Open Archives Initiative – Object Reuse and Exchange Report on the Technical Committee Meeting, January 11,12 2007

Edited by: Carl Lagoze & Herbert Van de Sompel

1 Venue

Butler Library, Columbia University, New York City

2 Final Agenda

Thursday, January 11

Time	What	Details	Who
9:00-9:30	Welcome	Logistics Motivation Agenda Overview Goals	Herbert and Carl
9:30-10:30	Round-table introductions	Who are you? How does this relate to your experience/job/background? What constituencies do you represent?	TC members
10:30-10:45	Break		
10:45-12:30	Problem Overview	White Paper	Herbert and Carl
12:30-13:30	Lunch	Provided	
13:30-15:30	Feedback (1)	Prepared remarks about white paper & ORE Problem space	TC members
15:30-15:45	Break		
15:45-17:00	Feedback (2)	Prepared remarks about white paper & ORE Problem space	TC members

Friday, January 12

Time	What	Details	Who
9:30-10:00	Review of Thursday & Planning for Friday	Progress Disagreements How to proceed?	Herbert and Carl (with TC members)
10:00-12:15 (with break)	Reaching Consensus (1)	Problem Statement Scoping Target Participants	All

12:30-13:30	Lunch	Provided	
13:30-17:00 (with break)	Reaching Consensus (2)	Use Cases Definitions Entities	All
16:00-16:30	Wrap up	What has been accomplished? When do we meet next? What we need to do between meetings? Reactions	Herbert and Carl (with TC members

3 Attendees: Introductions, Affiliations, Activities and Interests

TIM DILAURO

- John Hopkins University, Library Digital Programs/Digital Knowledge Center of the Sheridan Libraries
- Active interest in: interface between data and services, big science data repositories, humanities data repositories, digital manuscripts, entity matching, digital preservation
- Ongoing IMLS project with National Virtual Observatory regarding publications that include data

DAVE FULKER

- University Corporation for Atmospheric Research
- Co-founder of NSDL
- Founder of UNIDATA effort regarding exchangeable scientific data objects
- Active interest in: real-time data, data exchange with appropriate semantics, data visualization

TONY HAMMOND

- Nature Publishing Group (NPG), New Technology Group
- 10 years of experience in STM publishing
- Active involvement in specifying NISO OpenURL Standard, info URI
- Active involvement in efforts related to deploying the publishing infrastructure for the digital age including: identifiers (DOI, info URI), multiple resolution of DOIs, OpenURL (NISO standard and pre-standard), SRU, RSS
- Active involvement in NPG's Connotea, and Open Text Mining (OTMI) efforts

PETE JOHNSTON

- EduServ Foundation
- Proxy for Andy Powell
- Digital Library and eLearning environments
- Active involvement in efforts related to deploying the publishing infrastructure for the digital age including DCMI
- Previously worked at UKOLN

- Strong interest in semantic web, web architecture and their application to scholarly and learning repositories

RICHARD JONES
- Imperial College, Institutional Repository effort
- Previously at University of Bergen (Norway) and Edinburgh University
- Active in DSpace community as committer and member of DSpace Architectural Review Group
- Active interest in Electronic Thesis and Dissertation, moving objects across repositories, linking data, workflow
- Member of JISC Common Repository Interface Group

CARL LAGOZE
- Cornell University, Computing and Information Science
- Active involvement in specifying OAI-PMH
- Active involvement in other interoperability efforts related to web information systems including Dienst, Dublin Core, Fedora, and the ABC metadata ontology.
- Ongoing research in new forms of scholarly communication including the use of machine learning methods to recognize compound objects on the web and analysis of hybrid social/bibliographic networks.

PETER MURRAY
- OhioLINK
- Large-scale system for statewide higher education access to A&I databases, full-text collections, catalogues
- Statewide hosted digital content repository
- Active in Fedora community
- Active interest in digital preservation, workflow, versioning

MICHAEL NELSON
- Old Dominion University, Computer Science Department
- Previously responsible for NASA's Digital Library efforts
- Active involvement in specifying OAI-PMH
- mod_oai project to bring the power of the OAI-PMH to Web servers
- Ongoing research efforts in the realm of digital preservation

RAY PLANTE
- University of Illinois at Urbana-Champaign, National Center for Supercomputer Application, National Virtual Observatory
- Supporting astronomical research over the network: publishing data, resource registries, discovery of services, embedding data products in publications

ROBERT SANDERSON
- University of Liverpool, Department of Computer Science
- Active involvement in specifying SRU/W

- Active involvement in efforts related to deploying the scholarly communication infrastructure for the digital age including: SRU/W, unAPI, OAI-PMH, OpenURL
- Member of UK National Centre for Text Mining
- Close collaboration with San Diego Supercomputer Center
- Active interest in: text and data mining, XML, information retrieval, GRID computing, digital manuscripts

HERBERT VAN DE SOMPEL
- Los Alamos National Laboratory, Research Library, Digital Library Research & Prototyping Team
- Active involvement in specifying OAI-PMH, OpenURL (NISO standard and pre-standard), info URI, MPEG-21 DID and DII
- Active involvement in efforts related to deploying the scholarly communication infrastructure for the digital age including: OAI-PMH, info URI, OpenURL, SFX linking server
- Long-standing interest in helping establish the technical foundations for scholarly communication in the digital age
- Ongoing research efforts in the realm of digital preservation (NDIIP), digital scholarly communication (Pathways), alternative quality assessments of scholarly communication units (MESUR), repository and repository federation architecture (aDORe)

SIMEON WARNER
- Cornell University, Computing and Information Science
- Long-standing involvement with arXiv.org
- Strong interest in making arXiv an integrated part of the global research environment
- Active involvement in specifying OAI-PMH
- Involved in projects related to OAI-ORE: NSF Pathways (scholarly communication as a cross-repository workflow) and remote submission to arXiv (from CNRS France)

JEFF YOUNG
- OCLC Research
- Active involvement in efforts related to deploying the publishing infrastructure for the digital age including: registries (OpenURL, info URI), identifiers (info URI), OAI-PMH, OpenURL, SRU, RSS
- Developer of some of the most widely used toolkits for OAI-PMH and OpenURL deployment

4 Meeting Results

4.1 Intellectual property and OAI-ORE work

The TC made two decisions regarding the results of the OAI-ORE work:

1. All public documents will be covered under a Creative Commons license
2. The group made a handshake agreement that there would be no attempt to patent results from the collective OAI-ORE effort

4.2 Definition of OAI-ORE Objectives

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.

4.3 Compound Digital Objects

Digital content with multiple components that may vary on multiple axes including:

- Content (semantic) types including:
  - Text, image, video, audio
  - 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 the Global Digital Format Registry (GDFR)
- Network locations including content from:
  - Institutional repositories
  - Scientific data repositories
  - Social networking sites
  - General web
- Relationships including:
  - Lineage
  - Versions
  - Derivations

See Figure 1 for a sample (and simple) compound digital object. It depicts an imaginary object from arXiv.org from which several *views* are available. Informally a view can be considered an alternate presentation of the content or meaning of the digital object. In the case of this example these views are the article in various formats, a splash page in HTML, and Dublin Core metadata.

Figure 2 provides a more elaborate and complete depiction of the same object, which now contains a *component.* Informally a component can be considered a subpart of the main object; for example, a dataset, or a chapter. In addition, the figure shows a relationship (for example, a citation) to an external object. Note that the component is logically within the boundary, it is a part, of the primary object, whereas the cited object is outside the logical boundary.

From here onwards, we refer to both components and views of a compound digital objects as *members* of the compound digital object.

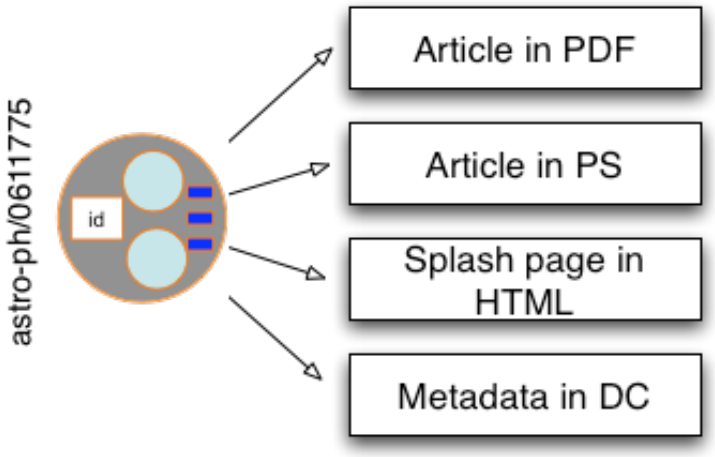

Figure 1: A sample (simple) compound digital object

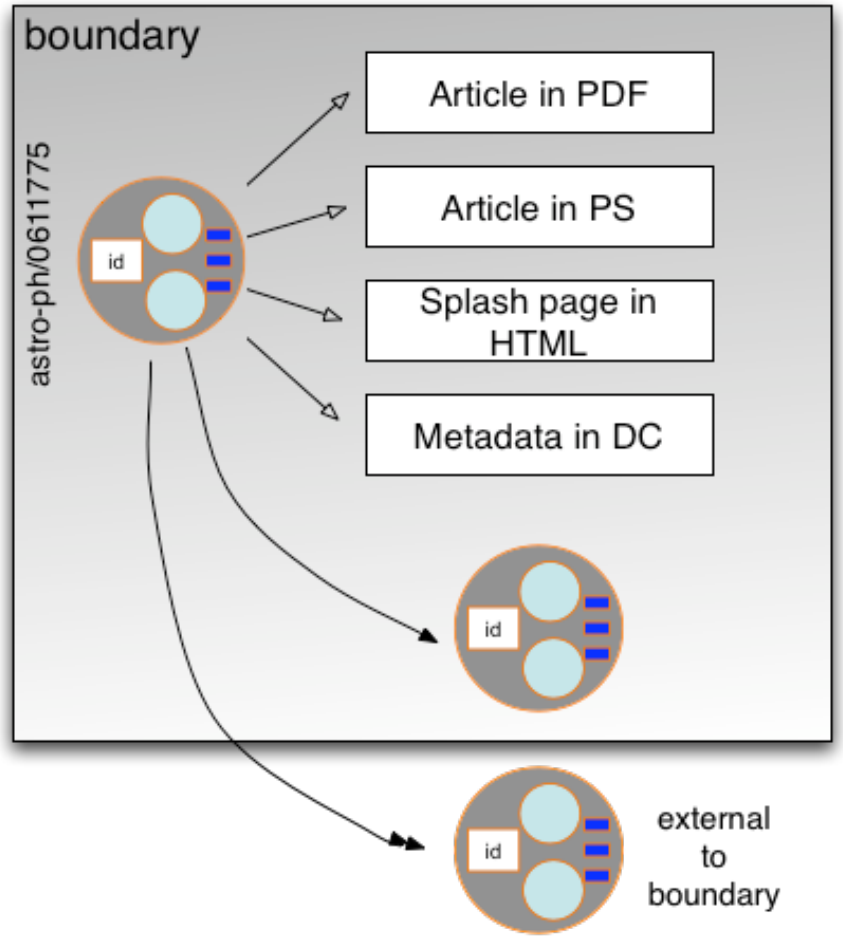

Figure 2: A more complete picture of a compound digital object

4.4 Scope of Use Cases and Applications

The standards and protocols endorsed, profiled or defined by OAI-ORE are intended to facilitate use and reuse of these compound digital objects and their components in the context of workflows supporting research and learning, while supporting notions of reference-ability, longevity, integrity, certification, and reproducibility that are foundations of scholarly communication. These workflows include a variety of services and applications that:

- facilitate discovery of these objects,
- reference (link to) these objects (and their members),
- obtain a variety of disseminations of these objects,
- aggregate and disaggregate these objects,
- enable processing by automated agents

4.5 Target adopters of OAI-ORE standards and protocols

Systems that manage content including:

- Institutional repositories
- Research-group and managed personal (ePortfolio) repositories
- Discipline-oriented repositories
- Publisher repositories
- Dataset repositories
- Cultural heritage repositories, including digitized museum and art collections
- Learning object repositories
- Digital and digitized text and manuscript collection management sytems

Systems that consume that managed content including:

- All the aforementioned systems because, in many cases, those systems also ingest content from other systems and/or provide services over the content they manage.
- Search engines
    - Specialized/discipline-specific
- General web applications
- Productivity tools including
    - Authoring tools
    - Citation management
- Indexing and abstracting services
- Aggregators
- Collaborative environments
- Object-based social network applications
- Data processing applications including
    - Data mining
    - Text mining
    - Scientific analysis tools
- Graph analysis applications including
    - Link checkers
    - Object-based Citation checkers
- Preservation services and other data management services
- Research assessment services
- Report generation

- Workflow tools

4.6 Proposed use cases

Over the next month the members of the technical committee will collaboratively (using a Wiki) develop a set of use cases against which the protocols and specifications will be tested. These use cases should be constructed as follows:

- One paragraph that describes a usage scenario and motivation from the perspective of an end user without describing the technical details of how the scenario is implemented. We note that each use case will inevitably require applications, registries, and services that extend beyond the scope of the OAI-ORE work.
- Commentary on how the protocols and specifications developed by OAI-ORE will facilitate the implementation of the use case.

The draft use cases and the TC members responsible for fleshing them out are as follows:

- Find, collect, analyze, relate, and publish data-oriented scholarly objects – Dave Fulker, Ray Plante
- Find, collect, analyze, relate, and publish text-oriented scholarly objects – Rob Sanderson, Tony Hammond
- Preservation of compound digital objects – Tim DiLauro, Michael Nelson
- Remote submission of compound digital objects – Simeon Warner, Jeff Young, Richard Jones
- Citation management – Herbert Van de Sompel, Tony Hammond
- Object equivalence recognition (de-deduping) to aid resource discovery – Pete Johnston, Andy Powell
- Graph-based quality assessment with eScience focus – Carl Lagoze, Peter Murray, Ray Plante

4.7 Relationship to web architecture

The TC spent considerable time analyzing the components of web architecture – URIs, *resources*, *representations* (see Figure 3) – and its implementation via HTTP. The analysis was based on a general consensus that results of OAI-ORE should align and not conflict with web architecture and should use that architecture as a foundation for the standards and protocols developed. As much as possible the OAI-ORE standards should be a specialization of existing web architecture concepts with the goal of meeting the requirements of the target adopters of OAI-ORE standards in the context of the defined use cases.

In the remainder of this report the usage of the terms *resource* and *representation* will be restricted to their definition in the web architecture (see http://www.w3.org/TR/webarch/).

Our analysis of the web architecture led to consensus around the following observations:

- The graph described by the web architecture document contains two types of nodes:
    - resources:
        - These are first-class objects with a standalone identity (URI)

        - They can be the target of links (or references), and the links may be typed to indicate the nature of the relationships between source and target. The following considerations apply for link typing:
            - It is not widely adopted/exploited in general Web applications;
            - Controlled vocabularies to define link types are not widely adopted;
            - The manner in which to express link types is specific to document formats such as HTML and XML that have their own link tags.
    - representations:
        - These are second-class objects that are identified only via the resource that they represent. A representation is the result of applying a service to an identifier of a *resource*. Since there is one-to-many relationship between a resource and its representations, the representations have no unique identity. As a result:
            - There is no means to link (or reference) a representation.
            - They are only accessible through protocol negotiation.
- Although specific document formats (e.g. HTML, XML) express a notion of composite documents, the web architecture does not itself address the question of how to describe the composition of a compound digital object that aggregates a number of resources in multiple content types. Specifically, it does not address the following:
    - What are the boundaries of a compound digital object, where that boundary contains a finite aggregation of members and a finite set of relationships among those members?
    - What are the types of relationships between the aggregated members of a compound digital object?

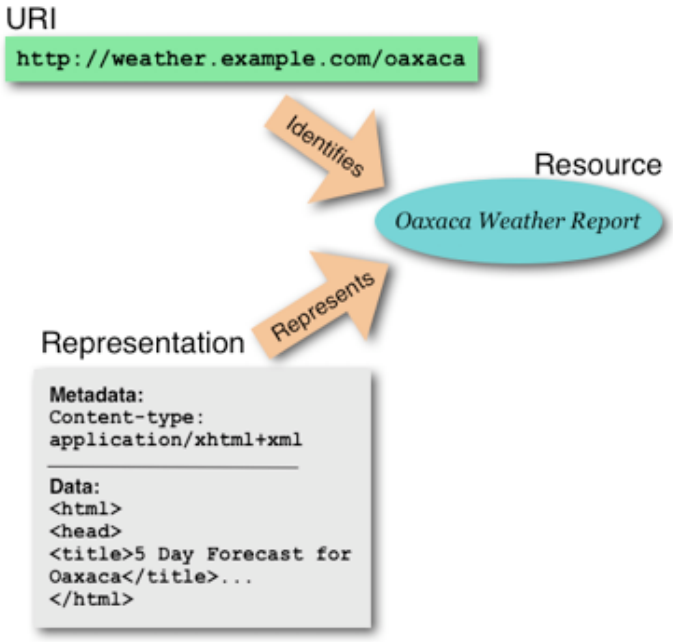

Figure 3: Web architecture (see http://www.w3.org/TR/webarch/)

4.8 Expressing compound digital objects using Web architecture concepts

As a result of this analysis we agreed on the following points that frame the requirements of the OAI-ORE work in relation to the design features of the web architecture.

A key requirement of the OAI-ORE scenarios for use and reuse of compound digital objects is the ability to unambiguously identify and reference both the compound digital object and its components (e.g.. dataset as sub-object as in Figure 2; chapters of a book; sections of a paper). A further requirement is the ability to unambiguously identify and reference "views" of these components (i.e. the pdf version or the ps version of a journal article). As a result, a compound digital object, its components, and these "views" must be modeled as resources (i.e. they must be identified by URIs) if they need to be available for re-use. **Error! Reference source not found.**Figure 4 contrasted with Figure 5 illustrates the need for this. In the former, the views are representations, without unique identity. In the latter, the views have been broken out to have a one-to-one correspondence to identifiable resources. Note that in both figures, the notion of the boundary of the compound digital object is not expressed.

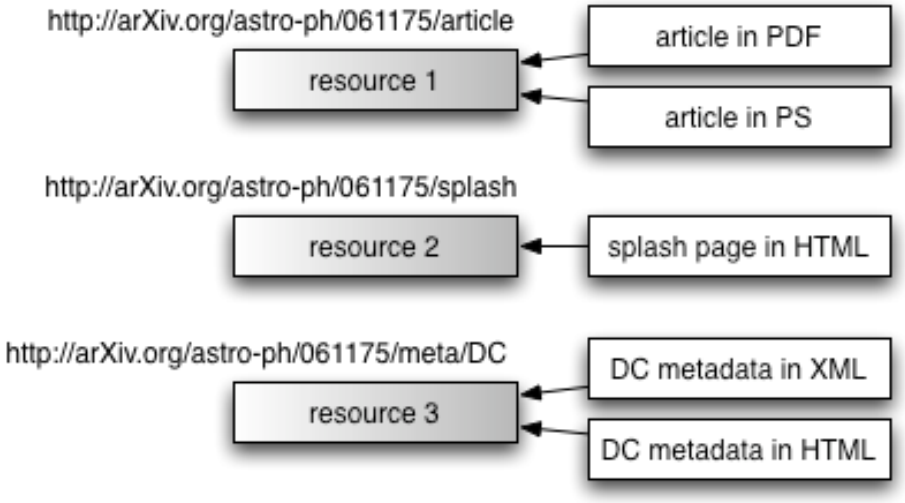

Figure 4: Compound digital object modeled according to Web architecture; specific views not reference-able; boundary of compound digital object not expressed

It follows that the compound digital objects that are the subject of the OAI-ORE effort must be bounded aggregations of resources and their relationships. This aggregation must itself be a first-class identifiable object, because the aggregation corresponds to a logical digital object, which should be reference-able, linkable, etc. Therefore, it must be rooted by an identifying resource, which we refer to as the *ORE resource* (see Figure 6).

The URI of an ORE resource serves as the access point for service requests upon the aggregation. A specific service request on this URI returns a representation that describes the members of the aggregation (including the ORE resource) and their relationships. In the remainder of this document, we refer to this representation as the *ORE representation* (see Figure 6), and to the aggregation described by an ORE representation as an *ORE aggregation*.

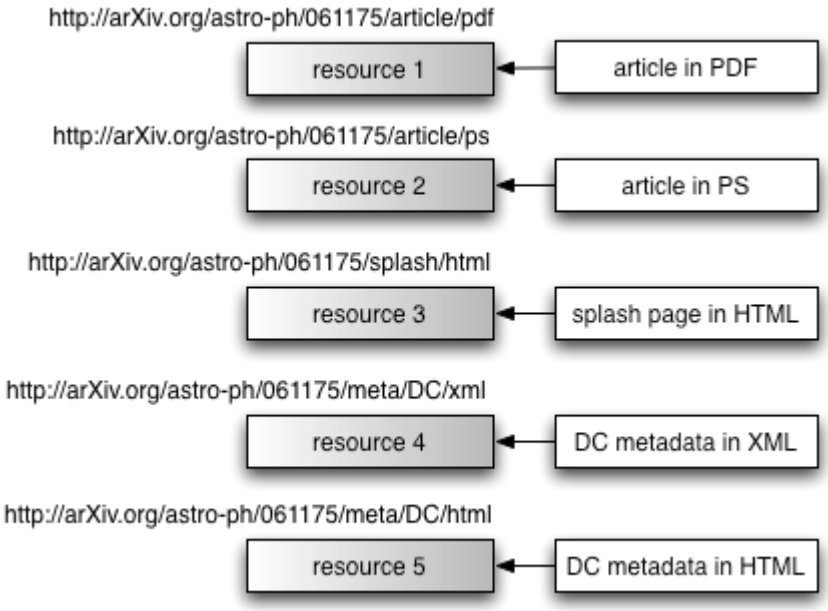

Figure 5: Compound digital object modeled according to Web architecture; specific views reference-able; boundary of compound digital object not expressed

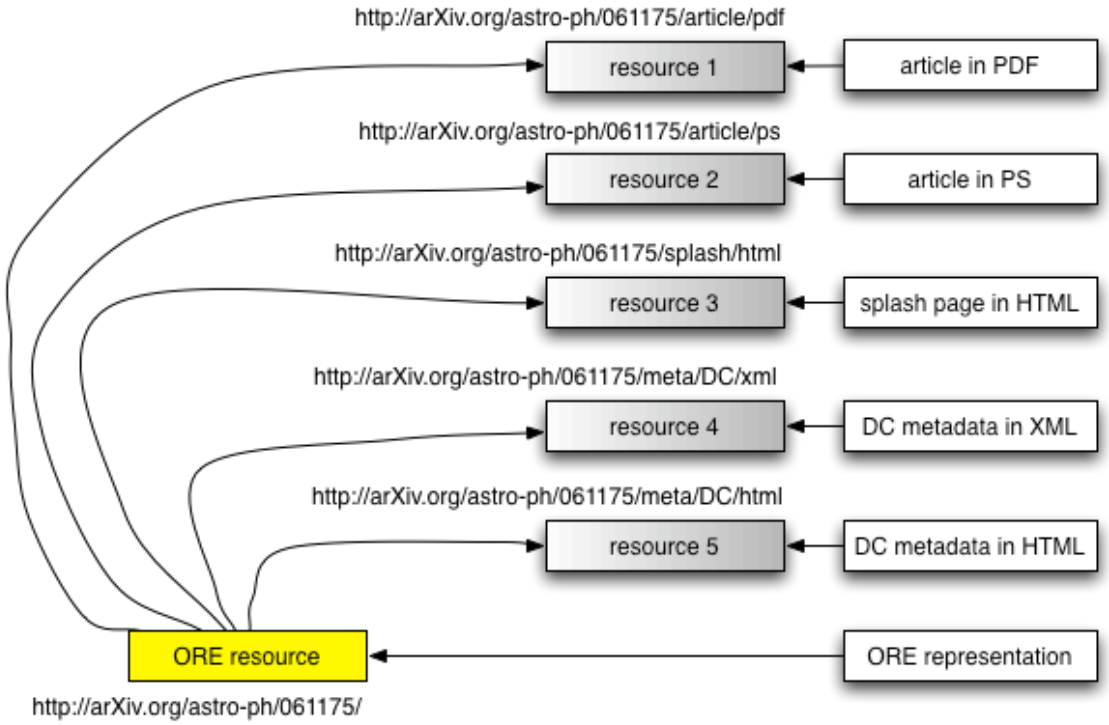

Figure 6: Compound digital object modeled according to Web architecture; specific views referencable; boundary of compound digital object expressed via an ORE representation.

Specifically, ORE aggregations form a sub-class of the set of possible aggregations of resources. This sub-class has the following distinguished properties:

- *An ORE aggregation has a boundary*: A key requirement of the OAI-ORE scenarios for use and reuse of compound digital objects is the ability to describe the boundaries – the finite set of resources and relationships – that correspond to compound digital objects. Note that these relationships internal to the boundaries of the ORE aggregation should have defined types that are specifically of a sub-class of all possible relationships - *intra-aggregation relationships (*Figure 7*)*. Therefore OAI-ORE must define a standardized model that can describe this boundary and that can be instantiated for ORE resources.
- *The resources in an ORE aggregation may have relationships to resources external to the aggregation*: A key requirement of the OAI-ORE scenarios for use and reuse of compound digital objects is the ability to express relationships between the members of a compound digital object (the resources in the ORE aggregation) and resources that are external to the aggregation. These relationships external to the boundaries of the ORE aggregation should have defined types that are specifically of a sub-class of all possible relationships - *inter-aggregation relationships (*Figure 7*)*. Therefore the model defined by the OAI-ORE must describe the typed relationships between resources in an ORE aggregation and external resources.

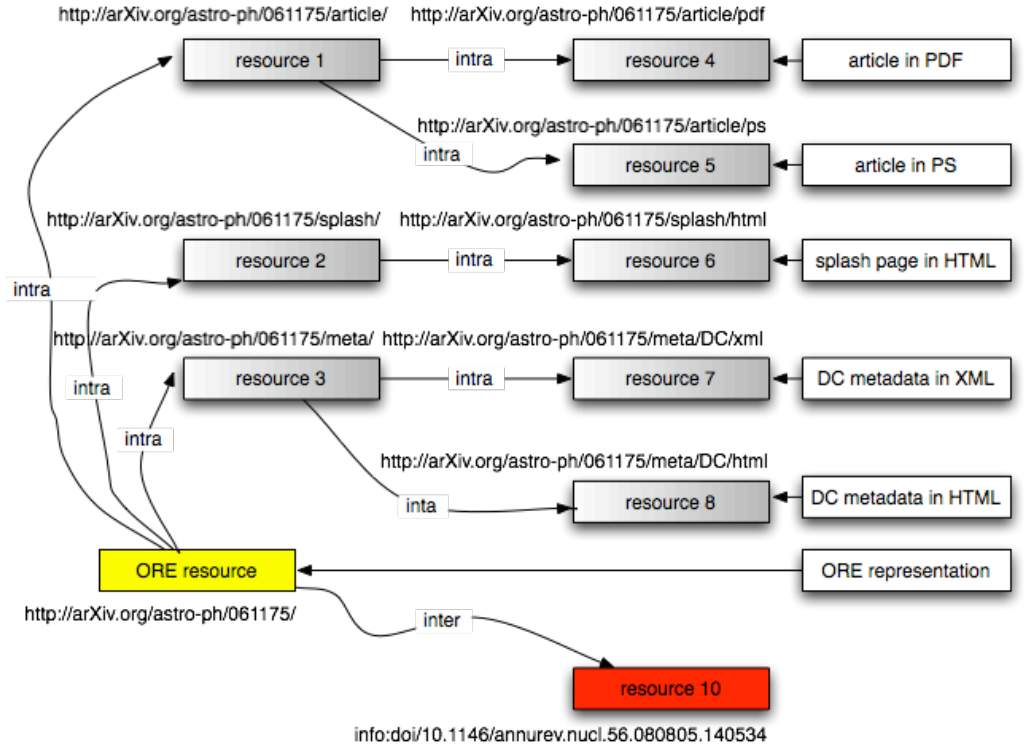

Figure 7: Compound digital object modeled according to Web architecture; specific views referencable; boundary of compound digital object expressed via an ORE represenation; inter and intra object relationships.

Representations should not explicitly be included in an OAI-ORE model for compound digital objects. Of course the end product of any web transaction applied to a resource is a representation. Since ORE service requests, described later are a sub-class of web transactions, they ultimately return representations.

4.9 A model for aggregating web resources and expressing their relationships

In order to fulfill the OAI-ORE requirements enumerated above, we **tentatively** define the following aspects of the OAI-ORE work:

I) Define a *model* (referred from here on as the *ORE Model)* that describes an ORE aggregation – a finite set of resources and the relationships among the resources of that aggregation – and the relationships between that ORE aggregation and its member resources and resources that are external to the ORE aggregation. The aspects of that model and the aggregations that it describes are as follows:

A) The model formally describes a connected sub-graph with nodes that are resources, with optional typing (e.g. content-type, media-type, etc.), and edges that are intra-aggregation relationships between the resources that fall into two classes:

1) *hasPart* that defines a containment relationship.
   (i) Examples are chapters of a book, sections of a research paper (e.g., abstract, introduction, etc.)
   (ii) This relationship is recursive; i.e., the target resource of a *hasPart* relationship can be the source of additional relationships.
2) *hasView* that defines a relationship whereby the target resource is a presentation or view of the source *resource*.
   (i) Examples are alternate formats of a scholarly paper (e.g., pdf, word, etc.) or a structured metadata (e.g., Dublin Core) presentation of that paper modeled as autonomous resources.
   (ii) The target of the *hasView* relationship is always a leaf node; i.e., it can never be the source of further relationships.

B) The connected graph has the following additional characteristics:

1) It may be a sub-graph of a larger connected graph since multiple digital objects may be logically aggregated within another digital object via the hasPart relationship.
2) It is a rooted graph with the root being a distinguished node that is an *ORE resource* with the following characteristics:
   (i) All other resources (nodes) in the sub-graph that is the respective ORE aggregation can be reached by following the directional relationships (*hasPart* and *hasView*) from ORE resource.
   (ii) The URI of that *ORE resource* is the URI of the logical unit that is the compound digital object.
   *(iii)*The ORE resource provides access to instances (serializations) of the model through OAI-ORE services as defined in the next section.

C) The ORE Model defines one other relationship, *hasRelationshipTo,* which expresses relationships between resources within the ORE aggregation (confined by the *hasPart* and *hasView* relationships) and *resources* external to the aggregation. We expect that

relationship will be specialized due to application or community based needs to accommodate semantics such lineage, derivation, citation, etc.

II) Define a format that can be used to serialize instances of the ORE Model. This format can be defined using schema mechanisms such as XML schema, OWL, etc.

III) Define a mechanism to associate instances of the model with ORE resources, and thereby provide the basis for services upon the defined aggregations of resources.

Figure 8 shows the compound digital object of the previous pictures expressed according to the above preliminary ideas regarding the ORE Model. Figure 9 shows an initial rendering of the ORE Model.

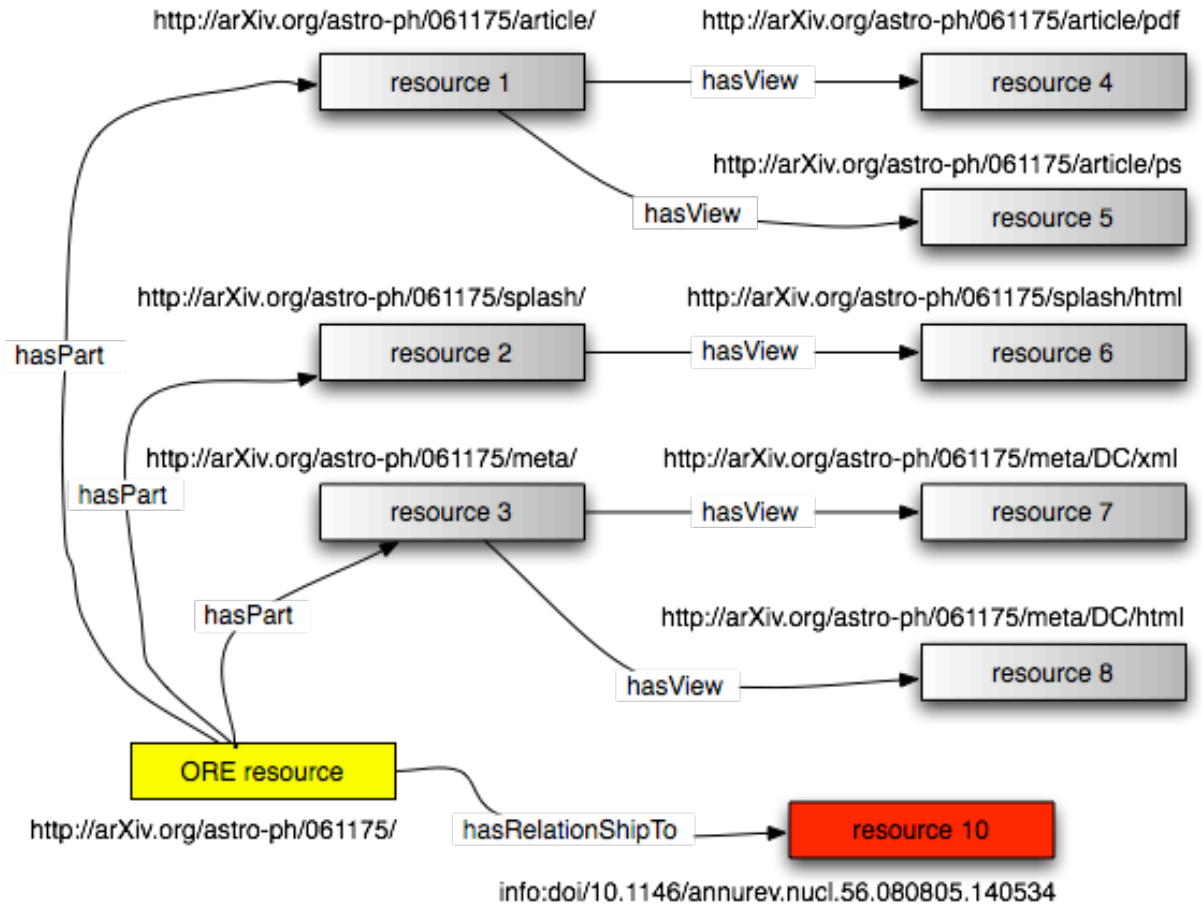

Figure 8: Compound digital object modeled according to (preliminary) ORE Model. Specific views referencable; intra (hasPart, hasView) and inter (hasRelationshipTo) relationships.

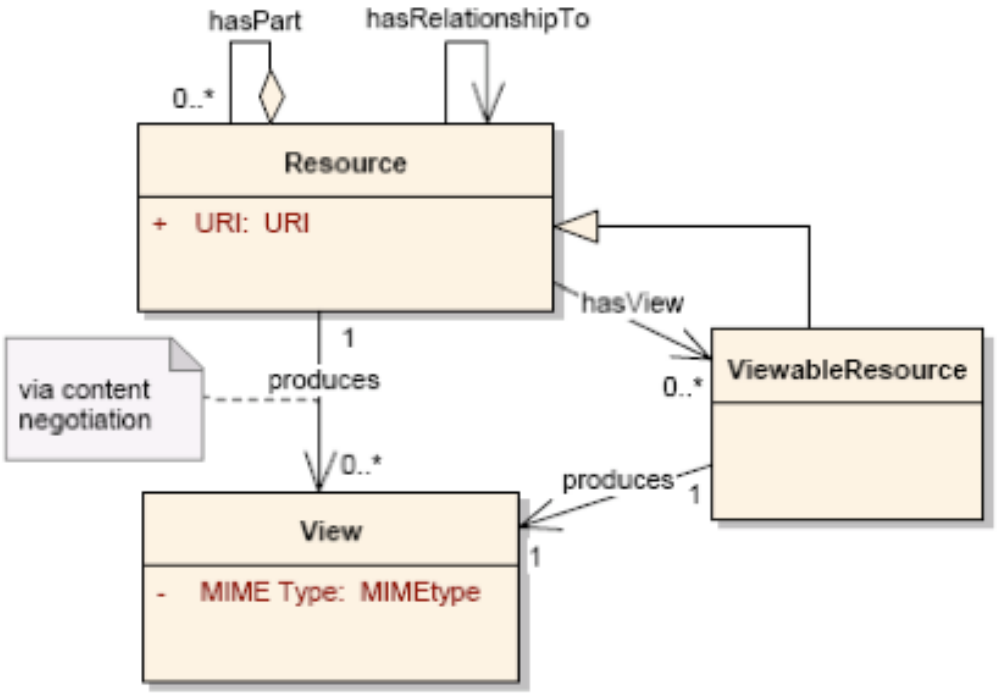

Figure 9: A preliminary rendering of the ORE Model

Collections and other aggregations

The model described above does not formally describe collections and similar resources. The group spent some time discussing these concepts, especially in the context of Pete Johnston's presentation. No formal agreement or consensus was developed around these concepts. Informally we agreed that they might be specialized forms of aggregations of objects that we wish to model (for example by typing resources). Further work clearly needs to be done in this area.

4.10 Defining OAI-ORE services

OAI-ORE services are transactions that exchange instances of the model described above. These instances are associated with ORE resources that, as described in the previous section, are the access point for operations on an aggregation of web resources. We define three classes of transactions that form the basis of the OAI-ORE service framework:

- *Harvest*: a request for a batch of instances that correspond to the ORE model from a set of ORE Resources. One typical application of this service is by robots from search engines.
- *Obtain*: A request for an instance that corresponds to the ORE Model from a specific ORE Resource. Typically this transaction will initiate an access transaction for the compound digital object or parts thereof.
- *Register*: A request to add new nodes or relationships to an ORE aggregation. Typically this will be take the form of adding a new compound digital object to a collection or repository.

5 Plans and Action Items

- Set up Wiki for collaborative work – assigned to Tim DiLauro for completion by January 19th
- Set up Connotea group for sharing of resources/citations – assigned to Tony Hammond for completion by January 19th
    - Everyone will set up accounts and send user name to Tony
    - All cites will be tagged oaiore
- Fleshed out use cases to be completed by February 2nd
- UML for proposed ORE model by February 13th
- Conference call – either Feb 12 or 13 for two hours starting at 11:00 GMT-5
    - Discuss use cases
    - Model review
    - Begin appropriate technologies discussion
- Two (monthly) additional conference calls leading up to next meeting
- Next meeting – two day meeting somewhere in the span of May 28-30, 2007.
    - Start implementation details