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

Carl Lagoze
Information Science, Cornell University
lagoze@cs.cornell.edu

Herbert Van de Sompel
Research Library, Los Alamos National Laboratory
herbertv@lanl.gov

***Note**:* This document is an exploratory white paper for internal use by the OAI-ORE Technical Committee, Advisory Committee, and Liaison Group. This is *not* intended for wider dissemination or publication. Contact the authors for permission to reuse this document.

1 Introduction and Motivation

Open Archives Initiative - Object Reuse and Exchange (OAI-ORE) is a new effort under the umbrella of the Open Archives Initiative. OAI-ORE is supported by a generous two-year grant from the Andrew W. Mellon Foundation, which began in October 2006. This funding is one outcome of an April 2006 meeting titled "Augmenting Interoperability across Scholarly Repositories"[1], which was sponsored by Mellon, Microsoft, the Coalition for Networked Information, the Digital Library Federation, and the JISC. The goal of this funding is to develop through an intensive international effort specifications for the exchange of information about Digital Objects between cooperating repositories and services. A web site containing up-to-date and legacy information about OAI-ORE is available at http://openarchives.org/ore.

As demonstrated by its positioning within the OAI framework, the goals of OAI-ORE are congruent with the original mission of OAI, formulated in 2001. The first sentence of this is "The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content". The original effort of OAI, the Protocol for Metadata Harvesting (OAI-PMH)[2], is just one approach to such interoperability, focusing on the exchange of structured metadata that is primarily bibliographic. OAI-PMH has achieved wide acceptance and is a de facto standard in the web-based information architecture. However, its metadata focus limits its utility in applications that focus more on the nature of the digital resource itself rather than structured descriptions about it. This resource-centric focus has become increasingly important as digital resources in e-Scholarship and cyberinfrastructure become more complex (i.e., multi-part, multi-media, distributed), rather than existing as a single (e.g PDF) file. We expect that over the long run, standards defined by OAI-PMH and OAI-ORE will co-exist within a spectrum of functionalities to share information in the networked environment.

Work on OAI-ORE takes place in the context of accelerating changes in the way that scholars produce, share, and access the results of their work and that of their colleagues. These changes entail the nature of scholarship and research itself, which has become more data-centric and collaborative, and in the mechanisms for storing and delivering the results of that scholarship, which include institutional repositories, tera-scale data repositories, and high-speed networks. These changes are also motivated by the broader set of social network applications that characterize the web today. The latest generation of scholars lives in a world where instant messaging, myspace[3], and youtube[4] are everyday communication vehicles. These *next- generation* scholars will demand that the *next-generation* scholarly environment mirror these new modes of collaboration.

[1] http://msc.mellon.org/Meetings/Interop/.
[2] http://openarchives.org/pmh.
[3] http://myspace.com.
[4] http://youtube.com.

This rapidly evolving context sets the stage for more advanced notion of scholarly communication that would fully compliment the way that scholars work. Such a system would allow flexible composition of distributed information units – text, data sets, image, etc. – and distributed services for the formation of new types of published results and new methods of collaborating. This would be a loosely coupled system based on an interoperability fabric that supported variety of workflows acting upon distributed units of scholarly information [13].

This future system will involve the coordination of multiple applications, services, registries, and the like. OAI-ORE is intended to provide the specifications of a base infrastructure layer that enables this coordination. As such, it focuses on the interaction of repositories, which we consider the fundamental building blocks of a future digital scholarly communication system. Specifically, this infrastructure should provide more consistent means to:

- Support the creation, management, and dissemination of new forms of scholarly resources.
- Facilitate the discovery of these new forms of scholarly resources.
- Establish reference links to resources
- Provide access to the representations of resources.
- Aggregate and disaggregate resources
- Re-use resources, or parts thereof, beyond the boundaries of the holding repository.

To develop the interoperability infrastructure to fulfill these requirements, the work plan for OAI-ORE over the two-year period from October 2006 through September 2008 will include the following components:

- The formation and management of an international working group to develop a set of specifications for a repositories interoperability framework. These specifications will describe common data models and interfaces for exchange of information based on these data models.
- The establishment and management of an experimental deployment community that will exercise the interoperability fabric in a variety of milieus, with the goal of empirically proving the interoperability fabric before wide-scale deployment efforts.
- The establishment of a sustainable community to support the widespread deployment and management of the standards fabric, and thereby make real and substantial changes in the scholarly communication system
- The development and publication of reference implementations of the standards built upon common repository packages such as aDORe , DSpace, ePrints , and Fedora . These implementations will be made available on the project web page as Open Source software under the terms of the Educational Community License.

2 OAI-ORE Organization

OAI-ORE will follow the same organizational model that led to the successful development, experimentation, and deployment of OAI-PMH [12]. This model combines

the leadership of an executive team that leads an international working group with consultation of an advisory committee of international experts.

The executive team for OAI-ORE is the same as OAI-PMH, Carl Lagoze[5] of Cornell Computing and Information Science and Herbert Van de Sompel[6] of the Los Alamos Research Library. Both Lagoze and Van de Sompel are supported by the Mellon grant for 50% of their total effort.

The OAI-ORE Advisory Committee (AC) provides strategic guidance regarding ORE directions and goals. They also provide outreach to a range of communities. The AC receives first hand information on the progress made by the ORE Technical Committee giving them the opportunity to provide early feedback on standards and protocols developed by the ORE. Members of the AC are leaders in various communities that we hope will deploy the products of ORE work. The members of the AC are:

- 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
- Clifford Lynch (chair) - Coalition for Networked Information
- Liz Lyon - UKOLN
- Peter Murray Rust - University of Cambridge
- Jim Ostell - National Center for Biotechnology Information
- Sandy Payette - Cornell University and Fedora
- Robby Robson - Eduworks
- MacKenzie Smith - MIT Libraries and DSpace
- Leo Waaijers - SURF Platform ICT and Research

The OAI-ORE Technical Committee (TC) is responsible for developing the data models, protocols, and standards that will be the product of ORE work. To accomplish this work the TC has periodic face-to-face working meetings and conference calls. Key members of the TC assist the OAI-ORE Executives in actual specification writing, which are then open for review by the entire TC. Members of the TC are key technical participants in target communities of the ORE work, thus ensuring that the resulting specifications and standards will have broad application. The members of the TC are:

- Les Carr, University of Southampton
- Tim DiLauro, Johns Hopkins University
- David Fulker, UCAR

[5] http://www.cs.cornell.edu/lagoze/.
[6] http://public.lanl.gov/herbertv/.

- Tony Hammond, Nature Publishing Group
- Richard Jones, Imperial College
- Peter Murray, OhioLINK
- Michael Nelson, Old Dominion University
- Ray Plante, NCSA and National Virtual Observatory
- Andy Powell, Eduserv Foundation
- Rob Sanderson, University of Liverpool
- Simeon Warner, Cornell University
- Jeff Young, OCLC

Finally, the OAI-ORE Liaison Group (LG) provides a linkage between the ORE work and international activities that share similar objectives. LG members act as a communication bridge between their respective projects and ORE technical work. LG members have access to ORE mailing lists and have early access to developing specifications. We expect that the LG will expand as the project matures. The members of the LG are:

- Leonardo Candela - DRIVER
- Tim Cole - DLF Aquifer and UIUC Library
- Rachel Heery - JISC
- Savas Parastatidis - Microsoft
- Thomas Place - DARE and University of Tilburg
- Robert Tansley - Google, Inc. and DSpace

3 Motivating Problems

Over the last decade and a half the architecture of the web has proven to be a marvelously flexible and extensible mechanism for general-purpose information dissemination. In the manner of TCP/IP, it provides a foundation layer upon which more specialized layers, tailored for selected communities and applications, can be built. The focus area of OAI-ORE, scholarly communication, is one of these areas and the specifications we develop over the next two years are intended to establish an infrastructure layer on top of the web that address the needs of this application domain.. As was the case with OAI-PMH we anticipate that the specifications we produce will also impact application areas beyond scholarly communication. In addition, we hope that the solutions we develop will continue to apply as the web evolves beyond its existing technologies.

The remainder of this section illustrates a number of cases in which the web architecture falls short of the needs of a high-integrity vehicle for storage and delivery of scholarly materials. This list is not meant to be exhaustive, but it illustrates that the existing web architecture is not sufficient for even for the most basic needs of scholarly communication. Many of these problems have their roots in missing functionality in HTTP, the core protocol of the web, and in URIs, the identification mechanism for the web. We will return to these motivating problems in a following section to describe how OAI-ORE might address them.

3.1 Exposing resources to robots

Search engines based on robot crawlers – Google, Yahoo, Windows Live Search – are by far the dominant applications on the web today. By definition then a measure of the utility of the web for scholarly activities must be the ability to use these search engines to find scholarly results. Andy Powell of Eduserv reports in his blog[7] an experiment that compares the effectiveness of the general Google search engine relative to a Google custom search engine offered by the OpenDOAR directory[8] for finding open access scholarly papers. Underscoring the importance of search engine exposure he notes: "First and foremost, our 'resource discovery' efforts should centre on exposing the full text of research papers in repositories to search engines like Google and on developing Web-friendly and consistent approaches to creating hypertext links between research papers."

The full results of the experiment are out of the scope of this report, but he notes a number of questions about the manner in which search engines treat scholarly documents in open access repositories. The first of these is: "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?" He notes that "that some repositories are only exposing the abstract page, not the full-text."

The core of this problem, which we will elaborate on later, is the fact that the web architecture does not give search engines a means of mapping from a URI that identifies a scholarly paper to the multiple *representations* (or dissemination types) for that paper. In even simple cases, where the scholarly publication is only a PDF file, many repositories such as the arXiv[9] expose additional representations of this document including a "splash page" (which in many cases is a human-oriented UI to the representations and metadata for the document), another file format for the document (e.g., PostScript), and an OAI-PMH based metadata record[10]. This problem will only become more complex as repository systems evolve from storage of simple single-file documents to aggregations of datastreams with both a variety of media types and a variety of intellectual content types including papers, datasets, simulations, software, etc. For example, consider the publications in the The Smithsonian/NASA Astrophysics Data System[11] that bind together text from multiple repositories, including the arXiv, with data sets from distributed data stores.

Some search engines may need to access all representations. Others that are more text oriented may need to access only those that are textual. And, others that are specialized such as the US National Virtual Observatory Datascope[12] may need to access just selected data components. We argue then that a core problem that we must address in

[7] http://efoundations.typepad.com/efoundations/2006/10/pushing_an_open.html
[8] http://www.opendoar.org/
[9] http://www.arxiv.org.
[10] For example, consider http://arxiv.org/abs/cs.DL/0610031 (splash page), http://arxiv.org/pdf/cs.DL/0610031 (pdf), http://arxiv.org/pdf/cs.DL/0610031 (ps), and http://arxiv.org/oai2?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:arXiv.org:cs/0610031 (OAI-PMH dublin core).
[11] http://adswww.harvard.edu/.
[12] http://heasarc.gsfc.nasa.gov/cgi-bin/vo/datascope/init.pl.

OAI-ORE is a consistent manner for repositories to expose the nature of resources and their representations to search engines.

3.2 Referencing resources

Citation is a fundamental aspect of scholarly communication. It helps establish the credibility of claims and distinguishes claims of new results from those already made in earlier work. This makes citation a primary tool for distinguishing plagiarism from the accepted practice of building on existing work. Citation analysis subsequently provides important data about the influence of scholars' work. Thus, the ability of the networked environment to accurately and persistently represent citation links is vital if the web is to be the primary environment for scholarly communication.

But, in the same blog entry, Andy Powell asks "Are we consistent in the way we create hypertext links between research papers in repositories?" He goes on to state that links are essential information for the ranking mechanisms in all modern search engines. In PageRank[3], hubs and authorities [10], and other ranking methodologies, the structure of hyperlinks to a specific web resource plays an important role in determining the "importance" of a resource.

Due to the "representation problem" noted above, researchers have a variety of options as they link from their paper to reference papers. In some cases, a reference may lead to the splash page, in another to the descriptive metadata, and in another to the full-text in PDF. Since crawlers are rarely able to "dedup" these reference (that is, determine that they actually reference the same resource), the actual weighting of the reference from the search engines' perspective is reduced. As Andy Powell notes: "If we could agree on a consistent way of linking to materials in repositories, we would stand to improve the visibility of our high-quality research outputs in search engines like Google."

3.3 From metadata harvesting to resource access

Metadata and metadata harvesting (i.e., OAI-PMH) in the web context has by and large been used to provide search and discovery functionality. Studies indicate that, despite support in OAI-PMH for multiple metadata formats, simple unqualified Dublin Core is in many cases the only format support by OAI-PMH data providers [5]. Correspondingly, most OAI-PMH service providers follow a "union catalog" model of aggregating metadata and making it searchable – for example, OAIster[13] and NSDL[14].

In the majority of cases, however, search is merely a precursor to access. And herein is where the problem arises. As noted by Chavez et al in a recent D-Lib paper [4] about the DLF Aquifer project[15]; "Metadata records harvested using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) are often characterized by scarce, inconsistent and ambiguous resource URLs. There is a growing recognition among OAI service providers that this can create access problems and can limit the range of services offered." Partly this is due to a weakness in unqualified Dublin Core, which provides no

[13] http://oaister.umdl.umich.edu/o/oaister/.
[14] http://www.nsdl.org.
[15] http://www.diglib.org/aquifer/.

mechanism for disambiguating among multiple "identifier" elements. However, the problem is also the result of the inadequacy of typing mechanisms for digital content, which should provide the means for distinguishing among multiple URLs associated with a resource described by a descriptive metadata record.

In one sense, this is another version of the multiple representation problem described in earlier sections. However, as the products of digital scholarship evolve towards compound, multimedia resources, the multiple representation problem becomes more complicated.

First, current content description, or typing, mechanisms on the web such as MIME Types[16] or even developing type registries such as the Global Digital Format Registry (GDFR)[17] do not have the power to express the semantics of these asset actions. An advanced scholarly communication environment on the web will need more expressive typing mechanisms and registries for those various types.

Second, as Chavez alludes to, services play an important role in the future scholarly environment. We envision an infrastructure in which representations are not only statically produced, but are dynamically available via a variety of web services that transform, analyze, visualize, etc. The "type" information for a representation needs then to describe these services, their provenance, legacy, etc.

3.4 Re-using resources and their representations

Reuse and refactoring of existing information has become increasingly popular in the context of what is called "Web 2.0". These are often characterized as a *mashup*. According to wikipedia "A *mashup* is a website or application that seamlessly combines content from more than one source into an integrated experience".[18] The growth of sites such as flickr[19] and youtube[20], when combined with the already massive corpus of "standard" web content provides a fantastic foundation of information from which mashups can be constructed. Furthermore, as massive book scanning services such as Google Book Search[21] and the Open Content Alliance[22] come on line, the opportunities for imaginative reuse of content will expand tremendously.

When applied to the scholarly process, these mashups give new meaning to the phrase "standing on the shoulders of giants", which is commonly used to characterize the manner in which new results incrementally build on and extended pre-existing results. The increasing online availability of research results – papers and data – in institutional repositories, publisher archives, and grid storage provides exciting new opportunities for reuse of existing results. However, this reuse must be easy and transparent, regardless of the source of the building blocks. Currently mashups are constructed using ad hoc or

[16] http://www.iana.org/assignments/media-types/.
[17] http://hul.harvard.edu/gdfr/
[18] http://en.wikipedia.org/wiki/Mashup_%28web_application_hybrid%29
[19] http://www.flickr.com/.
[20] http://youtube.com/.
[21] http://books.google.com/.
[22] http://www.opencontentalliance.org/.

propriety techniques such as the Google Maps API[23] or Amazon Web Services.[24] This is problematic in the scholarly domain where the building blocks for mashups are distributed across heterogeneous repositories.

In addition to scenarios in which scholars pull together existing building blocks, we'd also like to encourage the export or upload of scholarly resources to value-enhancing services. Consider slideshare[25] for example. As noted by Brian Kelly, the UK Web Focus: "… 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[26]."Like other "social networking services" it supports the development of communities of interest and their 2nd order products (annotations, discussions, reviews, etc.) around core resources, in this particular case slide presentations, which are frequent products of research and teaching.

Making this all possible will require a common set of interfaces (data models and APIs) that support both the ability to *obtain* components of existing documents, for construction of the mashup, and *put*, or request deposit, of scholarly products in value-adding services. These APIs should provide a clear record of the provenance of reused information. This provenance is a critical component of the integrity of scholarly information. In effect, reuse and refactoring must extend the notion of citation, which provides a record of provenance among textual scholarly artifacts.

4 Beyond the Web Architecture

In the previous section we demonstrated a number of requirements for scholarly communication that are not fulfilled by the basic web architecture. These include a shared data model with which we can introspect on a resource and its available representations, a means of representing the lineage among new resources, and common APIs for obtaining resources and representations and depositing newly constructed compound documents.

Nevertheless, the architecture of the world wide web [9] provides the foundation for our work. We, therefore, must examine its capabilities and, therefore, assess what must be built on this foundation. This, albeit, brief examination will cover the notions of identifiers (URIs), resources, and representations, will note how these are used in HTTP for access and deposit of information.

Figure 1, taken from [9] illustrates these basic web building blocks. As shown:

- A *resource* is an entity,
- with an *identifier* encoded as a URI,
- for which one or more *representations* are available,

[23] http://www.google.com/apis/maps/.
[24] http://www.amazon.com/gp/browse.html?node=3435361.
[25] http://www.slideshare.net/.
[26] http://ukwebfocus.wordpress.com/2006/11/09/slidesharenet-a-repository-for-slides/.

- which are transmitted as the payload in response to an access request (in the typical case an HTTP GET).

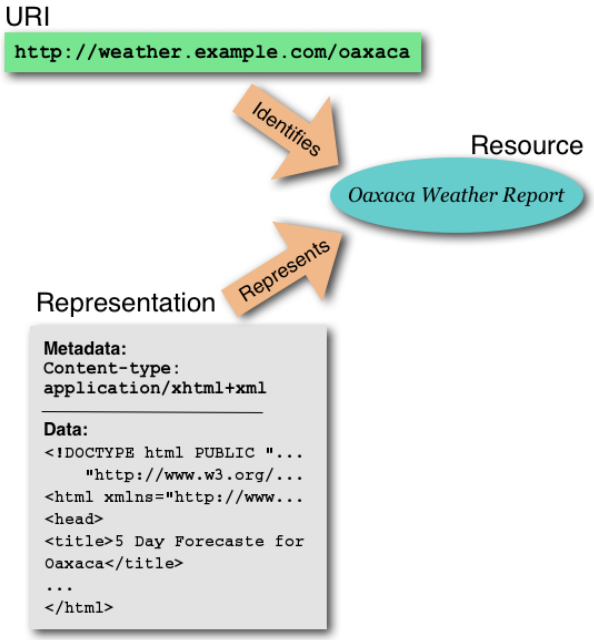

Figure 1 - Identifies, resources, and representations

As described in the previous section, this distinction between an identified abstract resource and its mapping to one or more concrete representations lies at the core of our work and deserves further examination. Figure 2 illustrates this relationship as a graph with typed relationships between the notions of identifiers, resources and representations.

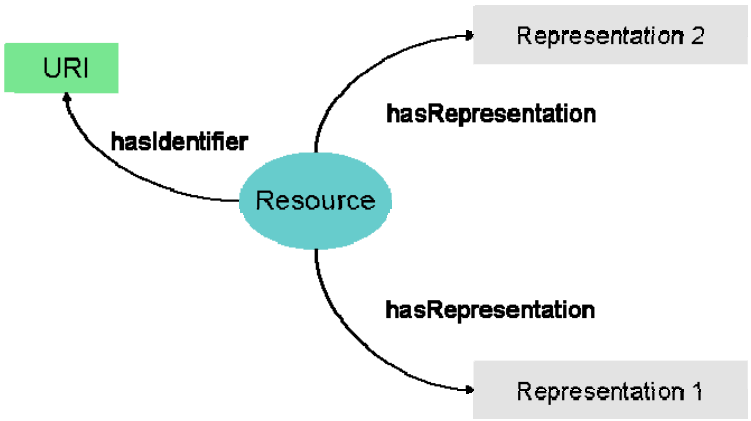

Figure 2 - Resources and representations

This abstraction is indeed a useful basis for building scholarly information environments with compound and complex resources. Unfortunately, HTTP implementation of resources and representations on the Web is incomplete, making it difficult to build more than trivial implementations of the multiple representation notions upon it. In particular, we note the following:

- Multiple Representations of a single Resource do not share a HTTP URI on the Web
- This is the result of:
    - HTTP URIs that play the dual role as abstract identifiers of resources and concrete resolution mechanisms – i.e., they are used by the Domain Name System to locate the network node that will process the HTTP request.
    - Weak and/or not implemented HTTP Content Negotiation capabilities, the result of which is that an HTTP URI resolves to a very limited set of representations.

A glance at *content negotiation* in both the abstract web architecture and its HTTP implementation indicates the intent of this functionality and its weaknesses. The W3C web architecture document [9] defines content negotiation as: "The practice of providing multiple representations available via the same URI. Which representation is served depends on negotiation between the requesting agent and the agent serving the representations." This corresponds to the identifier/resource/representation abstractions described above, and translates that set of abstractions to a definition for agent/protocol/server interaction through negotiation.

RFC 2616 [8] defines how this is done in HTTP. Summarized briefly, the specification defines two means for content negotiation:

- *Server-driven* in which the user agent (browser) includes a set of representation preference guidelines in the GET request, which the server than uses to provide the "best fit" representation. These guidelines fit into three classes: language, character set, and media (MIME) type.
- *Agent-driven* in which a request from the browser triggers a 300 (Multiple Choices) status code from the server with a list of alternative representations in the header field, and the browser, possibly with assistance from a human user, chooses the desired representation.

Server-driven negotiation is widely deployed on the web, especially with the spread of web browsers on handheld devices such as smart phones. For example, a browser request to amazon.com from a desktop browser and a phone-based browser produces two quite different pages. However, the capabilities of this form of negotiation are limited by the capabilities of the preference parameters (language, encoding, media type) in HTTP GET and the ability of the server to determine best representation in a very general sense. Clearly, this form of content negotiation is not sufficiently powerful for our application of interest, advanced scholarly documents in which the dimensions of representation variance include segmentation, service-oriented transformation, and ontology-based description.

Agent-driven negotiation (and its variant transparent negotiation which involves cooperation between a cache and a server) offers greater power. Unfortunately the mechanisms for this form of negotiation – essentially the 300 status code and "Alternates" header - seem to be undefined and therefore rarely if ever deployed. A

Google search on "HTTP alternates" uncovers an obscure HTTP Working Group Internet-Draft from 1998[27] that suggests a format for the alternates header field. This suggestion seems to have produced no further action. Indeed, in a W3C blog entry from October 2006[28] Olivier Thereaux notes that "Content Negotiation, a powerful yet fragile feature of the Web, seems well in need of a good sidekick." Alexey Feldgender responds: "Noone (sic) would dare return 300 or 406 nowadays. Even with the Alternates header, there is risk that a client's user agent doesn't understand that."

Perhaps the lack of further work in this area indicates some lack of relevance for general web content, although we doubt this. In fact an examination of some common web sites reveals a pattern where representations have multiple HTTP URIs, as illustrated in Figure 3, rather than following the one URI/multiple representation pattern suggested by the web architecture.

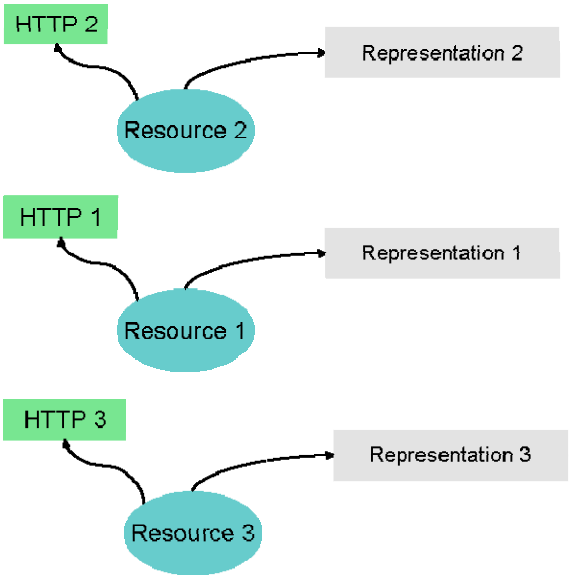

Figure 3 - Multiple representations/Multiple URIs

A look at the popular web site flickr[29] illustrates this. Each picture stored on flickr, in this example the Danpa Panda[30], is in fact a resource with three representations:

1. A slash page that shows the picture in the context of comments, related pictures, metadata, etc.
2. A thumbnail inserted displayed on a page showing a particular user's collection of photos
3. A high quality version of the image available for download and reuse.

As shown in Figure 4, these three representations are not joined by a common URI. Admittedly, flickr could implement a more rational URI syntax, whereby the URI each representation built from a common prefix. But this would still be local to flickr and

[27] http://tools.ietf.org/html/draft-ietf-http-alternates-01
[28] http://www.w3.org/QA/2006/10/missing_http_feature.html.
[29] http://www.flickr.com/.
[30] http://www.keoshi.com/weblog/more-of-danpa

would not provide the functionality that we desire – that is, a common URI on which a user or agent could introspect on available representations and thereby access them in a uniform manner.

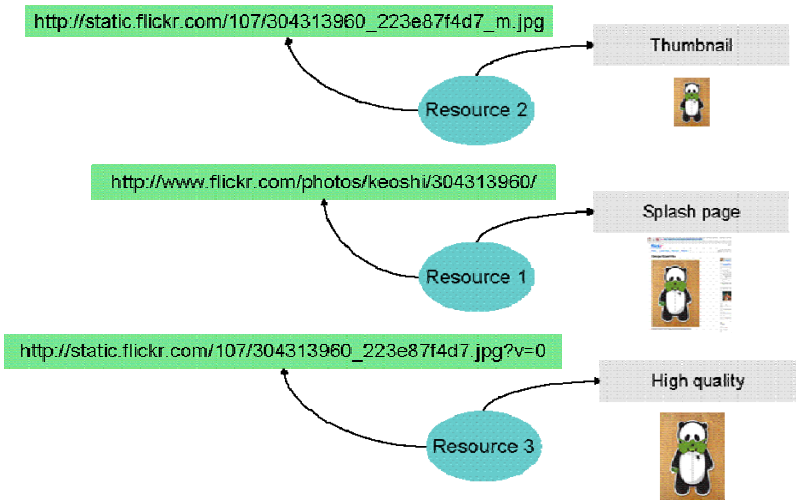

Figure 4 - Flickr "resource" and its "representations"

This pattern demonstrated in flickr is duplicated in most institutional repositories and eprint repositories, which at this point store and disseminate multiple views of relatively simple documents. This is shown in Figure 5 in which an eprint from the astro-ph section of arXiv has a representation as a slash page with abstract and metadata, a pdf version of the full document, and a page providing a user interface to choose from other available document formats, each of which has its own URI. We do see a more rational URI syntax here, but it is unique to the arXiv and no one URI provides a machine-actionable list of all representations of the resource. We note also that the MIME-type driven decisions in server-driven content negotiation would not work here. While some of the representations of this resource are distinguished by MIME type, others share the same type. Clearly, a more descriptive characterization of each representation is needed for any meaningful discrimination among the representations.

As we move from the relatively simple scholarly documents, shown in Figure 5, to compound documents, the shortcomings of existing, and partially implemented, content negotiation on the web become increasingly burdensome. Figure 6 illustrates a resource with identifier URI 4 that is an aggregation of representations from three pre-existing network-available resources – a paper from arXiv, an image from flickr, and a votable[31] formatted dataset. We envision other situations in which a resource may aggregate other resources rather than selected representations from them.

[31] http://www.us-vo.org/VOTable/.

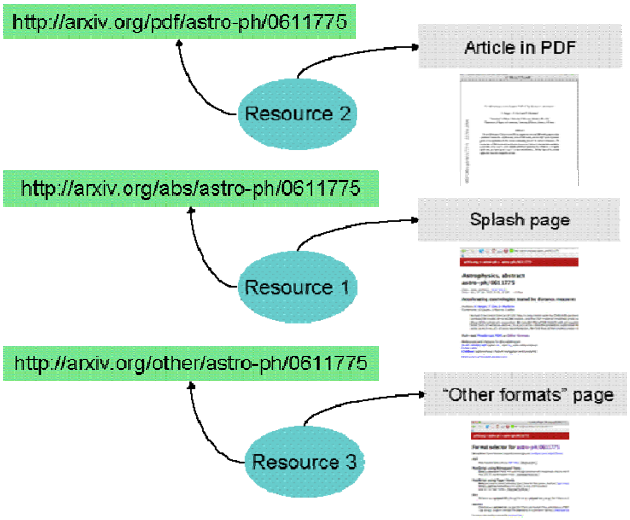

Figure 5 - Arxiv document and representations

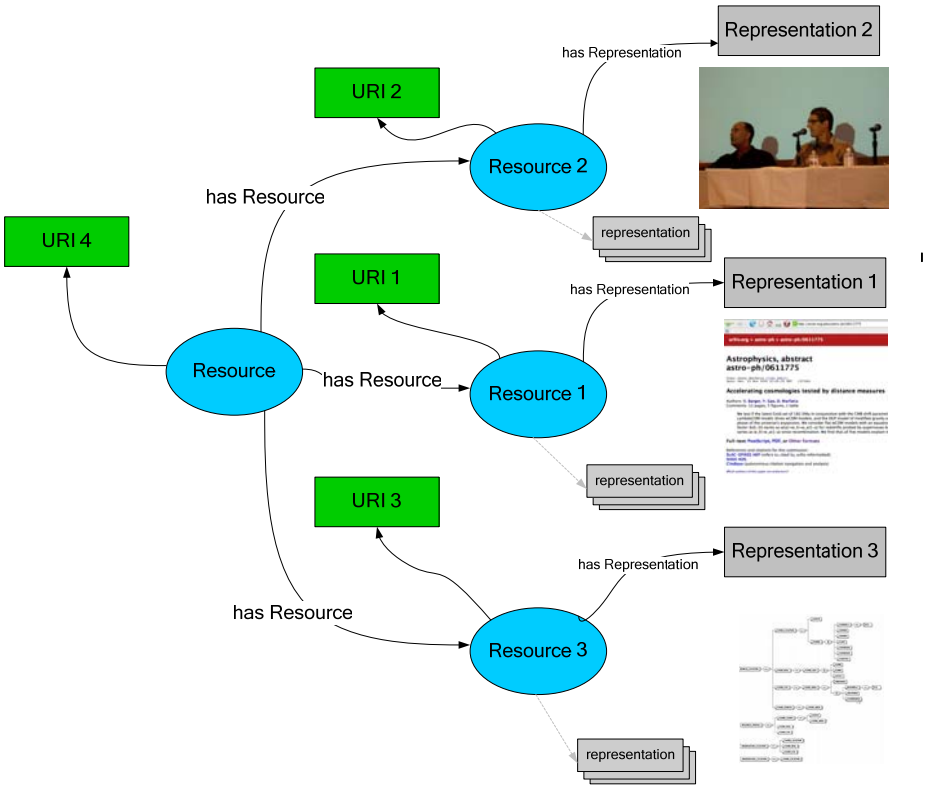

Figure 6 – Aggregate compound object

This figure introduces a number of new complications. First the representations of the resource identified as URI 4 are themselves representations of other resources with their

own URIs. Thus, we need to add to our model a new relation "has Resource" which co-exists as a recursive relation with the recursion terminated by a "has Representation" relation. Note that this new relation and its distinction from "has Representation" is not included in the web architecture. Furthermore, URI 4 only includes specific representations of the aggregated resources, which individually have other representations, as shown in the figure. Thus we need in our model to specify such selective inclusion in the resource/representation structure.

5 Thoughts towards a solution

Informed by the examination of the web architecture in Section 4 we can now look back on the motivating problems in Section 3 and suggest some high-level approaches toward a solution. At this point these are suggestions as to how the OAI-ORE work might proceed, and the actual solutions will result from joint work within the OAI-ORE technical committee. Nevertheless, we provide the following high-level guidelines for the work ahead.

Multiple Representation Issue – This transcends all the issues we have discussed thus far. The infrastructure we develop in ORE must permit:

- Listing of all available representations of a resource
- Qualifying the nature of each representation (beyond MIME type)
- Recursive nesting of resources thereby allowing transitive association of other resources' representations with a compound resource.

Harvesting Representations – Enabling repositories to successfully expose all or selected representations of resources to search engines (and other agents), as described in Section 3.1, will require some form of harvesting functionality. This is in one sense similar to the harvesting functionality in OAI-PMH, but resource-centric rather than metadata-centric. Furthermore, we argue that the harvesting functionality should have two additional features:

- It should not be focused on actual transfer of representations, but should reference representations (and provide sufficient information about those representations) so that harvesting agent (e.g., the search engine robot) can eventually access the appropriate representation itself. We suggest this since mandating transfer would require that participants in the infrastructure willingly exchange intellectual property, unnecessarily burdening the ORE task with IP issues.
- It should represent the relationship of a selected representation to other representations of the containing resources, making it possible for the harvesting agent to access those co-existing representations.

Linking among or referencing resources – Representation of citation links, described in section 3.2, and more advanced form of resource lineage, requires some mechanism for unambiguously identifying resources and expressing relationships among resources. These linkages should be more expressive than existing citation links, which really only express bibliographic (or work-centered) relationships. In our projected world of compound objects and dynamically constructed scholarly

resources, these links should be workflow and evidentiary based (e.g., they should include notions of time or version semantics, network location, and should distinguish between citation of the entire resource or selected representations of a cited resource).

Obtaining Representations – In section 3.4, we described the notion of reuse of resources, and representations thereof, in the construction of new scholarly resources and the citation of existing scholarly resources. We also motivate that these building blocks for reuse might be located across distributed repositories. Such functionality will require a common interface across repositories for obtaining a manifest of representations fo a resources and obtaining a selected representation.

Putting Resources – As noted in Section 3.4, we'd like to provide mechanisms with which scholars can use services external to institutional repositories such as slideshare. The infrastructure should therefore include a common service interface for deposit, or put, of resources.

Integration with the web architecture – Last, but foremost, whatever we do must to be embedded in the Web; we are not creating a parallel universe. Wherever possible and appropriate we should repurpose existing technologies, possibly with qualifications/extensions/modifications if required. Applications that are tightly integrated into the web architecture, such as robot-based search engines, should be able to easily transition between the standard HTTP-based web world and the enhanced infrastructure we suggest here. For example, the current Google crawler accesses individual representations. The mechanisms we devise should make it possible and easy for that crawler to transition from that fetched representation to the parent resource and the other available representations. Such a transparent transition would greatly enhance the heuristic deduping (finding alternate representations of a scholarly resource) already done by Google Scholar[32]. Finally, like the web architecture, and OAI-PMH, we need to keep our solutions simple; as simple as possible.

5.1 CaRF: Shared model for resources and representations

As described above, the notion of resources, representations, and content negotiation permeates the ORE problem space. We suggest then that a fundamental component of the ORE work should be the development of a shared model to represent the relationship between resources and their component representations. As described earlier, this model should be recursive, allowing resources to aggregate other resources and their respective representations. We identify this model as the *Canonical Representation Format*, or CaRF.

Repositories could then provide access, via an obtain (or alternatively Harvest) interface, to a distinguished representation for each resource, known as the *Canonical Representation*, or CaR, which conforms to the CaRF generalized model. The CaR is a representation itself that acts as per-resource manifest of all the representations of a

[32] http://scholar.google.com/.

resource, including metadata representations. In addition, the CaR should qualify each representation with a content descriptor – in the minimal case this could be a MIME type but we anticipate more expressive typing schemes (e.g., ontology-based URIs). Accessing applications and services (e.g., search robots) could then first obtain the CaR, introspect on the contents of the CaR, and determine which if any other representations of the resource to obtain. This use of the CaR is illustrated in Figure 7.

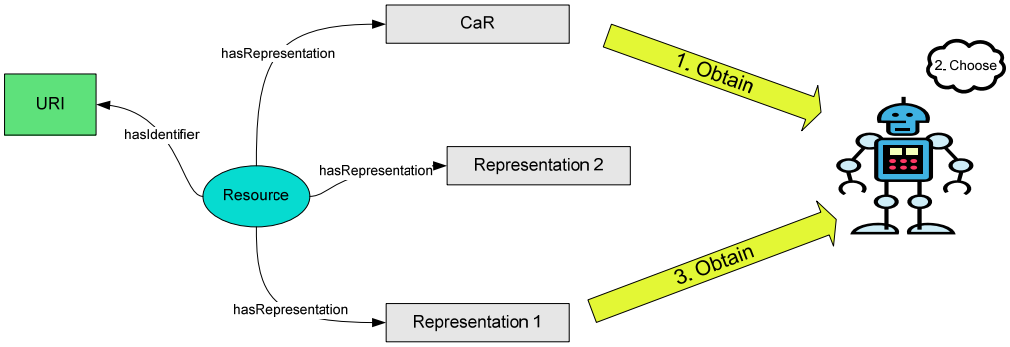

Figure 7 - Resource introspection via the CaR

Unlike bibliographic metadata, which is typically constructed through human effort, we imagine that at least the basic structure of a CaR could be machine generated. This will be increasingly possible with standardization of structured document formats such as the OASIS Open Document Format for Office Applications[33] and the Microsoft Office Open XML Format[34]. Furthermore, we have experimented with machine-learning based methods for recognizing online compound documents and describing their structure [7, 6].

We note that the concepts incorporated in the CaR/CaRF are at least partially implemented in a number of other projects and standards. We anticipate, therefore, that we will not have to "reinvent the wheel", and will be able to learn from previous experience and supplement it. A brief list, and not exhaustive, of similar technologies is as follows:

- ATOM [11] evolves the notion of web syndication, implemented originally in RSS. ATOM is distinguished from its predecessors by its facilities for extensibility and its general structure may be usable for CaR type resource manifests.
- DIDL (Document Item Declaration Language) [2] has been proposed as a standard for the structured description of multimedia resources.
- METS (Metadata Encoding and Transmission Standard) [1] is another structured description standard that, as noted in its name, focuses mainly on packaging metadata formats that describe a resource.
- Pathways Core [14] is the result of a NSF-funded research project at Los Alamos and Cornell. The project has experimented with common data model in the manner of the CaRF.

[33] http://www.oasis-open.org/committees/download.php/12572/OpenDocument-v1.0-os.pdf.
[34] http://office.microsoft.com/en-us/products/HA102058151033.aspx.

- The Digital Library Federation Asset Actions Experiment [4] has demonstrated the value of having a standard XML definition for the actionable URIs for a digital resource.

5.2 A Suite of Common Service Interfaces

In various parts of this document we have suggested a set of common service interfaces that are candidates for the ORE infrastructure. Each of these interfaces provide the framework for transactions based on CaRs; i.e., they initiate the transfer of common representations among repositories and services. Briefly summarizing, these service interfaces are:

- *Harvest* – batch retrieve of multiple CaRs.
- *Obtain* – retrieve an individual CaR.
- *Put* – Request deposit of an individual CaR.

Similar to the discussion in the previous section, these functions have been explored in other projects and standards, which should be carefully examined in the course of ORE work. A brief summary of these associated technologies is as follows:

- *Harvest:* Google SiteMaps[35], RSS, and OAI-PMH[36]
- *Obtain:* OpenURL[37] and unAPI[38]
- *Put:* ATOM publishing protocol[39], HTTP PUT, and WebDAV[40]

5.3 A Common Framework with Multiple Implementations

In conclusion, we propose an infrastructure based on an abstract data model, the CaRF, and three service abstractions – Harvest, Obtain, and Put. Each of these abstractions defines a set of functionalities that can be implemented via a variety of mechanisms, most which already exist in the web context. One interesting question is whether it is indeed more appropriate to support multiple mechanisms or mandate just one. We may conclude that at certain layers (e.g., the CaRF) it is more appropriate to mandate only one expression, while allowing for multiple mechanisms at other layers. This model is illustrated in Figure 8.

[35] http://www.google.com/support/webmasters/bin/topic.py?topic=8476.
[36] http://www.openarchives.org/OAI/openarchivesprotocol.html.
[37] http://www.niso.org/committees/committee_ax.html.
[38] http://unapi.info/.
[39] http://www.ietf.org/html.charters/atompub-charter.html.
[40] http://webdav.org/.

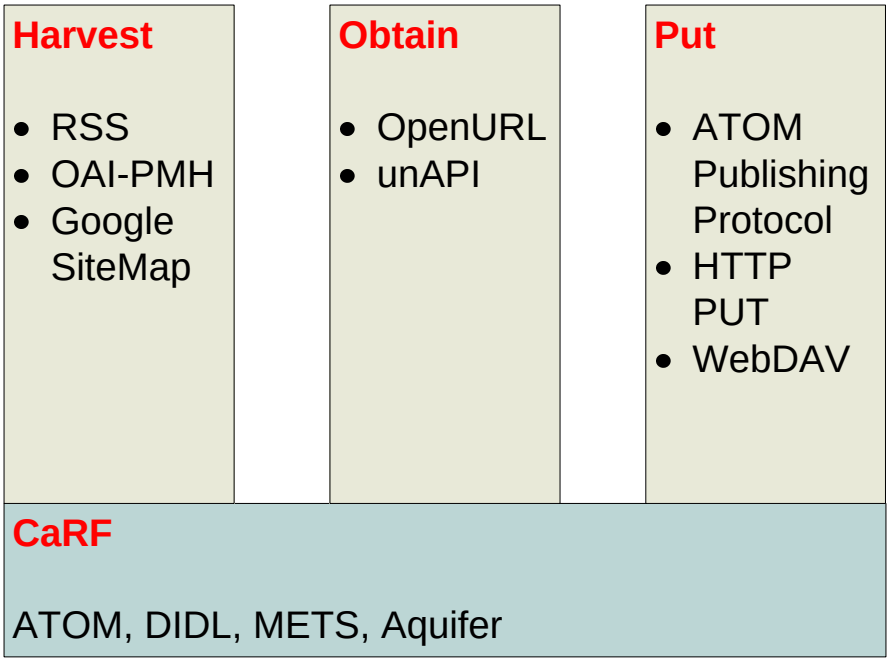

Figure 8 - Shared abstractions - multiple implementations

Some of the mechanisms will be more powerful than others. The end result then will be a spectrum of participation in the ORE infrastructure. Some participants will choose easy-to-implement minimalist mechanisms that offer relatively limited functionality. Others will choose more difficult-to-implement, but more functional mechanisms. However, because all are based on the same general abstractions, a degree of interoperability among all participants will be maintained.

The tasks then for the next two years are to define these abstractions and their expression through one or more mechanisms. With experimentation and prototype deployment we hope to prove the effectiveness of this model.

References

[1] *METS, Metadata Encoding and Transmission Standard*, Network Development and MARC Standards Office, the Library of Congress, 2002.

[2] J. Bekaert, P. Hochstenbach and H. Van de Sompel, *Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library*, D-Lib Magazine, 9 (2003).

[3] S. Brin, L. Page, *The anatomy of a large-Scale hypertextual Web search engine*, *Proc. 7th International World Wide Web Conference*, 1998.

[4] R. Chavez, T. W. Cole, J. Dunn, M. Foulonneau, T. G. Habing and W. Parod, *DLF-Aquifer Asset Actions Experiment: Demonstrating Value of Actionable URLs*, D-Lib Magazine, 12 (2006).

[5] T. W. Cole and S. Shreeves, *Lessons Learned from the Illinois OAI Metadata Harvesting Project*, in D. Hillmann, ed., *Metadata in Practice*, American Library Association, Chicago, 2004.

[6] P. Dmitriev and C. Lagoze, *Automatically Constructing Descriptive Site Maps*, *Eighth Asia Pacific Web Conference*, Harbin, China, 2006.

[7] P. Dmitriev, C. Lagoze and B. Suchkov, *As We May Perceive: Inferring Logical Documents from Hypertext*, *HT 2005 - Sixteenth ACM Conference on Hypertext and Hypermedia*, Salzburg, Austria, 2005.

[8] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach and T. Berners-Lee, *Hypertext Transfer Protocol -- HTTP/1.1*, The Internet Society, 1999.

[9] I. Jacobs and N. Walsh, *Architecture of the World Wide Web*, W3C, 2004.

[10] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, Journal of the ACM, 46 (1999), pp. 604-632.

[11] M. Nottingham and R. Sayre, *The Atom Syndication Format*, Network Working Group, Internet Engineering Task Force, 2005.

[12] H. Van de Sompel and C. Lagoze, *Notes from the Interoperability Front: A Progress Report on the Open Archives Initiative*, *6th European Conference on Research and Advanced Technology for Digital Libraries*, Springer-Verlag, Rome, Italy, 2002.

[13] H. Van de Sompel, S. Payette, J. Erickson, C. Lagoze and S. Warner, *Rethinking Scholarly Communication: Building the System that Scholars Deserve*, D-Lib Magazine (2004).

[14] S. Warner, J. Bekaert, C. Lagoze, X. Liu, S. Payette and H. Van de Sompel, *Pathways: Augmenting interoperability across scholarly repositories*, International Journal on Digital Libraries special issue on Digital Libraries and eScience, forthcoming (2007).