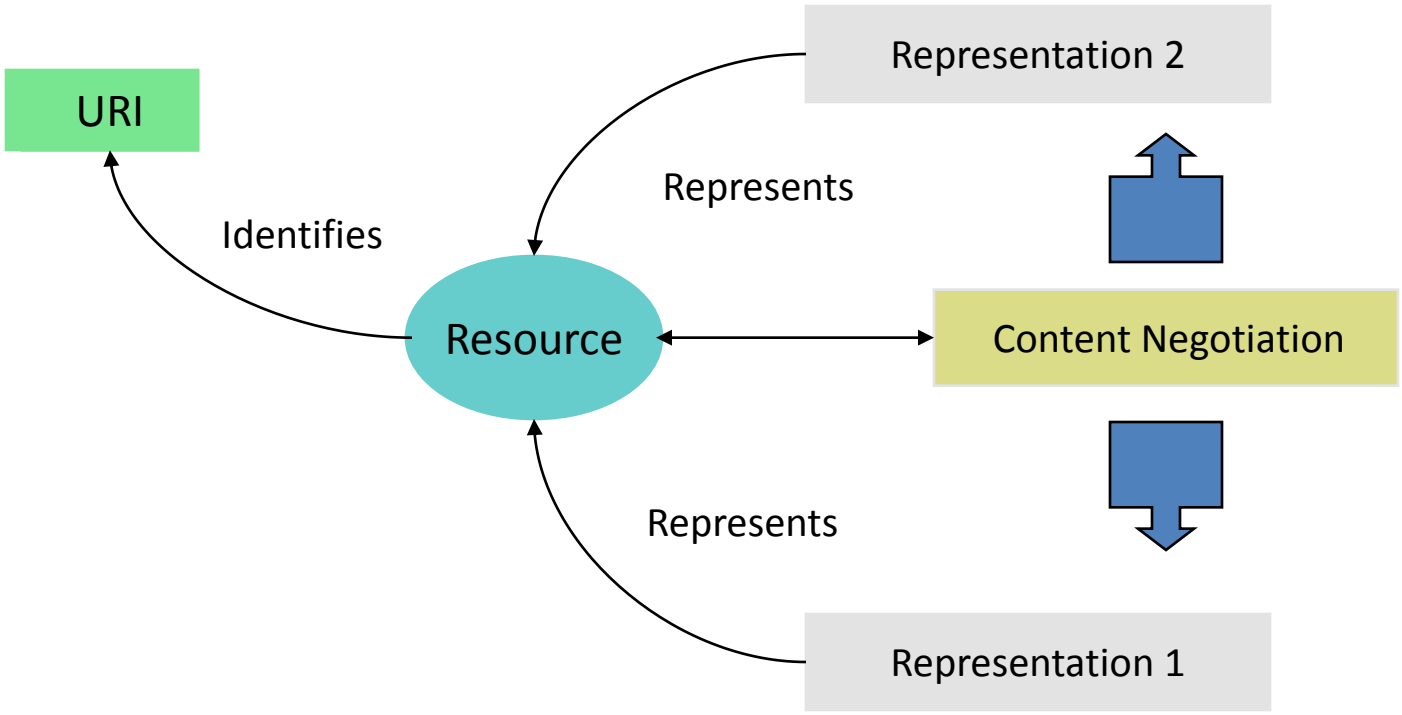
Systems that leverage managed digital objects

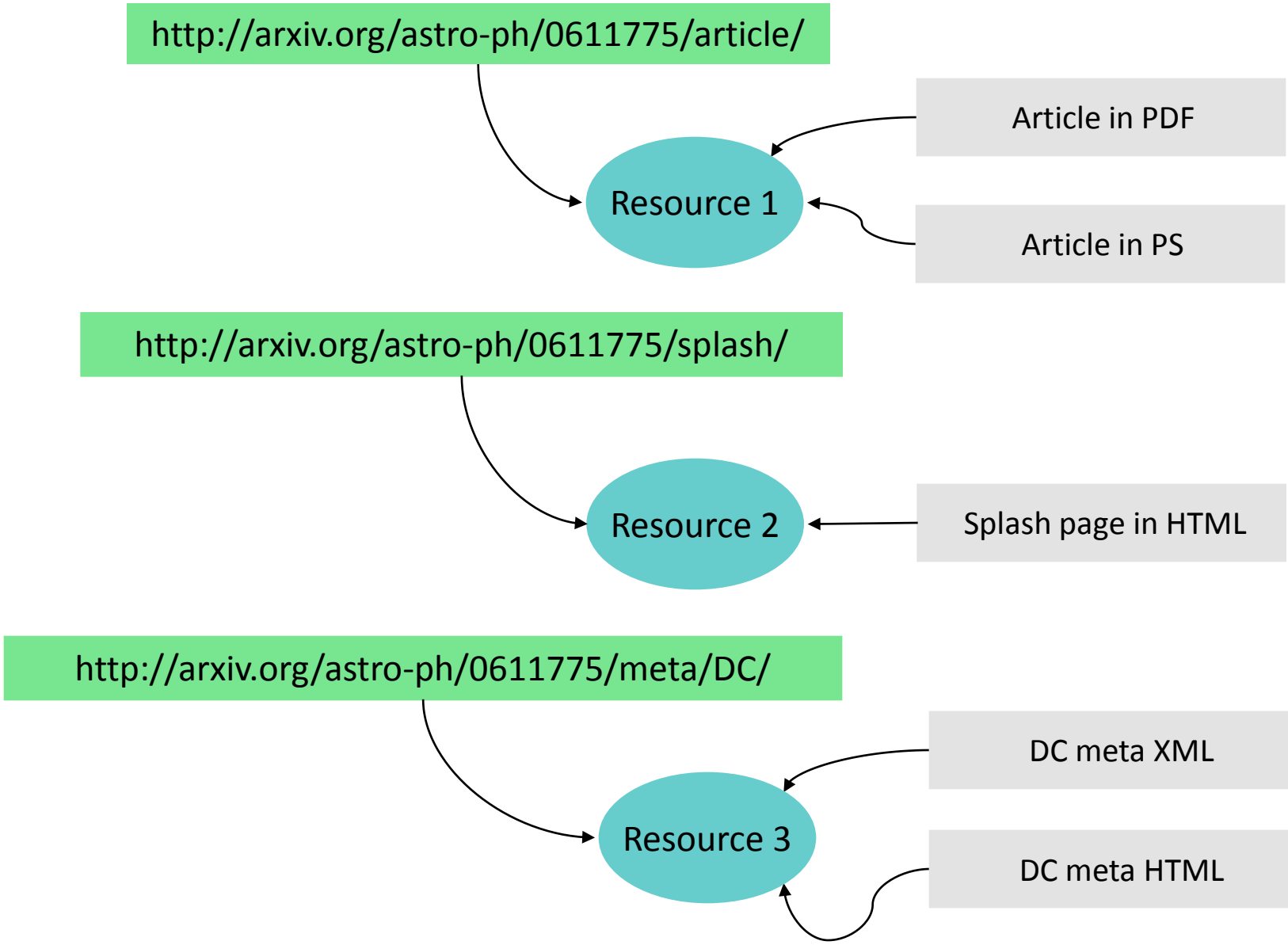
- Search engines
- Authoring tools
- Citation management
- Collaborative environments
- Social network applications
- Data/Text mining applications
- Graph analysis tools
- Preservation services
- Workflow tools
- Report generation tools

Working with the web architecture

- **Whatever we do it must be congruent with the web architecture**
  - Use existing capabilities where they are appropriate
  - Cleanly layer capabilities meeting the needs of our problem space
- Provide the infrastructure for web-based information systems that exploit/enhance and therefore overlay on the existing web.

W3C Web Architecture

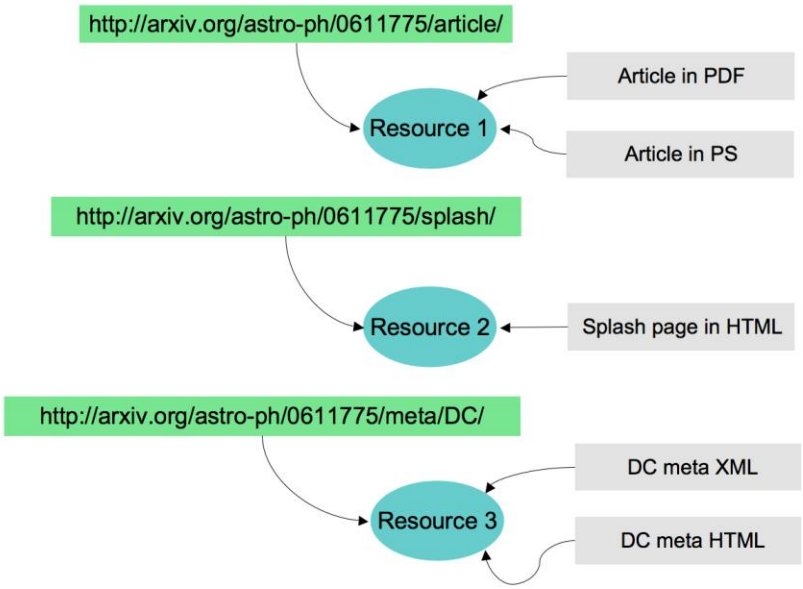



http://arxiv.org/astro-ph/0611775/article/
Article in PDF
Resource 1
Article in PS
http://arxiv.org/astro-ph/0611775/splash/
Resource 2
Splash page in HTML
http://arxiv.org/astro-ph/0611775/meta/DC/
DC meta XML
Resource 3
DC meta HTML

Compound Digital Object mapped to the Web

"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?"

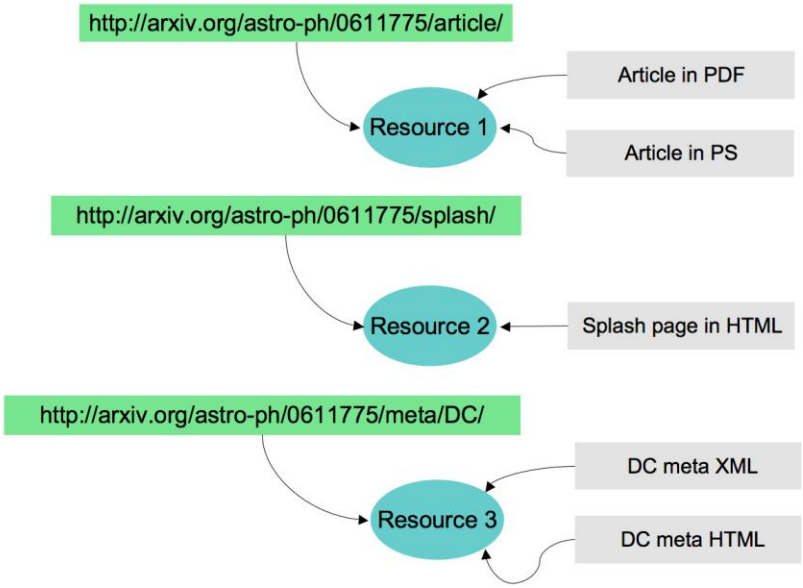
- Discovery: How does Google find all these resources that originate from the same digital object?
- Boundary: How does Google know these resources originate in the same digital object?



Compound Digital Object mapped to the Web

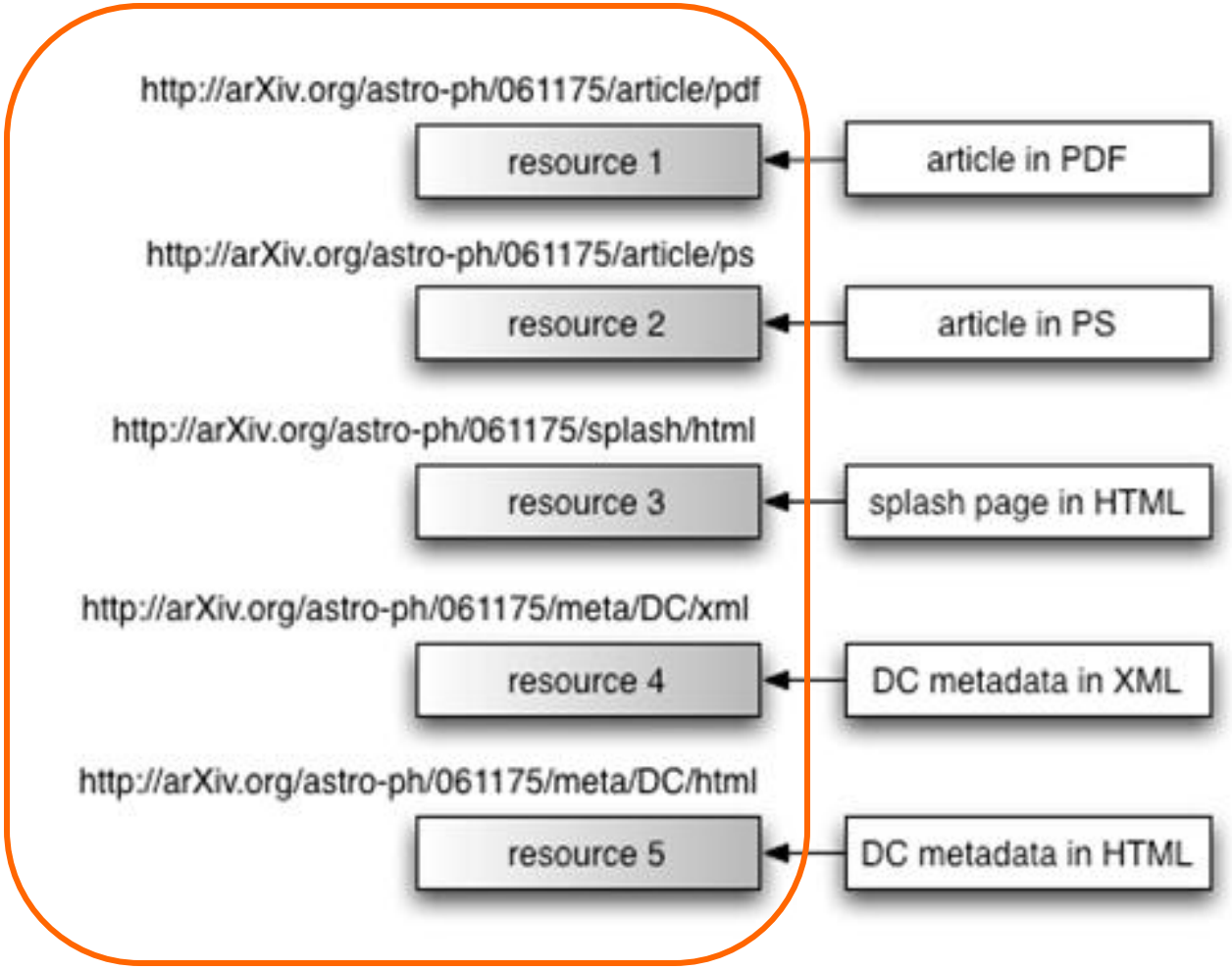
"Are we consistent in the way we create hypertext links between research papers in repositories?"

- Citation: Which Resource to link to?
- Citation: How to reference the PDF version (and not the PS version)?

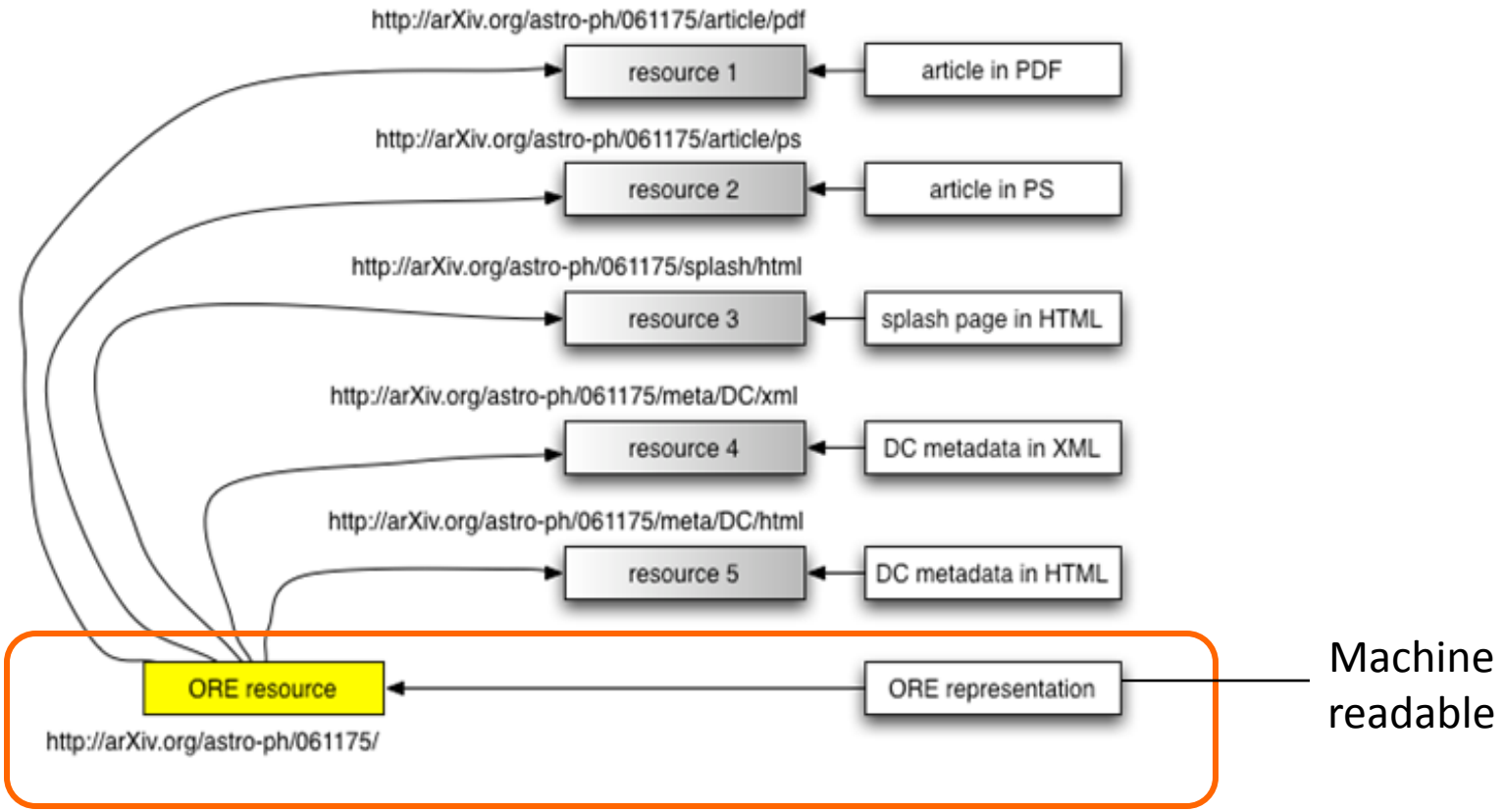


Observations (1)

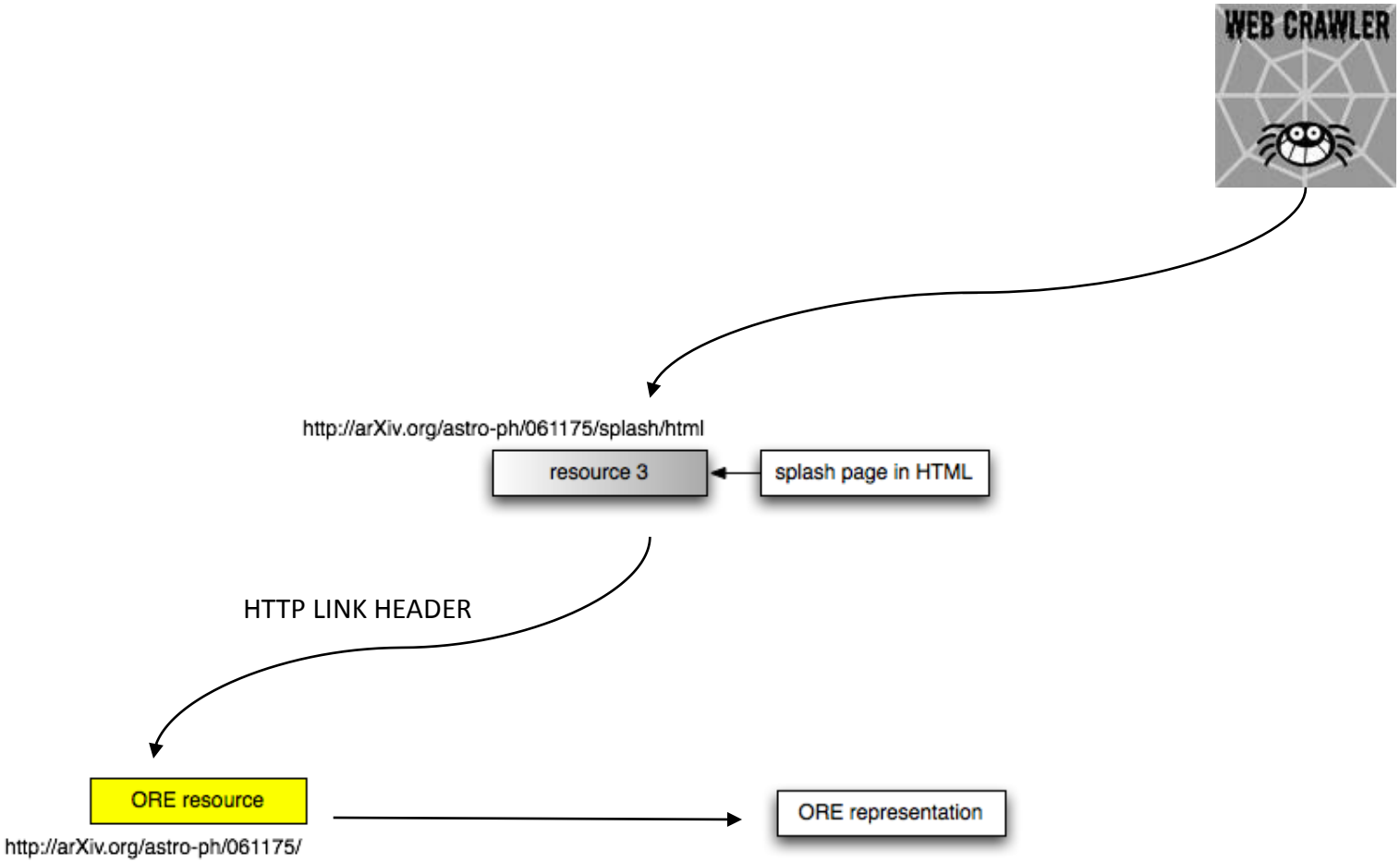
Views of digital object must be mapped to resources in order to be reference-able



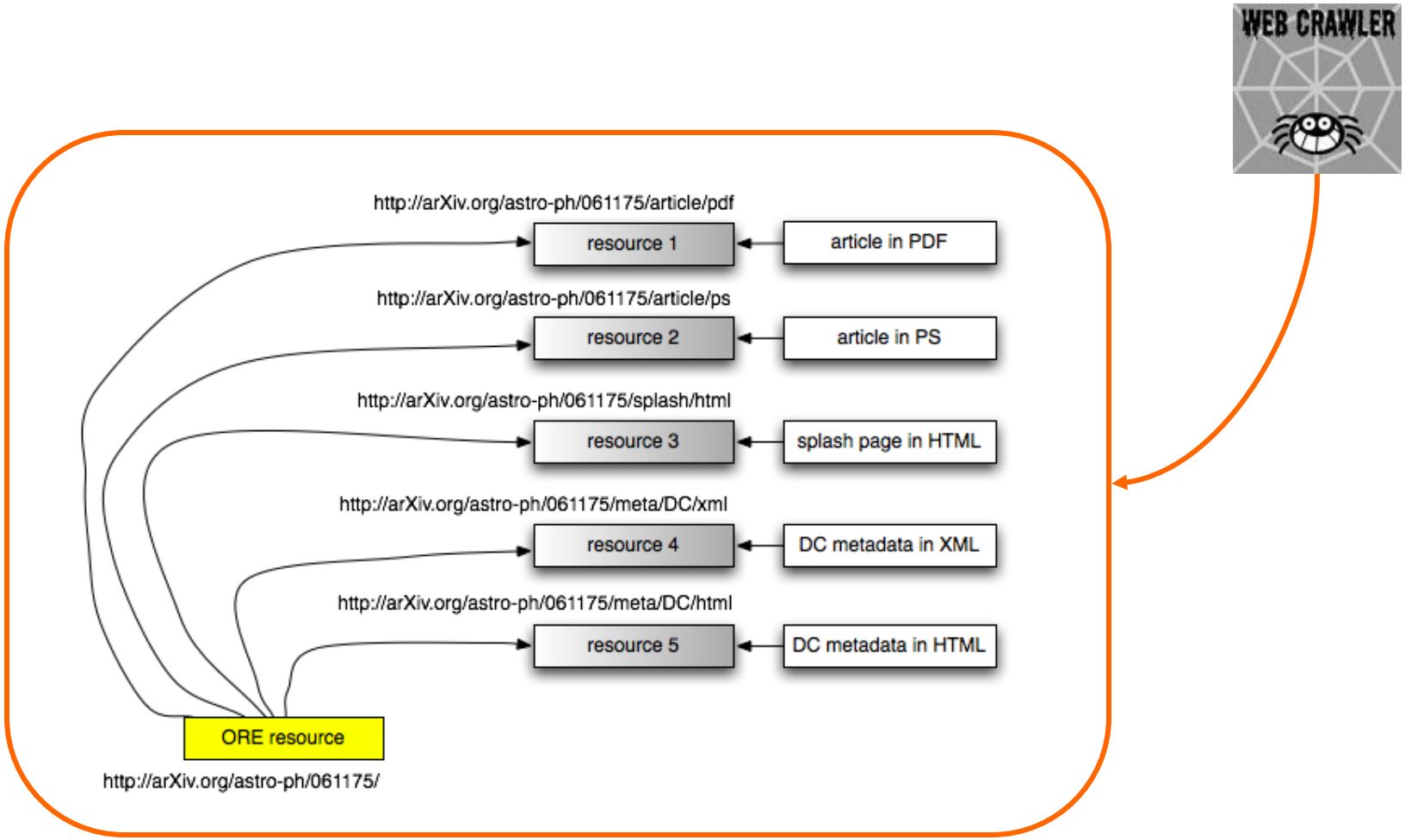
Observations (2):
Compound digital object must map to a resource with a representation that formally expresses the boundaries of the object



Observations (3):
Must allow for discovery of that representation (and hence of the Digital Object) by Web crawlers



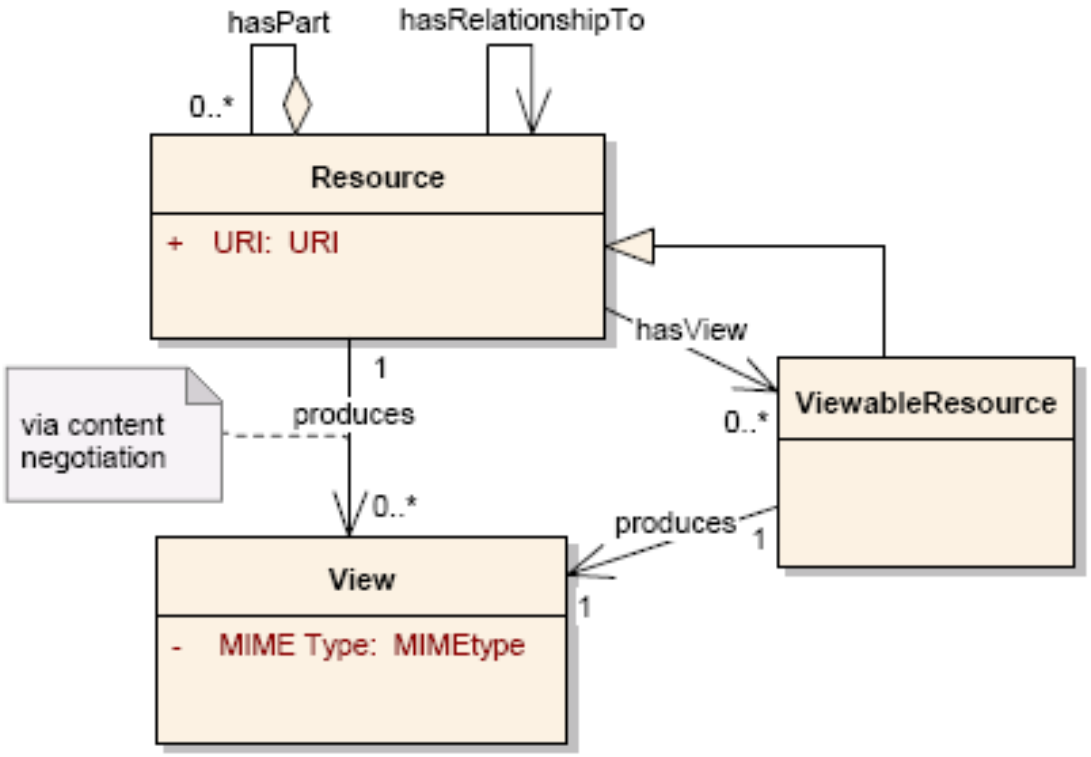
Observations (3):
Must allow for discovery of that representation (and hence of the Digital Object) by Web crawlers



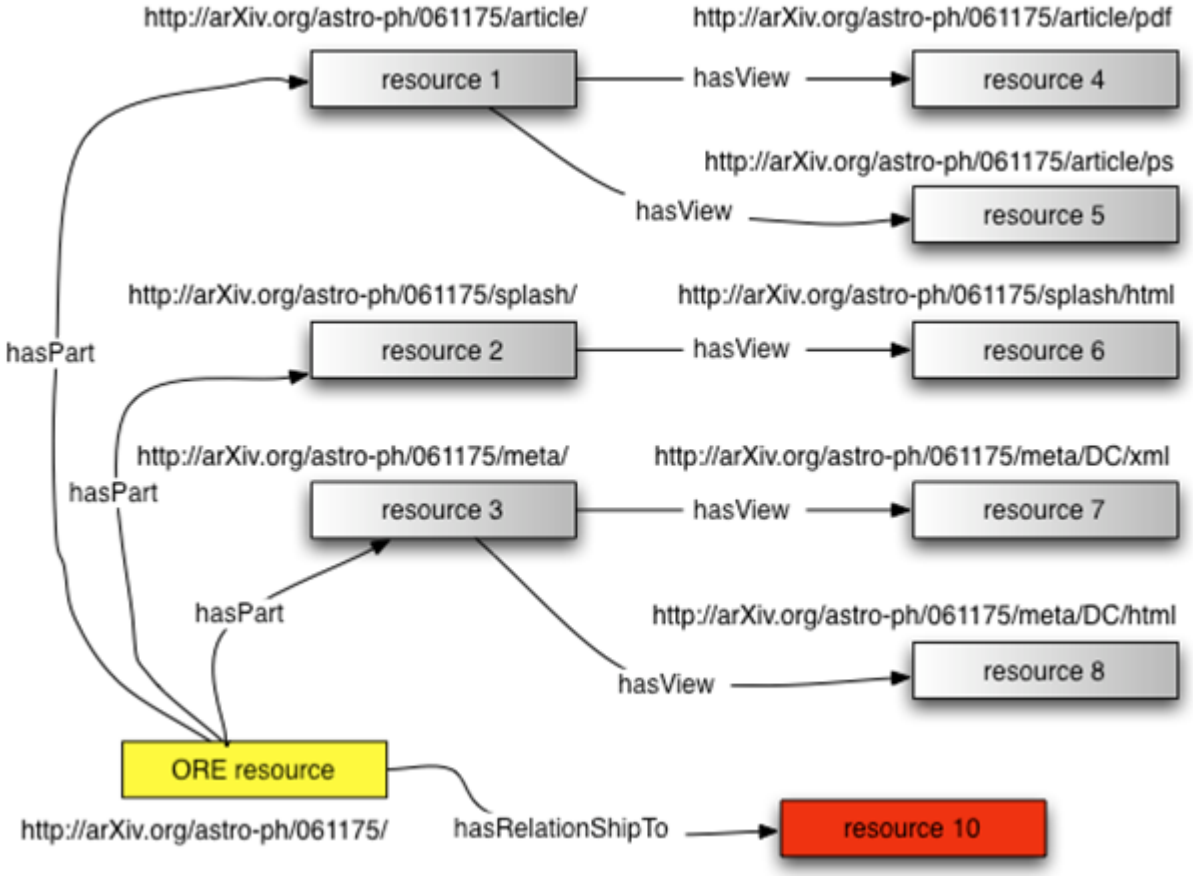
ORE representation based on ORE Model

- Formally expresses a bounded aggregation of resources and relationships that corresponds to a compound digital object
- Describes a connected graph:
  - finite set of resources and relationships among the resources
  - relationships among resources that are members of the aggregation and & resources are external to the aggregation
- Intra-aggregation relationships: *hasPart*, *hasView*
- Inter-aggregation relationships: *hasRelationshipTo with* community specialization

Preliminary ORE Model



Compound digital object modeled according to preliminary ORE model



OAI-ORE : Current focus

- Definition of the ORE Model
- Review of appropriate technologies for ORE Model and ORE representations
  - ATOM
  - DID, DIDL
  - Dublin Core Abstract Model
  - …
- Identification

OAI-ORE : Afterwards

- Look into core services, all based on exchanging ORE representations
- Three classes:
  - *Harvest*: a request for a batch of ORE representations from a repository, one per Digital Object.
  - *Obtain*: A request for an ORE representation for a specific Digital Object (represented in the Web as an ORE resource).
  - *Register*: A request to add new nodes or relationships to an ORE aggregation.

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

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)

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)
- Pete Johnston – Eduserve Foundation (UK)
- Richard Jones - Imperial College (UK)
- Peter Murray - OhioLINK
- Michael Nelson - Old Dominion University
- Ray Plante - National Center for Supercomputing Applications
- Rob Sanderson - University of Liverpool (UK)
- Simeon Warner - Cornell University
- Jeff Young - OCLC

Questions