Open Archives Initiative – Object Reuse and Exchange
Report on the Technical Committee Meeting, May 29, 30 2007

Edited by: Carl Lagoze & Herbert Van de Sompel

1 Venue

Google Inc., New York, NY

2 Final Agenda

Tuesday, May 29

Time	What	Details	Who
9:00-10:00	Welcome	Introductions, Motivations, & Meeting Goals	Carl, Herbert
10:00 -11:00	Review of Named Graph White Paper	Summary of technical approach, Summary of open issues	Carl, Herbert
11:00 – 11:15	Break		
11:15-13:00	Round-table discussion regarding the white paper	Concepts - the Named Graph. Enumeration of Open Issues	All
13:00-14:00	Lunch	Provided	
14:00-15:30	Reports from Technical Exploration Groups	Findings on explorations regarding named graphs, serializations, discovery.	TC members
15:30-15:45	Break		
15:45-17:00	Reports from Technical Exploration Groups, continued. Round-table discussion.	Findings on explorations regarding named graphs, serializations, discovery. Discussion of findings and issues.	TC members, All

Wenesday, May 30

Time	What	Details	Who
9:00-11:00	Presentation of serialization excercises and discussions thereof.	Some participants were asked to present approaches to serialization for 3 types of compound objects.	TC members
11:00-12:30	Reaching Consensus (1)	Working towards core agreements regarding named	All

		graph, serialization requirements, referencing.	
12:30-13:00	Lunch	Provided	
13:00-15:00	Reaching Consensus (2)	Working towards core agreements regarding named graph, serialization requirements, referencing.	All
15:00 – 15:15	Break		
15:15-15:45	Recap of discovery	Exploration of discovery in light of core agreements that were reached.	All
15:45 – 16:00	Wrap-up	What has been accomplished? Reactions.	Cliff, Herbert

3 Attendees

ORE coordinators:

- Carl Lagoze, Cornell University
- Herbert Van de Sompel, Los Alamos National Laboratory

From the ORE Technical Committee:

- Tim DiLauro, Johns Hopkins University
- Leigh Dodds, Ingenta
- David Fulker, UCAR
- Tony Hammond, Nature Publishing Group
- Pete Johnston, Eduserv Foundation
- Richard Jones, Imperial College
- Peter Murray, OhioLINK
- Michael Nelson, Old Dominion University
- Rob Sanderson, University of Liverpool
- Simeon Warner, Cornell University
- Jeff Young, OCLC

From the ORE Liaison Group:

- Tim Cole, DLF Aquifer and UIUC Library
- Julie Allinson, UKOLN and JISC
- Jane Hunter, DEST and University of Queensland
- Savas Parastatidis, Microsoft
- Thomas Place, DARE and University of Tilburg
- Robert Tansley, Google, Inc. and DSpace

From the ORE Advisory Committee:

- Jane Hunter, DEST and University of Queensland
- Clifford Lynch, Coalition of Networked Information

4 Identifying Core Questions

4.1 Initial feedback to the Compound Information Objects white paper

The white paper that was prepared to frame discussions for this meeting is at:

- Carl Lagoze, Herbert Van de Sompel. Compound Information Objects: The OAI-ORE Perspective. http://www.openarchives.org/ore/documents/CompoundObjects-200705.html

The following considerations were shared during the round-table discussion regarding the white paper:

- Issues related to author identification should be considered out of scope for the initial bootstrap work, but should be flagged for later, or allied, efforts.
- Issues related to trust and authenticity should be considered out of scope for the initial bootstrap work, but should be flagged for later, or allied, efforts.
- The white paper leaves out issues regarding *transactions* pertaining to compound objects, e.g. the notions of Harvest, Obtain, Register that were described in previous versions of the white paper are missing. With this regard:
    - There is agreement that these remain important.
    - The white paper does cover Harvest to an extent in Section 7.7 "Discovery of Named Graphs" of the white paper. It lists Harvest (Sitemaps, RSS, OAI-PMH) alongside link-based discovery mechanisms.
    - In the proposed approach, the notion of Obtain translates to dereferencing URIs, i.e. the URI of the named graph and the URIs referenced by the named graph.
    - The notion of Register (update a compound object, create a new compound object, and/or the notification to a component resource of its membership in a compound object) is indeed not addressed in the white paper.
- Concern was expressed regarding the relationship between semantic web activities and the ORE effort. This concern is inspired by the fear of alienating potential adopters that remain skeptical regarding the semantic web activities. How explicitly should ORE communications and specifications reference and/or use semantic web technologies?
- Concern was expressed regarding the scope of named graphs and Resource Maps: do these only reference resources that are part of a compound object, or also resources that are not considered part of the compound object, but that have relationships to (components of) the compound object.
- It was suggested that different applications may adhere to different notions of a boundary (within the confines of a named graph), for example, constrained by link types.

- The need for simplicity of a solution was emphasized. Yet, the simple solution must be able to deal with complex scenarios.
- The notion of metadata pertaining to a named graph is important, specifically datetime of assertion (related to versioning) and authorship.
- There is a need to be able to point at a specific resource in a named graph / Resource Map.
- There is a need to express that a resource is part of a specific named graph.
- There is a need to be able to include (sub-)named-graphs as components of named graphs.
    - Note: Although this requirement was formulated as stated here, the underlying issue was really how to re-use a compound object (that has a corresponding named graph), as part of another compound object (that has a corresponding named graph). And the same question pertains to components of a compound object.
- The containment node and its identification remain puzzling issues.

4.2 Core questions

The following fundamental questions were crystallized from the round-table discussions, and were further explored during the meeting:

In the ORE domain:

- Do named graphs only reference resources that are part of the compound object, or also other, related ones?
- Do named graph resources always have a containment node?
- How to name the containment node?
- How to reference, the named graph, the compound object, a component of the compound object?

5 Meeting Results

5.1 Principles regarding the goals of ORE

A first principle regarding the core goal of ORE is proposed: **Enriching the web graph with boundary information**.

This first principle could be enriched with the secondary principle: **Adding meaning to the web graph through relationships and links**.

These principles help clarify the following regarding the scope of ORE activities:

- Use the notion of a simple manifest.
- Restrict ORE-defined relationships to be of a structural (boundary) nature.
- Allow communities to define other relationships and allow these to be used in the manifest.

5.2 Core Agreeements

<u>Introductory Note:</u>

Several comments received in response to a draft version of this report expressed a level of confusion regarding the distinction between named graph, Resource Map, and serialization of Resource Map.

In the white paper, and in the New York meeting discussions, the following perspective was taken (using Web architecture terminolgy):

- *Named graph == resource*
- *Resource Map == serialization of named graph == representation*

In order to present a faithfull report of the meeting, the core agreements listed below are expressed using this perspective. However, feedback was presented at the meeting that indicates a desire to evolve this perspective as follow (using Web architecture terminolgy):

- *Resource map == resource*
- *Serialization of Resource Map == representation*

<u>**In the ORE domain**, we agree upon the following:</u>

1. URIs exist that, **by convention**, are interpreted to identify compound information objects. This is, for example, the case of the HTTP URI of a DSpace splash page. According to the web architecture, however, such a URI identifies a resource that is the splash page, not the resource that is the compound information object for which the splash page is a convenient human entry point.

2. The existence of a compound object in the ORE domain is dependent upon the existence of a named graph resource, i.e. there exists the notion of a named graph that corresponds with a compound object.
    a. A named graph resource **must** be identified by means of a protocol-based URI which can be resolved to a Resource Map.
    b. A named graph resource **must** have only Resource Maps as representations.
    c. Multiple Resource Maps **may** be available from the URI that identifies a specific named graph resource. If this is the case, all such Resource Map representations **must** be equivalent serializations in different formats of the same named graph resource.

3. Regarding the expressiveness of the Resource Map:
    a. The most primitive Resource Map **must** simply express the full set of resources that are considered part of a compound object.
    b. In addition to the expressiveness of (3a), Resource Maps **may** also:
        i. Express resources that are not part of a compound object (according to the author of the named graph), but that are referenced by the named graph

that corresponds with the compound object. If it does, it must distinguish between those resources that are part of the compound object and those that are not.

ii. Express the relationships between the resources referenced by the named graph.
iii. Express the types of the relationships between the resources referenced by the named graph, i.e. label the arcs.
iv. Express other information related to the named graph and to the resources that it references such as metadata, etc.

4. There **must** be ways to determine whether a representation available from a protocol-based URI is a Resource Map. We recognize the following possibilities:
   a. MIME type indicating a representation is a Resource Map as available from e.g. HTTP header.
   b. Explicit statement in a representation indicating that the representation is a Resource Map.

5. We recognize 3 types of reference:
   a. Reference using the URI of a named graph – this is a reference to the named graph that corresponds to a compound object.
   b. Reference using the URI of a non-named graph resource, including a reference to a resource that is part of a compound object – this is a reference to "only" that resource.
   c. Reference using a URI double consisting of the URI of a resource and the URI of a named graph, i.e. ($URI_{resource}$, $URI_{namedgraph}$) – this references the resource in the context of the named graph, i.e. it references the resource in the manner that it is referenced by the named graph. This approach can uniformely be used for both the containment node and any other resource of a compound object. For example:
      i. *a frog*

5.3 Further discussions

Regarding the containment node:

No clear conclusion was reached regarding the issue of the containment node. There is the notion that somehow the named graph (serialization) should reflect the "is-ness" of the compound object that the named graph corresponds with. To put it differently, somehow the "aggregation" aspect, the 'compoundness' as it were, that binds the components of a compound object must be represented. This is, in essence, the motivation for considering the concept of a containment node (see also white paper use cases).

For example, in named graphs that include more than just those resources that are considered to be part of the compound information object, a need was felt to express an internal boundary within the larger named graph boundary that distinguishes the resources hat are part of the compound object from other resources referenced in the graph. In other instances there may be resources that are unconnected to any other resource in the named graph, but which must be referenced in the named graph because the graph's author wants to assert that they are part of the compound information resource.

The serialization excercises have shown that these and other objectives can be achieved in various ways, including the explicit introduction of a URI-identified containment node, the use of a blank node for a ‘bag’ of resources without explicit containment relationships, or the use of an existing resource as the ‘containment node’. Which direction to take may depend on document modeling considerations and serialization approaches.

Regarding the re-use of resources in multiple compound objects

Related to the referencing mechanisms listed under 6.2.5 above, there was discussion over the re-use of individual resources as parts of multiple compound objects, and how this relates to the named graphs that corresponds to these compound objects. To illustrate re-use approahces, let:

- U be the URI of a resource,
- X and Y be URIs of named graphs that reference U (i.e. that are compound objects to which X and Y correspond of which U is a part), and
- Z be the URI of a named graph corresponding to a compound object that wants to re-use U.

The following re-use approaches can be distinguished:

- Re-use utilizing referencing approach 6.2.5.b above: in Z, reference U
- Re-use utilizing referencing approach 6.2.5.c above: in Z, reference U in the context of X, or reference U in the context of Y. These are re-uses of U in the way that it appears in a certain “context”, i.e. the context specified by the named graphs X and Y, respectively.

As part of this discussion, the following was stated:

- Serialization techniques need to support referencing a resource in the context of a named graph.
- It would be helpful to have services that can respond to the question “In the context of which named graphs is this specific resource U referenced?”

Regarding named graph discovery separate from Harvest-based solutions:

Two approaches were presented to support discovery of named graphs using HTTP response headers (see Michael Nelson’s materials):

- Referencing the named graph itself in the HTTP header in response to a HTTP HEAD/GET issued against a component resource of a compound object. This could, for example, be achieved using an ORE specific header (e.g. ‘X-OAI-ORE-Named-Graph’), or using an ORE-specific implementation of the HTTP LINK header (e.g. Link: <URI of named graph>; rel="info:ore/type/named_graph").
- Referencing a **discovery node** in the HTTP header in response to a HTTP HEAD/GET issued against a component resource of a compound object. It was agreed that this approach was useful in case only one resource of the compound object (for example the containment node) would know about the existence of a named graph, while all other

components would only know about their mutual parent node (for example the containment node). This could, for example, be achieved using an ORE specific header (e.g. 'X-OAI-ORE-Discovery-Node'), or using an ORE-specific implementation of the HTTP LINK header (e.g. Link: <URI of discovery node>; rel="info:ore/type/discovery_node").

Regarding ontologies:

The original agenda listed a discussion of:

- Vocabularies for expressing types of links between resources denoted by the nodes in a named graph.
- Vocabularies for expressing properties of resources denoted by nodes in a named graph, especially semantic type, media type, and media format.

Hardly any time was spent on these, but there is consensus that the ORE effort itself should take a very minimal approach towards defining vocabularies, i.e. should restrict itself to vocabularies that are necessary and sufficient to bootstrap implementation and adoption. More expressive vocabularies should be community-defined.

6 Plans and Action Items

- An HTML version of the white paper will be publicly posted. (Done at http://www.openarchives.org/ore/documents/CompoundObjects-200705.html).
- In order to support the launch of projects that want to build on ORE concepts (e.g. a potentially Microsoft-funded eChemistry project), an effort will be launched to release alpha specifications by the end of September 2007. With this regard:
  - An ORE sub-group will work towards sharing a draft alpha specification with the ORE Technical Committee and ORE Liaison Group by early September 2007.
  - Members of the ORE Technical Committee and the ORE Liaison Group should reserve time to review and comment upon this draft alpha spec (and iterations thereof) in the course of September 2007.
- The scope and direction of the ORE effort has shaped up enough to merit a conference call of the ORE Advisory Committee. This will be scheduled post the NYC meeting.
- There will be an ORE related panel at JCDL 2007 involving Clifford Lynch, Neil Jacobs (JISC), Carl Lagoze and Herbert Van de Sompel.