Scholarly Communication

Evolution and Revolution

Information Science

Carl Lagoze

DELOS/NSDL Summer School

June 1, 2007

Joint work with …

- Dean Krafft – Cornell/NSDL
- Sandy Payette – Cornell/Fedora Commons
- Herbert Van de Sompel – Los Alamos
- Simeon Warner – Cornell/arXiv
- Members of the Open Archives Initiative Object Reuse and Exchange Technical Coomiitee

Funding From

- Andrew W. Mellon Foundation
- National Science Foundation
- Microsoft Corporation

Structure of talk

- What is scholarly communication
- A system under pressure
- Evolution Open Access and Institutional Repositories
- From Evolution to Revolution
- Web Architecture as a Foundation
- Open Archives Initiative Object Reuse and Exchange

Beyond Search and Access

As suggested by Borgman [14-16], digital libraries should match and indeed dramatically extend traditional libraries. As such, they should be much more than search engine portals. Like any library they should feature a high degree of *selection* of resources that meet criteria relevant to their mission, and they should provide *services*, including search, that facilitate use of the resources by their target community. But, freed of the constraints of physical space and media, digital libraries can be more adaptive and reflective of the communities they serve. They should be *collaborative*, allowing users to contribute knowledge to the library, either actively through annotations, reviews, and the like, or passively through their patterns of resource use. In addition, they should be *contextual*, expressing the expanding web of inter-relationships and layers of knowledge that extend among selected primary resources. In this manner, the core of the digital library should be an evolving information base, weaving together professional selection and the "wisdom of crowds" [54].

D-Lib Magazine, November 2005

What is scholarly communication

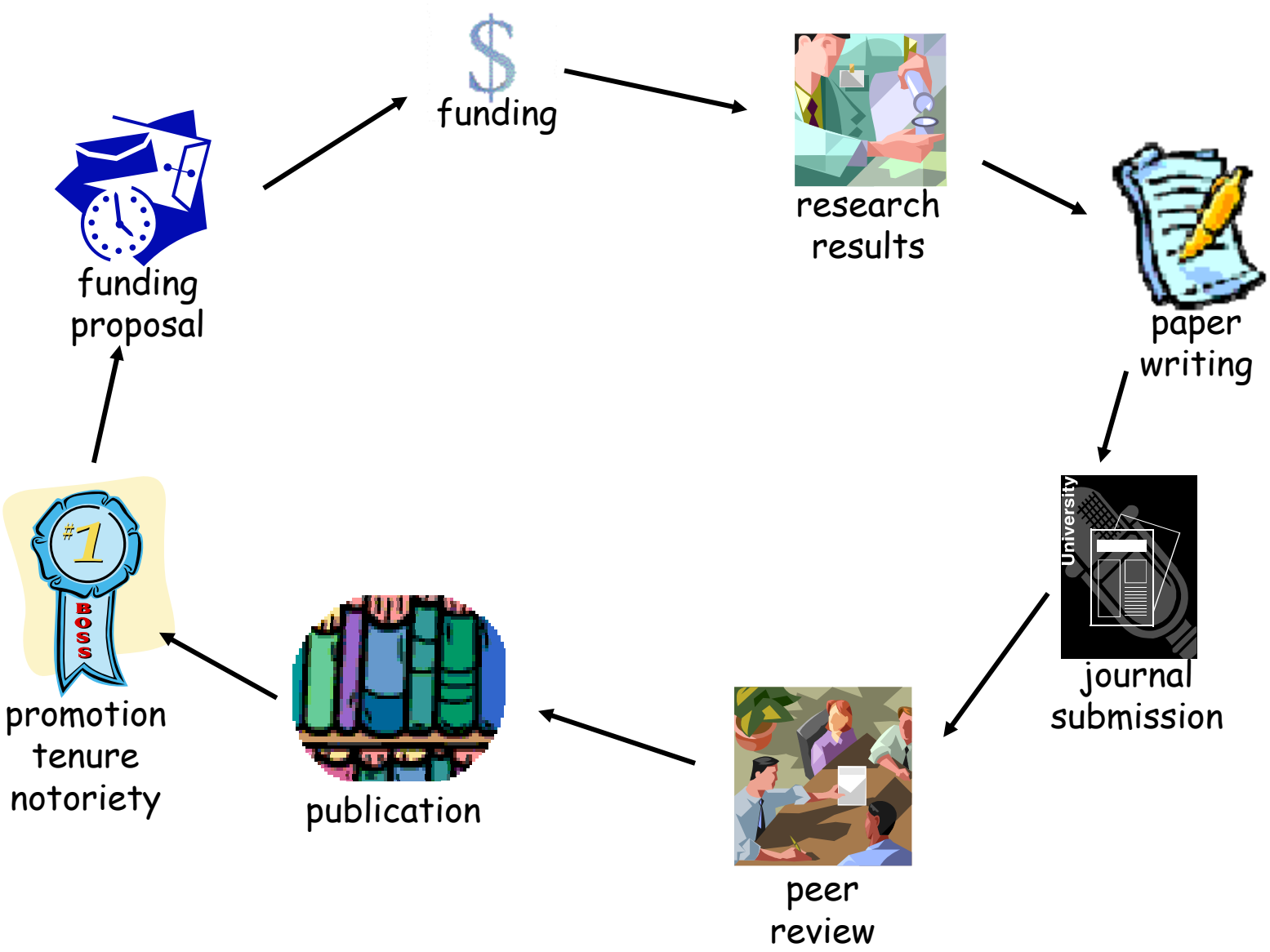
(Very) short history of scholarly communication

- Pre-history: Scholarship through personal communication
- 1665: first scholarly journal
  - From face-to-face communication to more open accessible system
  - Anselm Strauss: social worlds built on texts
- Late 20th century: Monopolization
  - Distortion of journal model
  - “Serials crisis”
- 1990’s: Digital Emergence
  - Web, E-journals, e-Print archives, institutional repositories
  - Reassertion of democratization
  - Access uber alles
- 21st century: ??

Why do scholars publish?

- It is the tangible product of our work
- Our funders expect it – big publication lists always look good on reports
- It is our responsibility to our colleagues
- It is good for our egos
- It is the/a key to tenure, promotion, and hiring

How the system works



Who are the role players

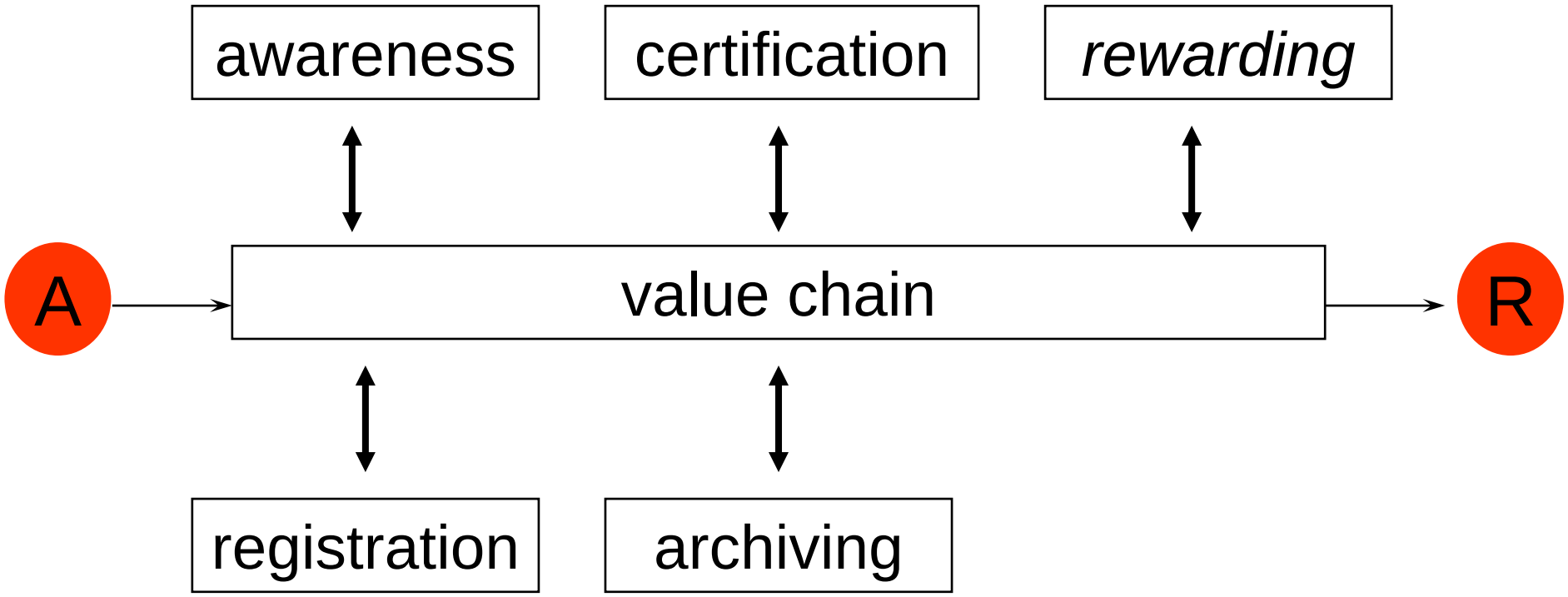
- Scholars
  - Faculty
  - Researchers – Commercial, Academic, Government Labs
- Publishers
  - “Big” for-profits: Elsevier, Springer-Verlag (Kluwer)
- Learned and Professional Societies
  - ACM, APS, AMS
    - Publishing operations often subsidize other operations
    - Some are hard to differentiate from for-profit publishers – e.g., IEEE, American Chemical Society
- Libraries
  - In paper system the sole distribution point for publications
  - Archiving and preservation role

Functions of scholarly communication

- Registration – to establish intellectual priority
- Certification – to certify quality and validity
- Awareness – to ensure accessibility
- Archiving – to endure availability for future use
- Rewarding – for tenure, promotion, compensation

(Roosendaal & Geurts)

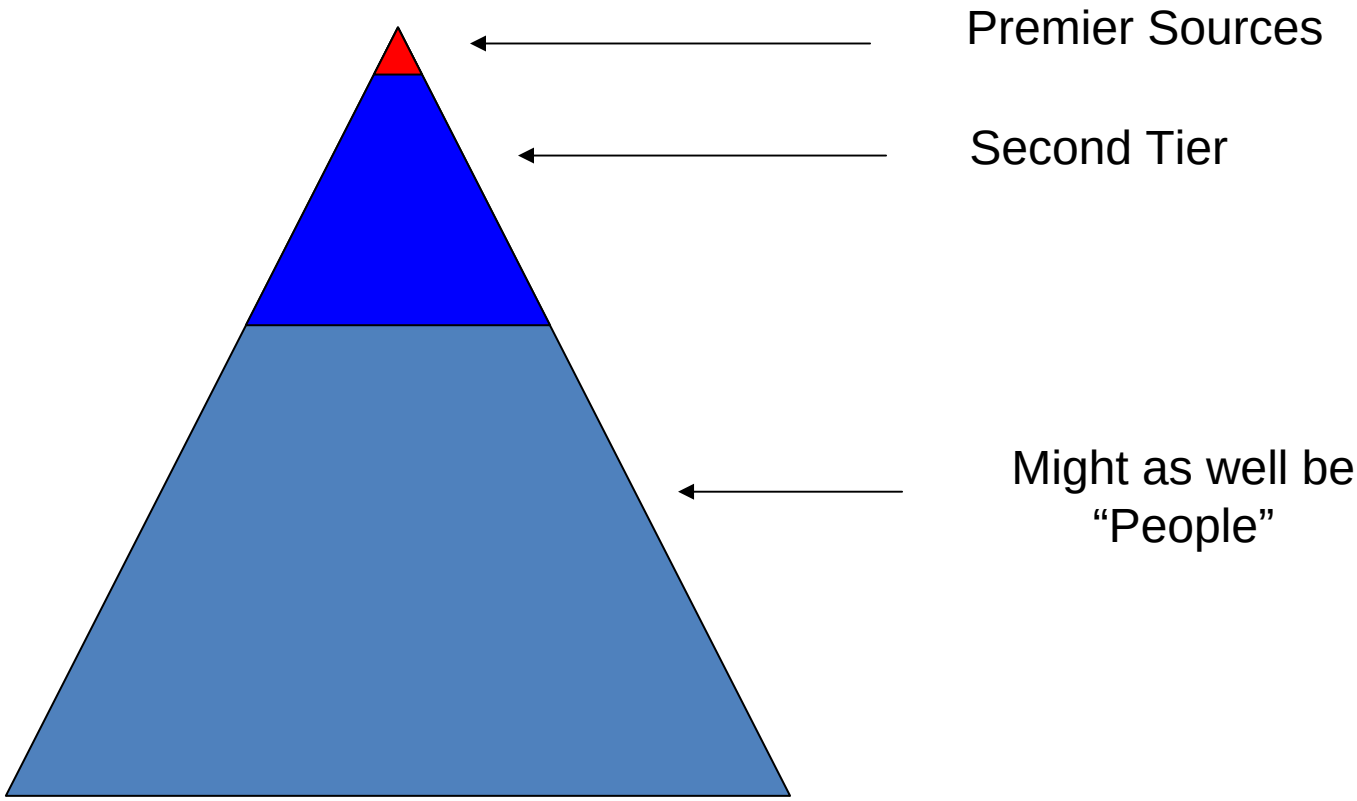
Value chain perspective of scholarly communication system



Peer Review

- Claimed to be the basis of quality in the system
  - Is it really fair and “objective”?
  - Is it the only measure of quality?
- Almost entirely volunteer
- Blind or visible

Scholarly publishing is extremely hierarchical



Establishing Premier Journals – Citation Analysis

- A citation is a reference from one work to another [as a hyperlink: a citation link]
- Citation Graph – nodes are works, vertex is citation
- Citation analysis uses citation relationships to analyse patterns in research
- ‘Bibliometrics’
  - (study of patterns in literature)
- Eugene Garfield
  - ISI Science Citation Index (SCI) identify “hottest” journals

Assumptions in current scholarly publishing system

- Publications are difficult to produce
- Publications are difficult to distribute
- Readership is by closed community
- Quality assessment is by closed community
- Archiving and management is by closed community

Some “side effects” of the current system

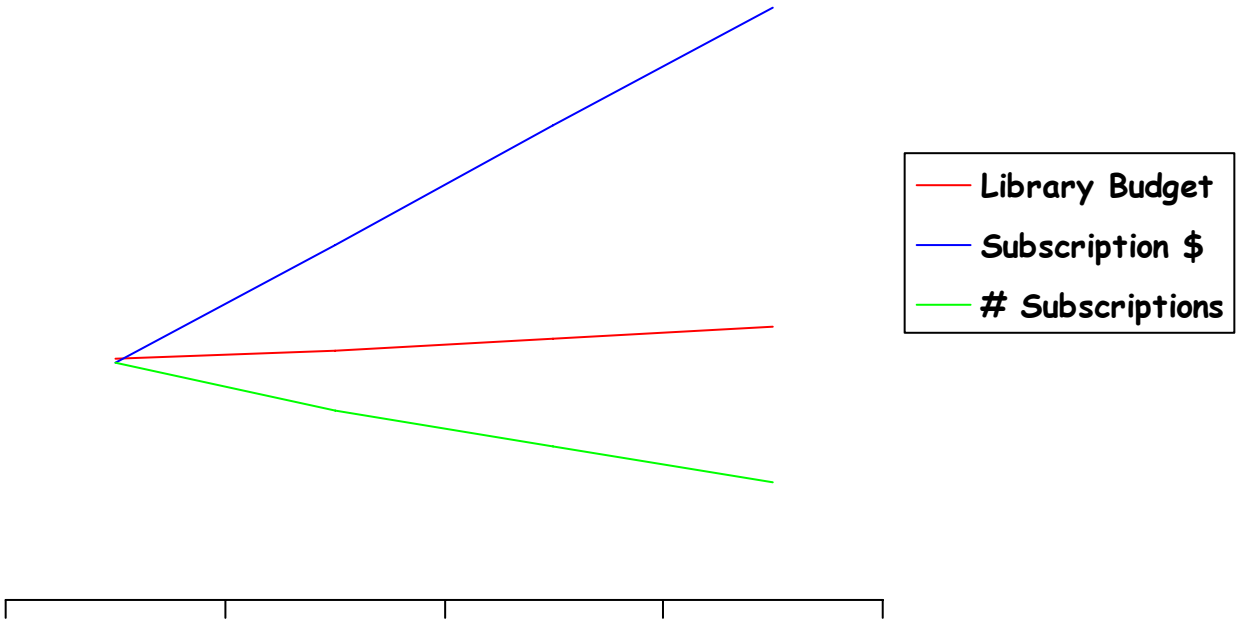
- *Rich get Richer!*
  - Best known scholars have an advantage in peer review system
  - Riches institutions in richest countries can best afford journal prices
  - High prestige journals are self-sustaining due to SCI factors
- Global scholarly divide worsens
  - Research institutions in developing countries can’t afford subscriptions
  - Intellectual capital flees
- Hierarchy gets more stratified
  - Unpublished papers disappear
  - Entry into the system is difficult

A System under Pressure

Issues and Changes

- Exponentially increasing amount of information produced by scholars
- Growth in both dimensions
  - Horizontal
    - Increased specialization
    - New and more specialized journals
      - 5000 peer reviewed journals in education research
  - Vertical
    - Diminish single source reliance
    - Facilitate multi-uses for single source
- Compressed time for “relevance” of results, increased demand for rapid delivery
- Changes in the type of publication
  - demand for data availability

Broken Economics



Some facts about subscription prices

- Average journal subscription price has gone up 7-10%/year over the past 10 years
  - 1986-2002 US CPI increased 57%, research library journal subscription budget increased 227%
- Some journals have gone up 20-40% of the past 5 years!!!
- Some journals cost 5K-10K per year
- Many societies have raised subscription prices 20-25% over the past several years
  - “Catch up” to the private publishers
  - Fund research into digital initiatives
  - Cover the rest of their operations
- Elsevier’s price rise per year equates to one less faculty member per year at Cornell (according to Bill Arms)
- http://oap.comm.nsdl.org/10most.html

Where are the costs in the print system

- Publishers
  - Copy-editing
  - Production
  - Administration of review system
  - Production
  - Distribution
- Libraries
  - Cataloging
  - Preservation
  - Binding
  - Shelving

Economics have changed!

- Distribution in electronic system is basically free
  - Fundamental assumption of paper system is eliminated
  - “Publishing” by everyone should be encouraged and supported
- Services need to be disambiguated from distribution
  - Free distribution doesn’t mean that there isn’t an economic model
  - Systems like review, filtering, awareness can be built on top of a free distribution system

The Scholars have changed (or are changing):

- The web 2.0 generation is growing up
- Systems that combine social activity and information are the norm
- Will they accept our norm?

Open Access and Institutional Repositories

Open Access

- Various proclamations
  - Budapest, Berlin Open Access
  - Harnad's "Subversive Proposal"
- Products of Scholarship should be controlled by the scholars
- Scholarship works through analysis, reuse, and adaptation
  - Standing on the shoulders of giants.
- Openness of systems allows it to flexibly adapt to changing conditions and contexts
  - Think about the web
- Open Access does NOT mean free access

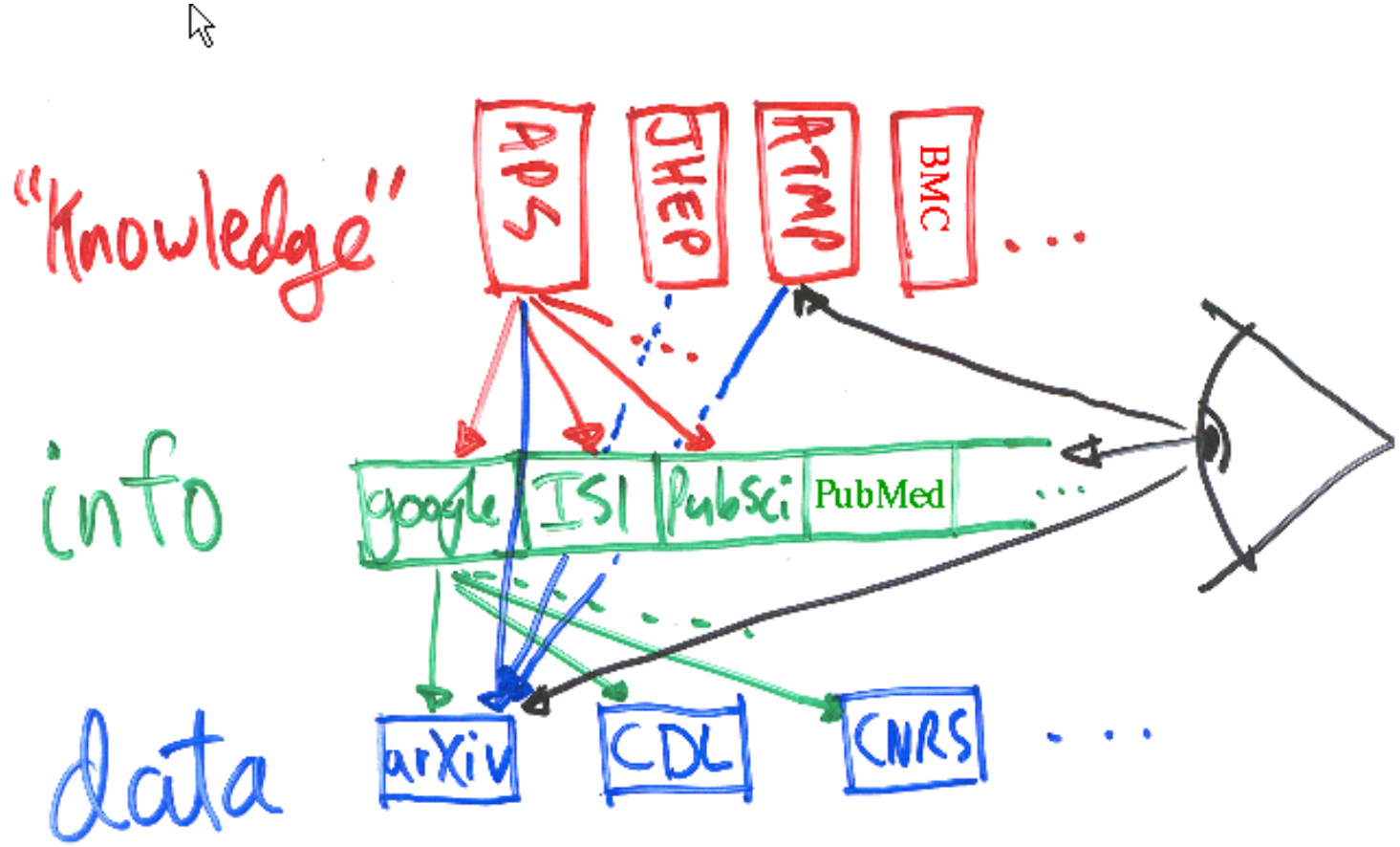
Institutional Repositories

- Various technologies: Dspace, Fedora, ePrints
- Universities, laboratories act as agents for open access
- Retention of intellectual property at the institutional level
- Creates a data layer for construction of higher level services.

Federating the data layer

- Interoperability Protocols and Standards are basis of Federation
- Interoperability provides means of exchange and interaction with heterogeneous systems.
- Common interoperability standards
  - Dublin Core
  - OAI-PMH

Building on The Data Layer



Acks. P. Ginsparg

What are the implications of this model?

- A marketplace of ideas
- People choose appropriate entry points into the system
  - Troll for free at the lowest layers
  - Pay for guided entry at upper layers
- Exposure of the “long tail”
- Money can be made by synthesizing information
- Standards for interchange amongst layers are important (e.g., OAI-PMH)

Have open access, institutional repositories, and the web solved all our problems?

What has it accomplished?

- **Early Dissemination**:
  - Enhance upstream scholarly communication
- **Open Access**:
  - Bypass of traditional publisher model
- **Document Discovery**:
  - Increased exposure to commodity search engines (Google Scholar)
- **Storage and Archiving**:
  - New models for distributed preservation (e.g., LOCKSS)

But these changes are *evolutionary*, not *revolutionary*

- An adaptation of the traditional publishing paradigm
  - Submit documents
  - Gain access to documents
  - Share results earlier in the scholarly process, and electronically

- Unit of discourse and dissemination is still the traditional (largely static) *document*
  - Store documents to provide access and archiving
  - Index documents to promote search and discovery
  - Citation analysis to understand relationships of documents

Why is this not enough?

- What about process and workflow that is at the heart of scholarship?
- Aren't scholarly results richer than the static artifacts of traditional publishing?
  - What about data, visualization, simulation?
- Shouldn't the system help scholars "stand on the shoulders of giants"?
  - Mechanisms for reuse, refactoring, and re-aggregation of existing scholarly artifacts are too limited.
- Where are the tools for collaboration, commentary, annotation – knowledge sharing?
  - Shouldn't the 'object-centered sociality" of blogs, myspace, wikipedia, etc. extend to the scholarly domain?
- Shouldn't we be able to apply the algorithmic methods that have revolutionized web search to scholarly communication?
  - Can't we do more than citation analysis?

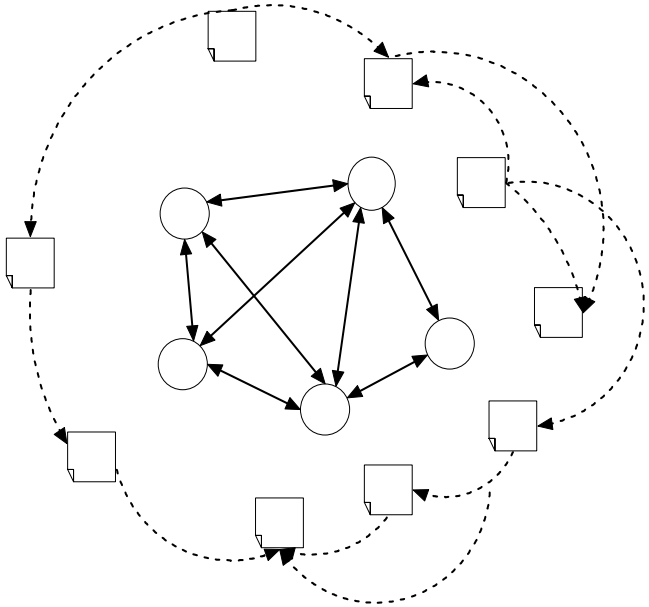
From Evolution to Revolution

What do we want to be able to do after the revolution?

- Content aggregation:
  - combining information entities in novel ways
- Information reuse:
  - allowing secondary, tertiary products
- Information transformation:
  - combining information entities with computational services
- Collaboration and contribution:
  - exploiting the wisdom of crowds through annotation, commentary, etc.
- Knowledge integration:
  - capturing semantic and factual relationships among information entities

Build a revolutionary scholarly communication system that resembles the nature of scholarship itself.

Building Rich Scholarly Knowledge Networks



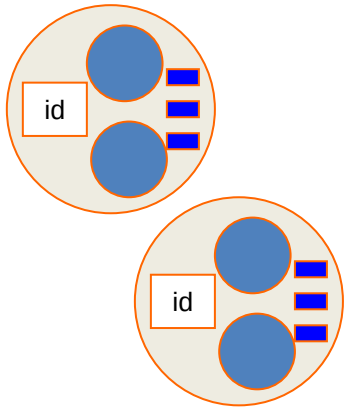
Disconnected networks:

- **formal publication network**
- **social network (actors)**

Translating to functional requirements

- Redefine the *information unit* of scholarly communication
- Redefine the repository from storage and access to *service provision* over distributed components
- Support the *exchange of complex information* across *independent value-adding services*
- Record the *workflow* (provenance) of information units as they move across value-adding services
- Provide *open-source protocols and models* enabling automated analysis (beyond Page Rank)

New Information Unit

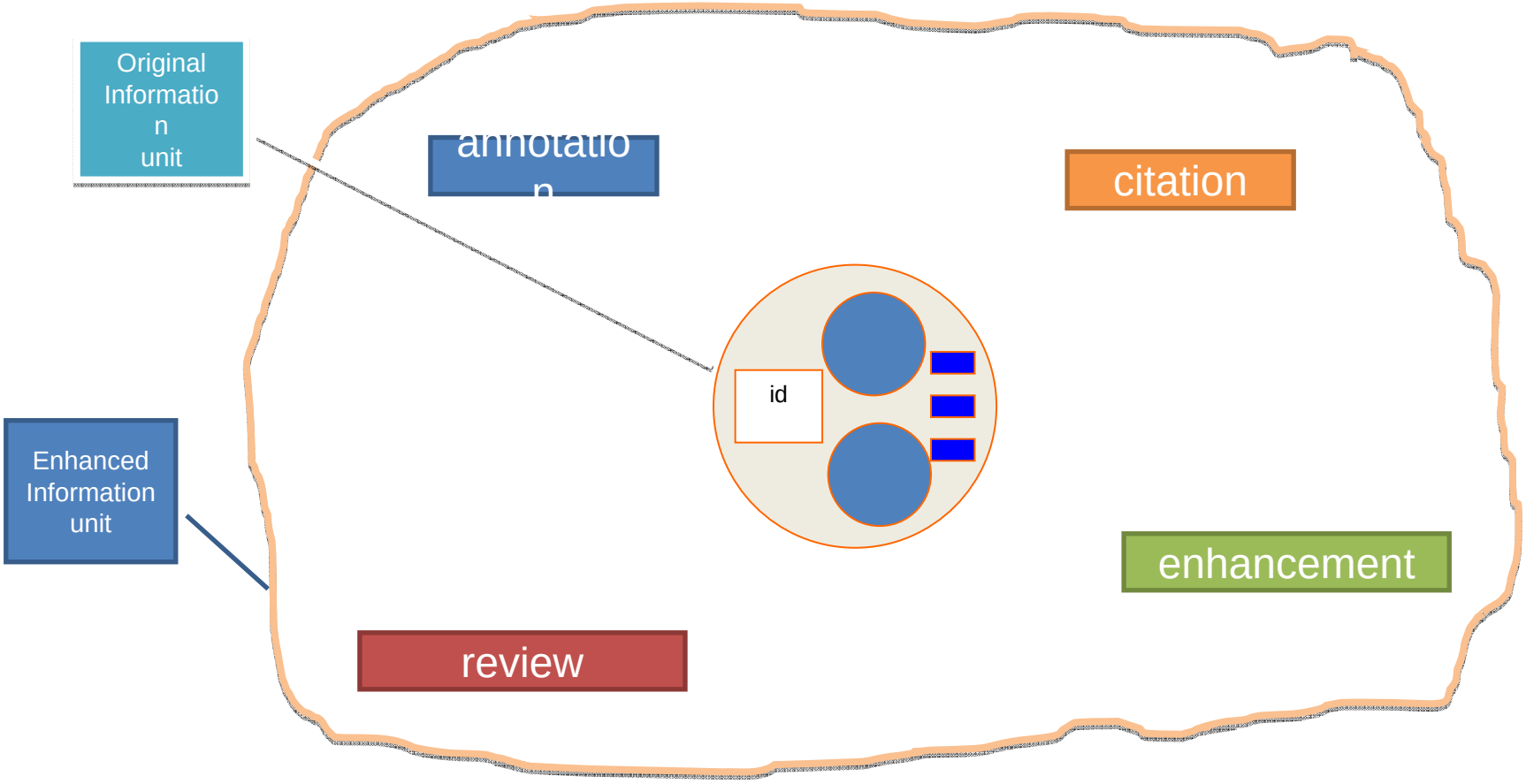


Digital Objects

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

That grow in value over time



Shameless Promotion
(but based on objective analysis)

http://fedora.info

But also…

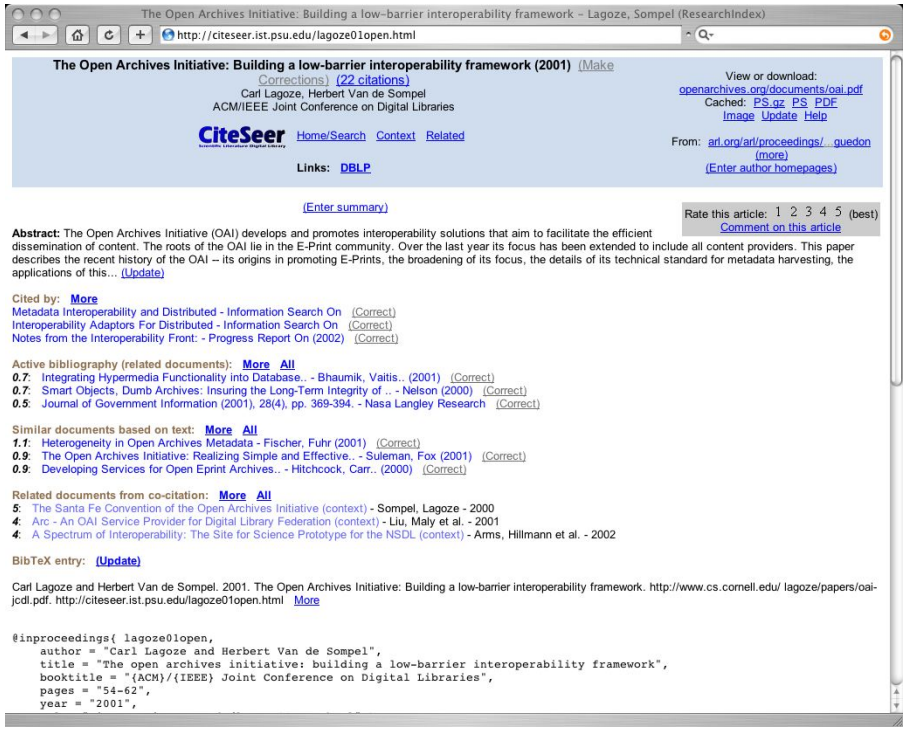

Apache
HTTP SERVER PROJECT

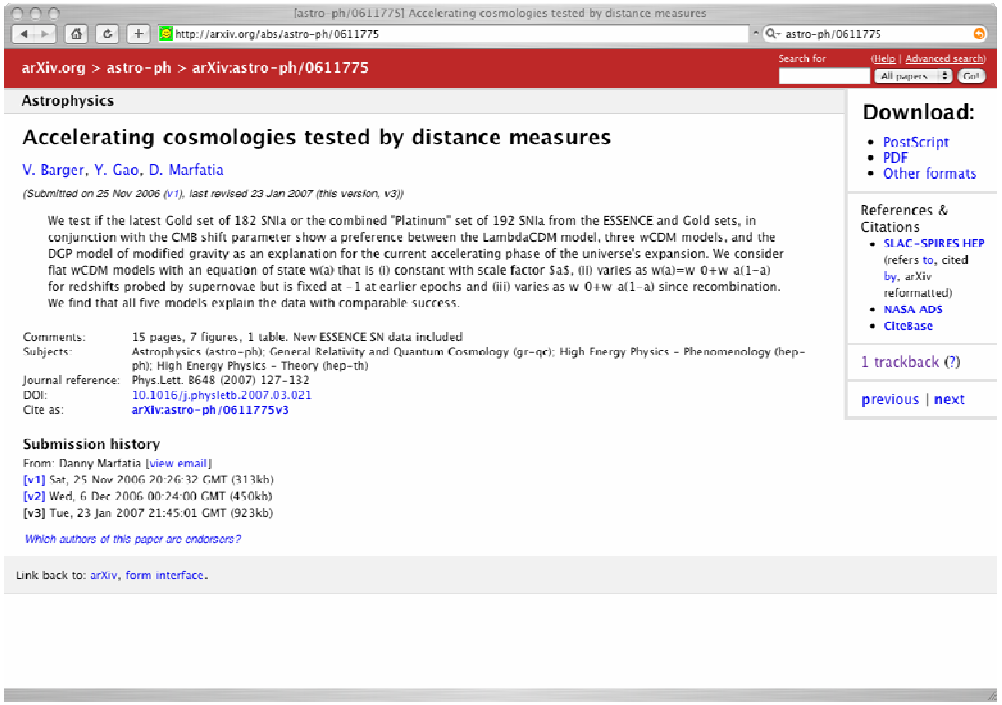

DSPACE


eprints


Cornell University
Library
arXiv.org

Scholarly Examples

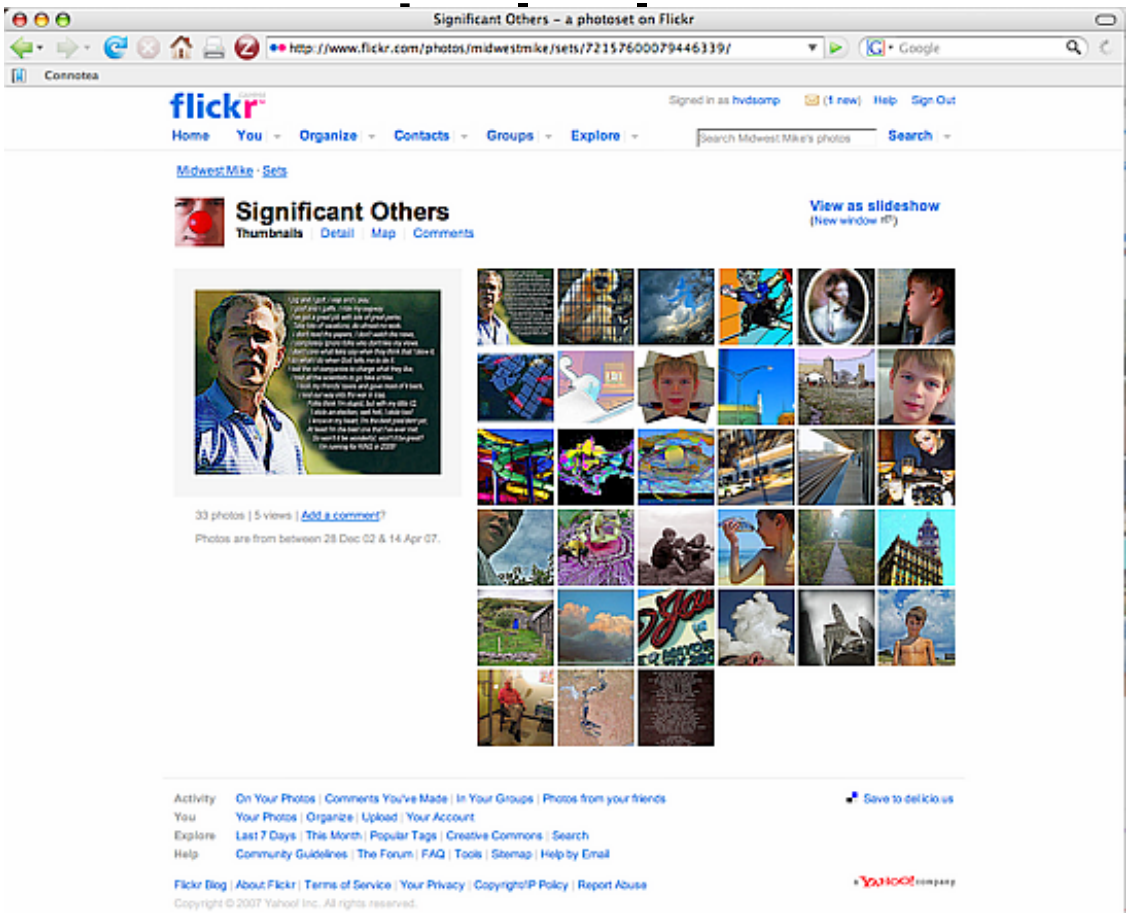




http://citeseer.ist.psu.edu/lagoze01open.html

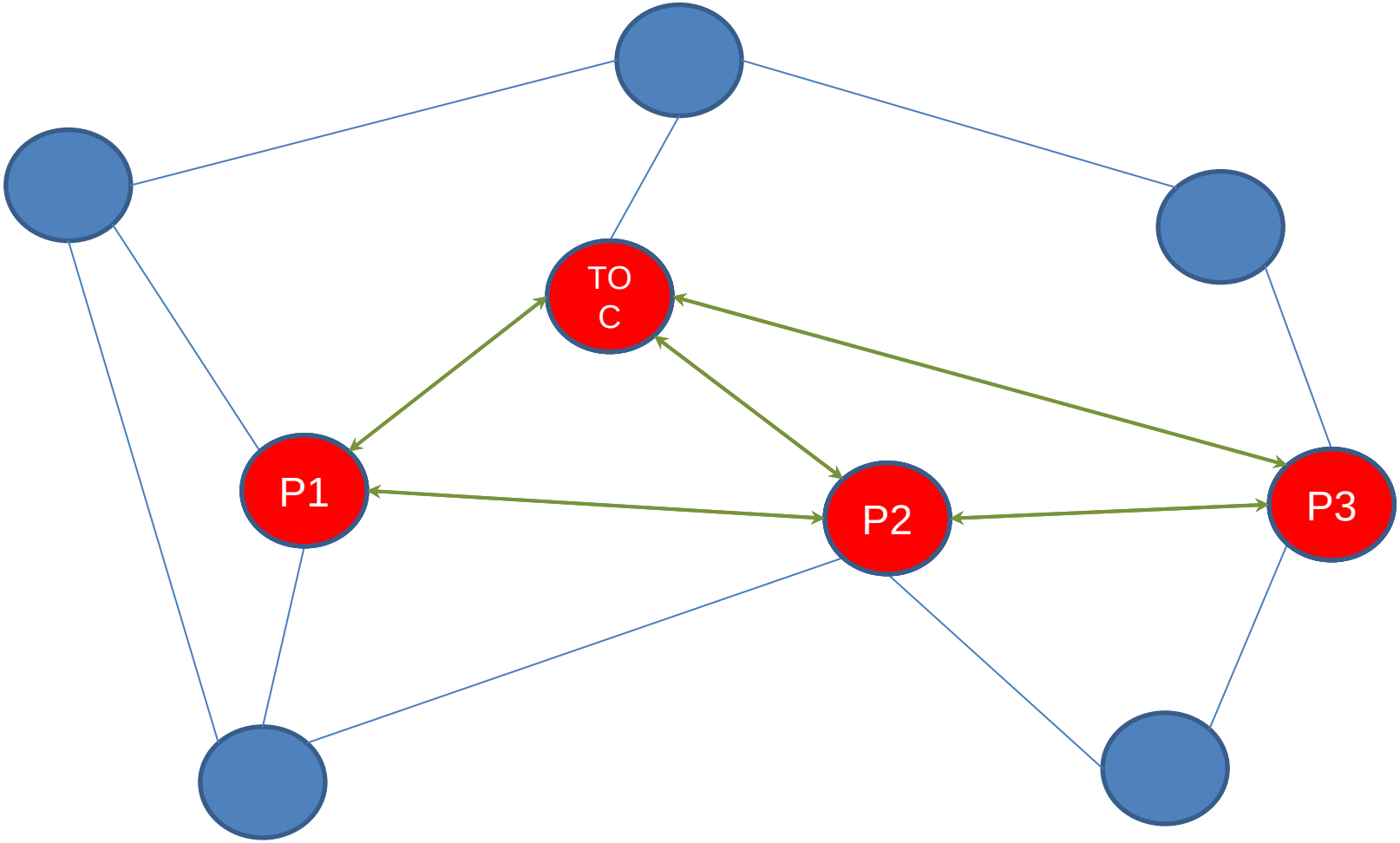
http://arxiv.org/abs/astro-ph/0611775

But these things are not only

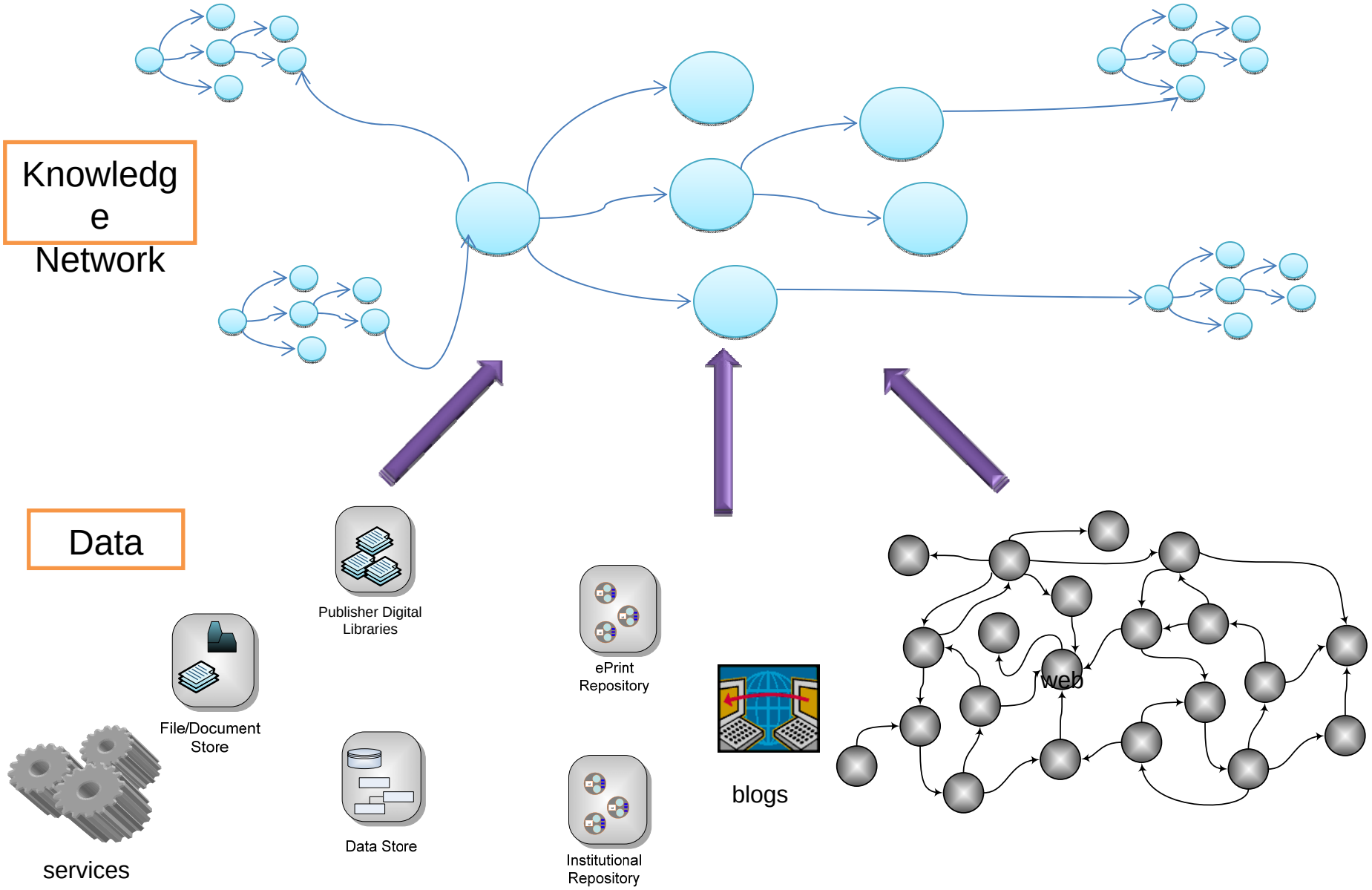


http://www.flickr.com/photos/midwestmike/sets/72157600079446339/

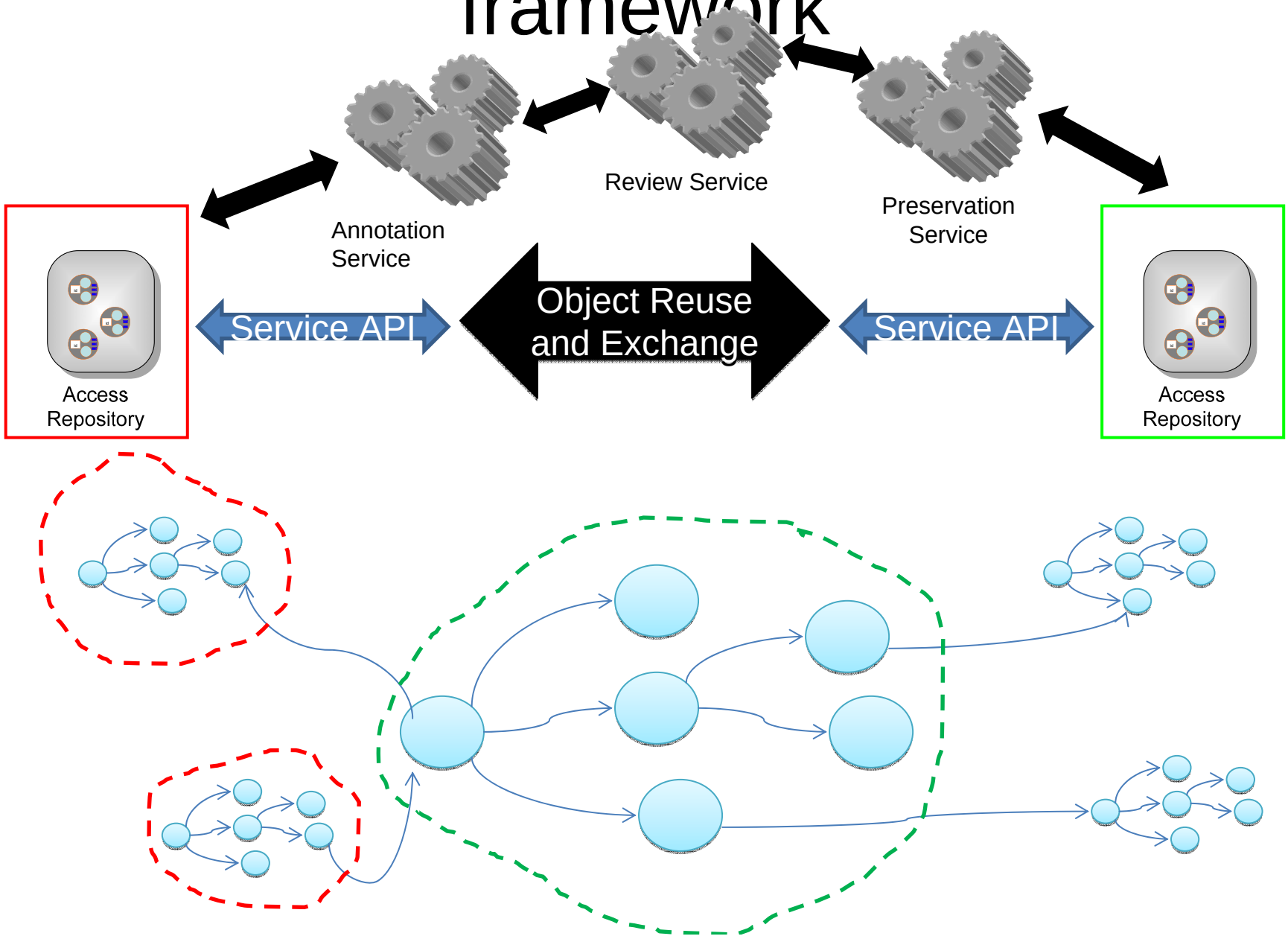
And in fact we use compound objects every day on the web



Moving outside the repository boundary



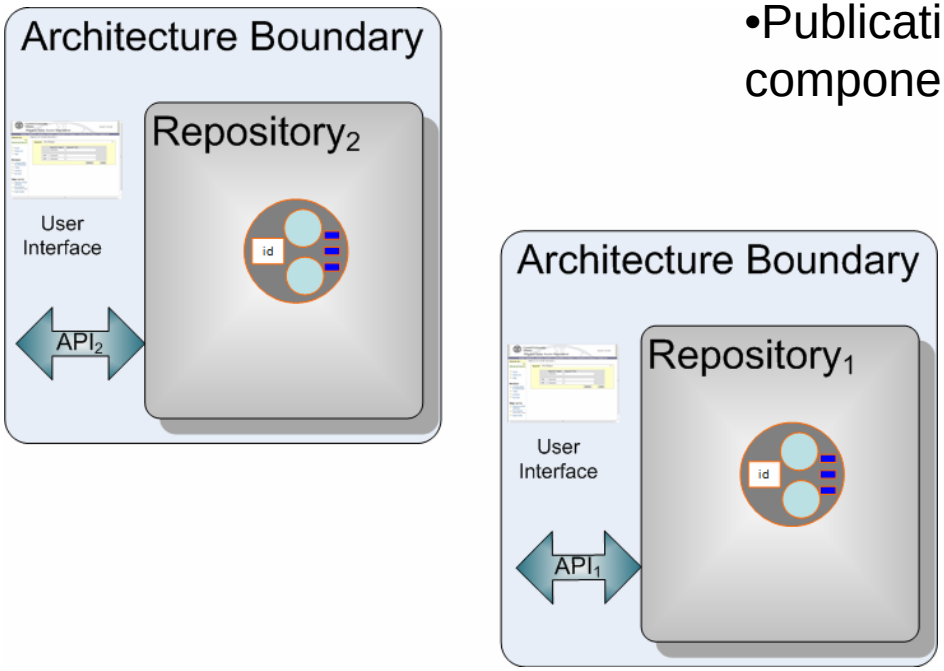
Repositories within a service framework

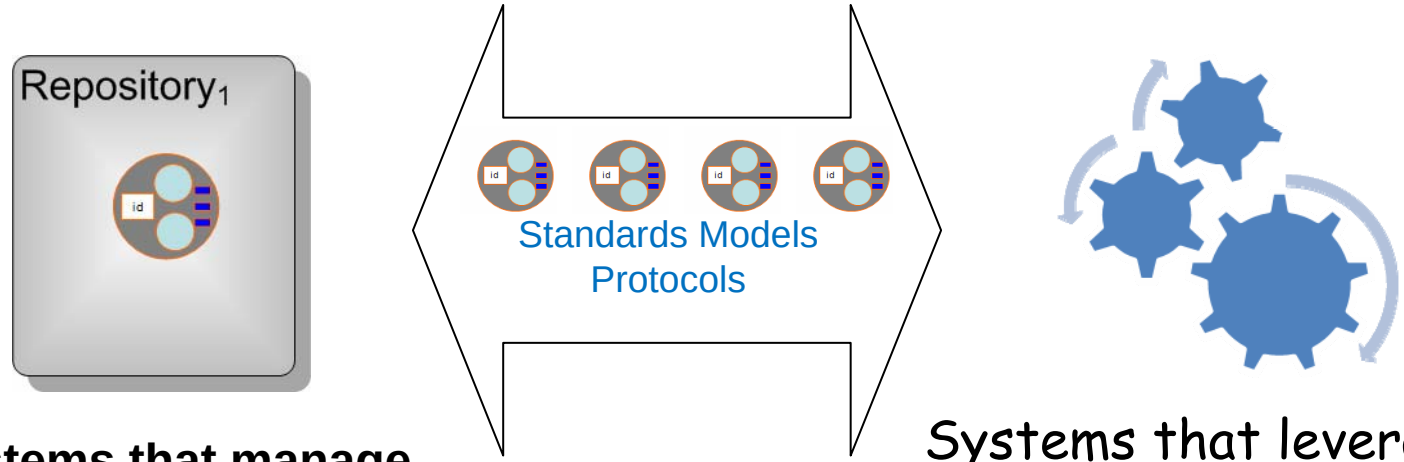


Access Repositories

Without standards, repositories expose compound objects in manners specific to the repository architecture:

- Interfaces (API & user-oriented)
- Identification schemes
- Publication of compound objects and components to the Web





Systems that manage digital objects

- Institutional repositories
- Discipline-oriented repositories
- Publisher repositories
- Dataset repositories
- Cultural heritage repositories
- Learning object repositories
- Digitized book and manuscript collections
- Image repositories
- …

Systems that leverage managed digital objects

- All repositories from left column
- Search engines
- Authoring tools
- Citation management tools
- Collaborative environments
- Social network applications
- Graph analysis tools
- Preservation services
- Workflow tools
- …

Web Architecture as a Foundation

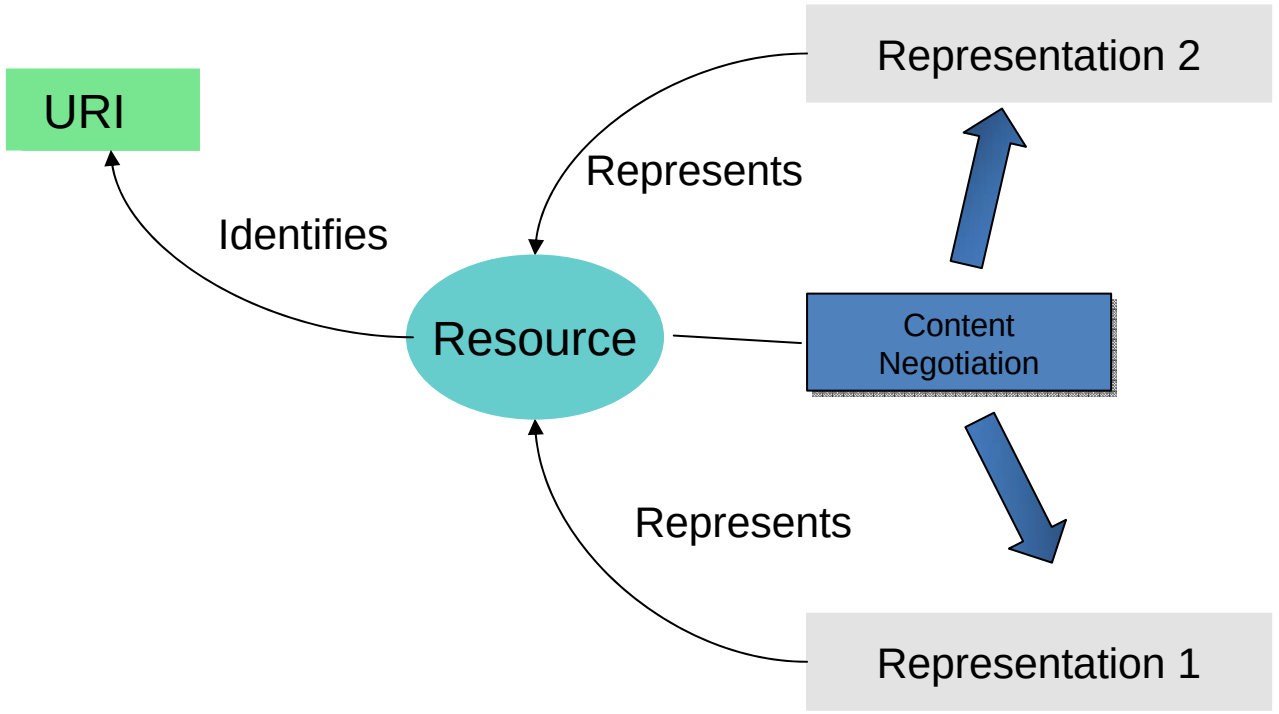
The web is a notably successful instance of interoperability

- URIs
  - Resources
  - Representations
- HTML
  - CSS
- HTTP

Working with the web architecture

- Whatever we do 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.
- (Digital Libraries must be congruent with evolving trends of “web culture”)

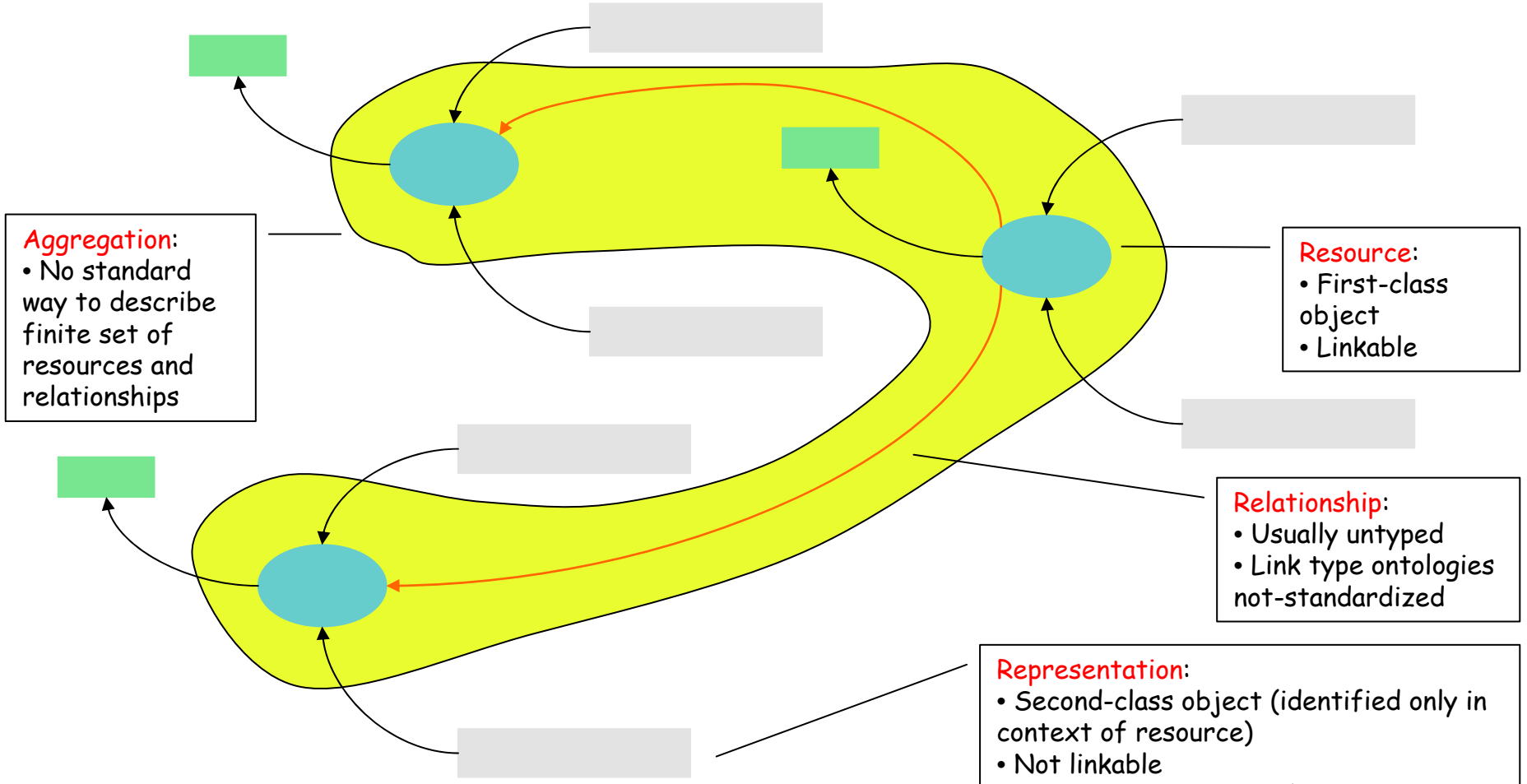
Nature of web resources



Think for a second…

- When I access google.com on my cell phone it looks different than on my desktop
- When I access google.com from Paris it looks different than when I access it from Ithaca

W3C Web Architecture: details



So what does this mean in our context?

- We need the notion of aggregations of resources represent compound objects
- We need support for citing compound objects and their parts
- We need to express well-defined relationships among these objects and their components

Open Archives Initiative Object Reuse and Exchange

OAI Object Re-Use and Exchange

- OAI-ORE is a new interoperability 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:
  - Coordinators: Carl Lagoze & Herbert Van de Sompel
  - ORE Technical Committee: 13 international members
  - ORE Liaison Group: 8 international members
  - ORE Advisory Committee: 16 international members
  - Representing: scholarly publishers and aggregators, eScience, eHumanities, education, search engines, various repository systems, digital library efforts, related standardization efforts, etc.

- See http://www.openarchives.org/ore/

OAI is not just about metadata anymore

OAI-PMH	OAI-ORE
Repository structure	Object structure
Repository centric	Web centric
Metadata centric	Resource centric
Metadata harvesting	Object re-use (obtain, harvest, register)

OAI-PMH and OAI-ORE are complimentary;

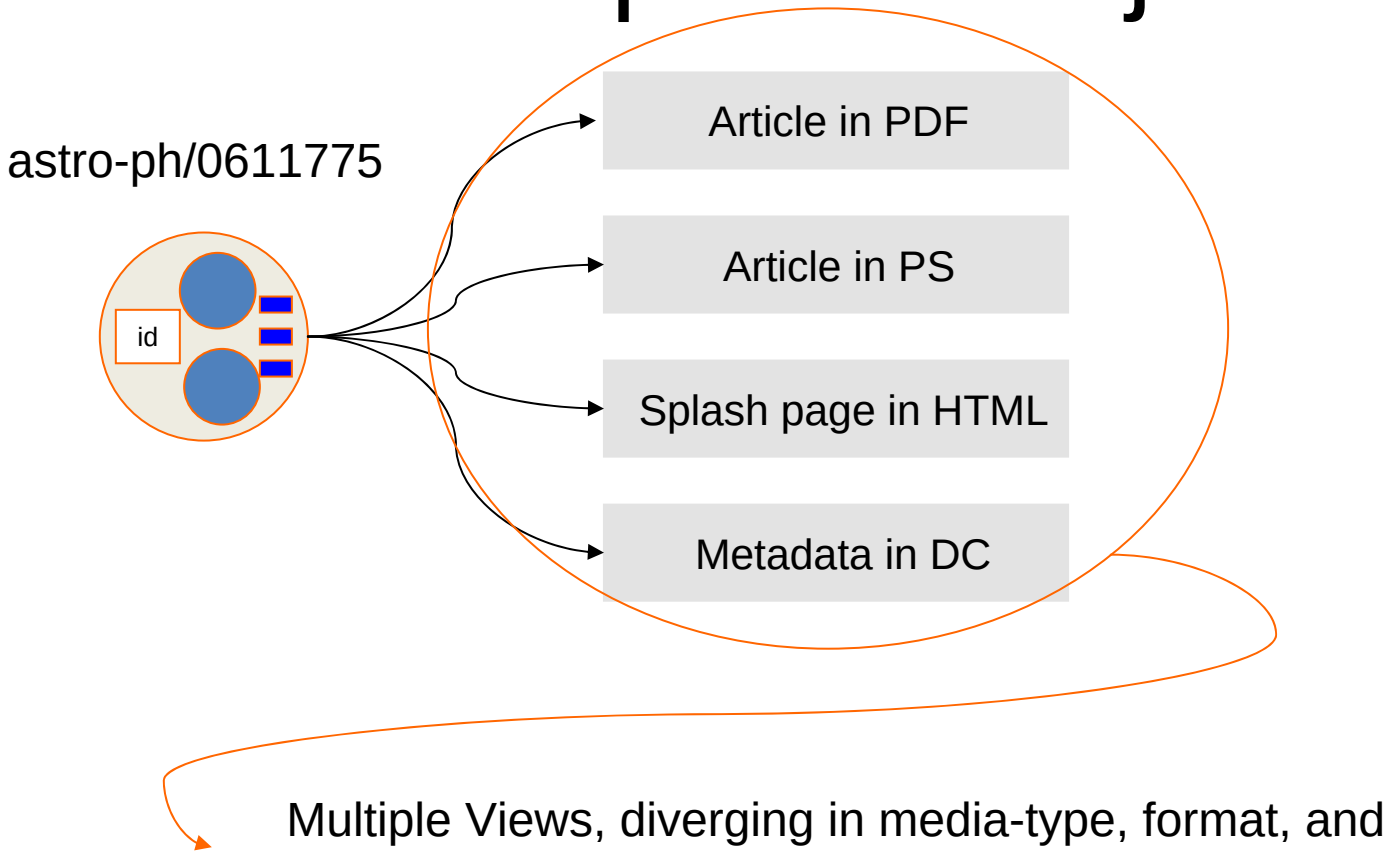
- you can do one without the other
- you can do them together

OAI Object Re-Use and Exchange

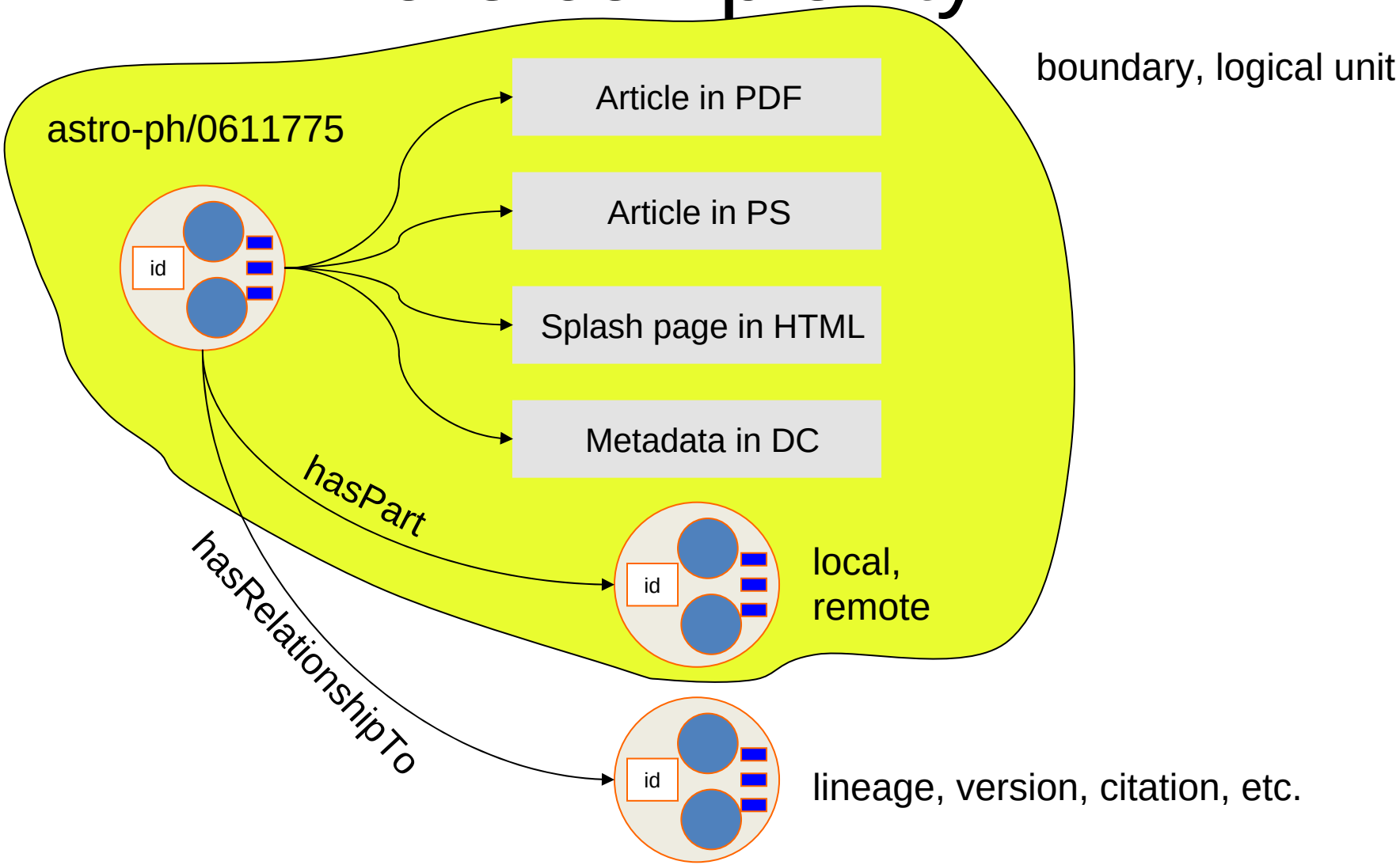
- Develop, identify, and profile extensible standards and protocols to allow repositories, agents, and services to interoperate in the context of use and reuse of compound digital objects beyond the boundaries of the holding repositories.

- 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,
  - provide the foundation for more advanced information environments

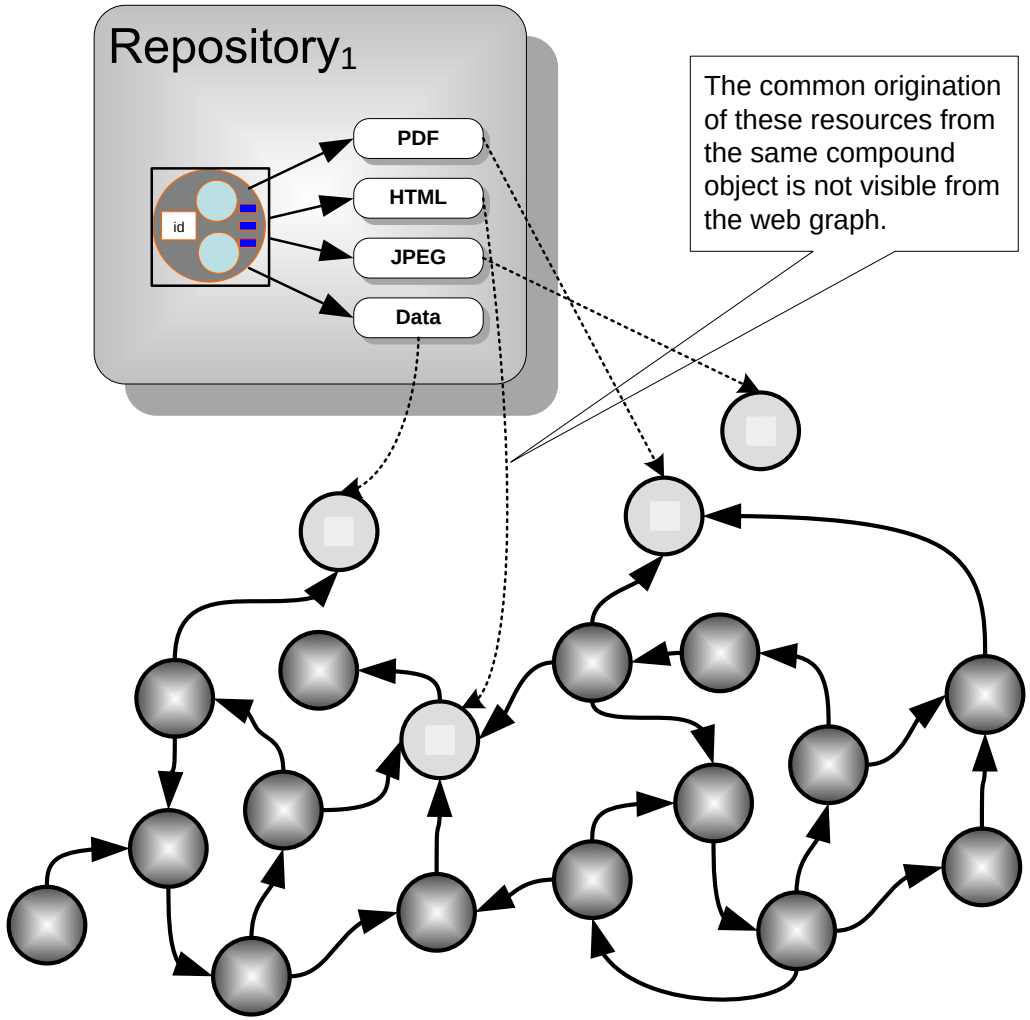
Compound Object



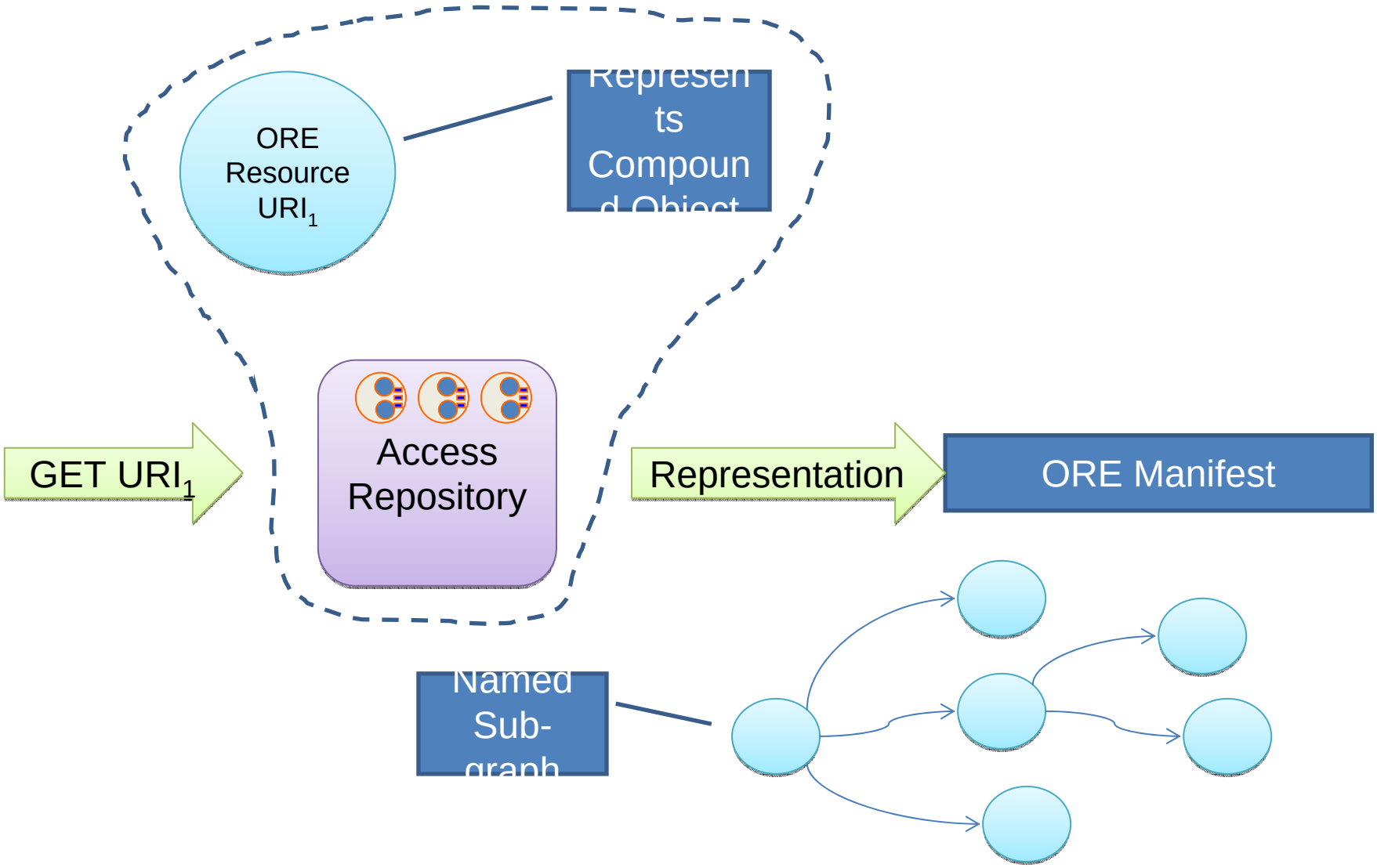
More complexity …



Exposing the components of a compound object as web resources (with URIs) solves one problem, but…

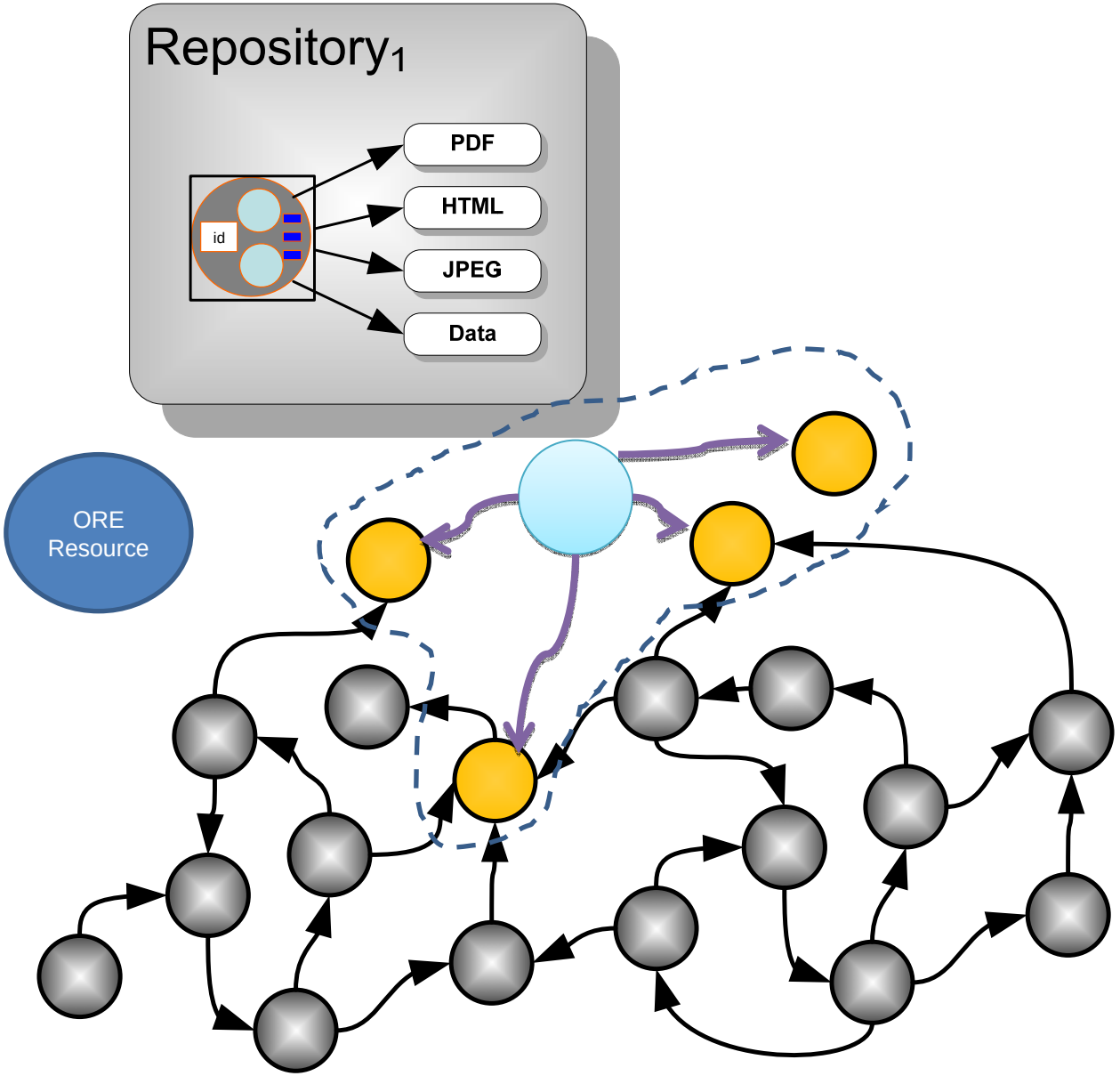


ORE Resource, ORE Manifest

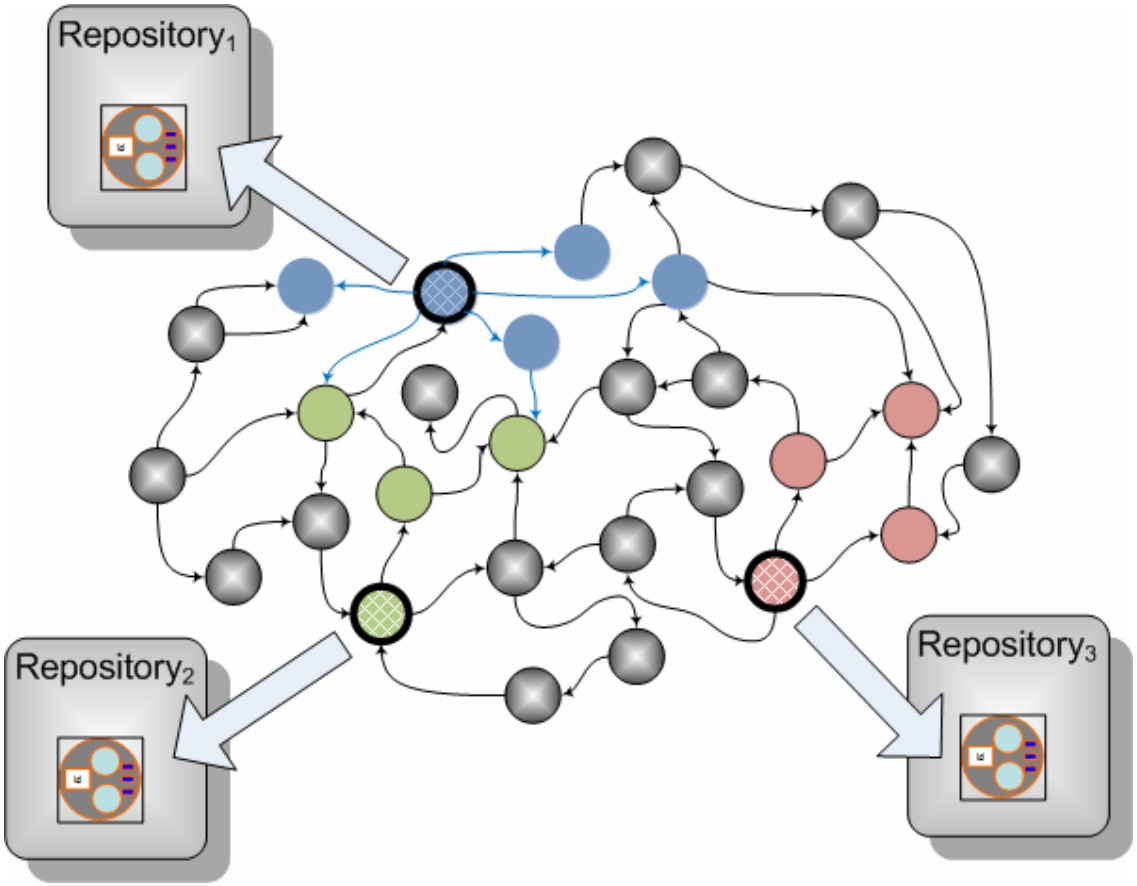


Serialization of the ORE Manifest

- RDF/XML
- Atom
- OAI-PMH
- DIDL


Repository1
PDF
HTML
JPEG
Data
id
ORE
Resource

ORE Resources and manifests reveal compound objects on the web



ORE and the path to a Process-oriented Scholarly Communication System

- Decompose the traditional process (Roosendaal & Geurts)
  - Registration (establish intellectual priority of result)
  - Certification (certify quality and validity of result)
  - Awareness (ensure accessibility)
  - Archiving (ensure availability for future use)
  - Rewarding (means to support tenure, promotion, compensation)

And more…

- Add new services to the mix
  - Workflow
  - Collaborative functions (e.g., annotation, re-use)
  - Data mining and analysis
  - Preservation monitoring and migration
- The result: services cooperate to turn data into information and knowledge.

Analysis of rich knowledge networks

- Topic detection
- Quality and influence
- Evolution of ideas over time

Conclusion

- The web, institutional repositories, data repositories, etc. provide the building blocks for new knowledge networks
- Building these network requires common models and protocols for exchange of information about complex information units
- This infrastructure will provide new ways to share information, knowledge, and wisdom