Exploring Object Reuse and Exchange

Abstract

This chapter of "Object Reuse and Exchange (OAI-ORE)" will introduce and explain some basic elements of RDF in order to "bootstrap" our exploration of ORE. The National Digital Newspaper Program at the Library of Congress will be used as a real-world example to complement a sequence of graphs that will illustrate Aggregations and Resource Maps, metadata, Aggregated Resources, and nested Aggregations.

Building Blocks: RDF Triples

The Resource Description Framework (RDF) was originally designed as a standard for encoding metadata but has grown in its scope and application to be used more generally for modeling information.[1] It is especially useful for describing entities in terms of their relationship with other entities. RDF statements are called "triples" and take the form of subject-predicate-object (see figure 2). In other words, something (the *subject*) is described by or related to (*predicate*) something else (an *object*). For example, you could describe a particular document that is a page in a newspaper by making a series of sentence-like statements about it:

1. The document was issued on December 14, 1918.
2. The document is titled "The St. Joseph observer. - 1918-12-14 - 1"
3. The document is ordered first.
4. The document belongs to a particular issue.
5. The document is a newspaper page.

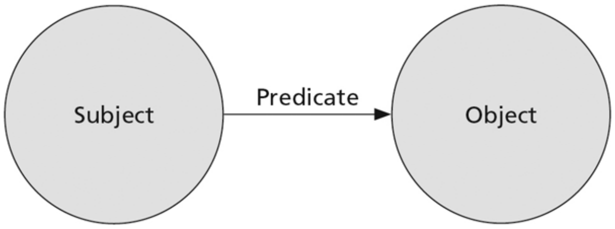


Figure 2
Graphical representation of an RDF triple

Taken together, collections of these statements can provide robust and relational descriptions of resources. In dealing with Web resources using RDF, the subject is denoted by a URI. The predicate is also denoted by a URI. The object can be denoted by a URI or a literal (a string of text).[2] If the newspaper page in our previous example was digitized and available on the Internet, our statements could be:

1. *The document was issued on December 14, 1918.*
   subject: http://chroniclingamerica.loc.gov/lccn/sn90061457/1918-12-14/ed-1/seq-1#page
   predicate: http://purl.org/dc/terms/issued
   object: 1918-12-14

2. *The document is titled "The St. Joseph observer. - 1918-12-14 - 1"*
   subject: http://chroniclingamerica.loc.gov/lccn/sn90061457/1918-12-14/ed-1/seq-1#page
   predicate: http://purl.org/dc/terms/title
   object: "The St. Joseph observer. - 1918-12-14 - 1"

3. *The document is ordered first.*
   subject: http://chroniclingamerica.loc.gov/

lccn/sn90061457/1918-12-14/ed-1/seq-1#page
predicate: http://chroniclingamerica.loc.gov/terms#sequence
object: 1

4. *The document belongs to a particular issue.*
subject: http://chroniclingamerica.loc.gov/lccn/sn90061457/1918-12-14/ed-1/seq-1#page
predicate: http://www.openarchives.org/ore/terms/isAggregatedBy
object: http://chroniclingamerica.loc.gov/lccn/sn90061457/1918-12-14/ed-1#issue

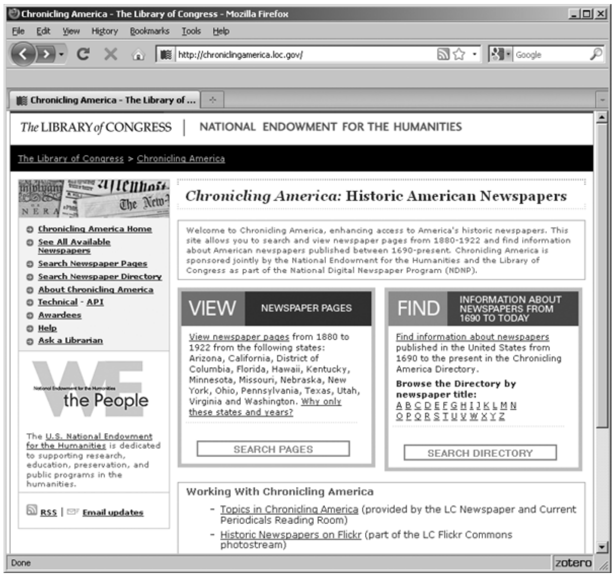
5. *The document is a newspaper page.*
subject: http://chroniclingamerica.loc.gov/lccn/sn90061457/1918-12-14/ed-1/seq-1#page
predicate: `rdf:type`
object: http://chroniclingamerica.loc.gov/terms#Page

Literals can be any plain text, or they may be formatted by referencing a data type, such as YYYY-MM-DD for a date. Furthermore, vocabularies such as those offered in Dublin Core (see "dcterms" in the examples in the next chapter) can be used to establish the meaning of the resource or the relationship between it and another resource in a manner that is formal and can be referenced. If an existing vocabulary does not meet your needs, you can create your own

Continuing our newspaper example, let's take a look at a real-world implementation of ORE with the National Digital Newspaper Program at the Library of Congress. After a summary of the program and an overview of its general workflow, we'll see how Aggregations was defined in the program's data model. We'll graph the relationship between an Aggregation and its Resource Map and then add more information sequentially to illustrate the primary ORE entities through subsequent graphs.

National Digital Newspaper Program

The National Digital Newspaper Program (NDNP) is a partnership between the Library of Congress and the National Endowment for the Humanities (NEH) to digitize and preserve newspapers that were published in the United States between 1836 and 1922.[3] The NEH grants funds to state projects to select and digitize newspapers of historical and cultural significance, which are then aggregated, preserved, and made accessible by the Library of Congress.[4] The first phase of the project is complete, and the NDNP currently contains 214 newspaper titles with approximately 192,000 issues and 1.8 million pages of content.[5] The digitized newspapers are made available through the Chronicling America website (see figure 3), which acts as a portal for searching and browsing the collection.


Chronicling America - The Library of Congress - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
http://chroniclingamerica.loc.gov/
Google
Chronicling America - The Library of ...
The LIBRARY of CONGRESS
NATIONAL ENDOWMENT FOR THE HUMANITIES
The Library of Congress > Chronicling America
Chronicling America Home
See All Available Newspapers
Search Newspaper Pages
Search Newspaper Directory
About Chronicling America
Technical - API
Awardees
Help
Ask a Librarian
National Endowment for the Humanities
WE the People
VIEW
NEWSPAPER PAGES
FIND
INFORMATION ABOUT NEWSPAPERS FROM 1690 TO TODAY
A B C D E F G H I J K L M N
O P Q R S T U V W X Y Z
SEARCH PAGES
SEARCH DIRECTORY
RSS | Email updates
Done
zotero

Chronicling America: Historic American Newspapers

Welcome to Chronicling America, enhancing access to America's historic newspapers. This site allows you to search and view newspaper pages from 1880-1922 and find information about American newspapers published between 1690-present. Chronicling America is sponsored jointly by the National Endowment for the Humanities and the Library of Congress as part of the National Digital Newspaper Program (NDNP).

View newspaper pages from 1880 to 1922 from the following states: Arizona, California, District of Columbia, Florida, Hawaii, Kentucky, Minnesota, Missouri, Nebraska, New York, Ohio, Pennsylvania, Texas, Utah, Virginia and Washington. Why only these states and years?

Find information about newspapers published in the United States from 1690 to the present in the Chronicling America Directory.

Browse the Directory by newspaper title:

The U.S. National Endowment for the Humanities is dedicated to supporting research, education, preservation, and public programs in the humanities.

Working With Chronicling America

- Topics in Chronicling America (provided by the LC Newspaper and Current Periodicals Reading Room)
- Historic Newspapers on Flickr (part of the LC Flickr Commons photostream)

Figure 3
Chronicling America website at the Library of Congress

Chronicling America
http://chroniclingamerica.loc.gov

The Library of Congress has a fairly straightforward workflow and data model. Digitized content and metadata are physically mailed to them from the projects on hard drives, and the group of files on each hard drive is referred to a "batch." There are Aggregations defined for batches, newspaper titles, issues, and pages. For example, a page Aggregation includes a derivative JPEG 2000 image, the raw text output of an optical character recognition (OCR) scan of the page, a structured capture of the OCR encoded in XML, a PDF file of the page, and a thumbnail JPEG image.[6]

Pages are then aggregated by issues, and issues are aggregated by newspapers, which are aggregated by title. So, from our previous example, page 1 would be aggregated by the issue published on December 14, 1918, that is aggregated by the *St. Joseph Observer* (the title of the newspaper). This may seem to suggest a hierarchical data model; however, it is not: issues are also aggregated by batches. In other words, issues from a newspaper may come from more than one batch. When you think about large-scale digitization projects, this makes sense because all of the issues from a newspaper may not be digitized at the same time or by the same facility. Also, large runs of a given newspaper's issues will span multiple drives. While end users may not

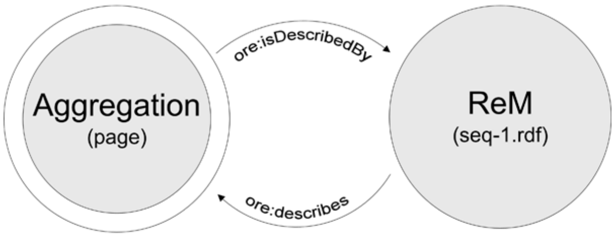


Figure 4
An Aggregation and Resource Map

care which batch an issue came from, this information is potentially important for the library's maintenance of the technical provenance of the objects.

RDF graphs, which are based on triples, can be used to explore ORE entities in an intuitive way. We will begin exploring ORE entities by graphing the relationship between an Aggregation and the Resource Map that describes it. Then we will add some metadata and Aggregated Resources. We'll graph a nested Aggregation and conclude by putting all of these pieces together in a single graph. The examples are tied to the way that the NDNP mapped ORE to its data model, with Aggregations defined for pages, issues, newspaper titles, and batches. It may be helpful to flip ahead to the next chapter on Resource Maps to see these examples continued through to their serializations in RDF/XML.

Aggregations and Resource Maps

Generally speaking, an Aggregation can be considered to be a collection of Web resources, and a Resource Map is like a shopping list that describes what is inside an Aggregation. To be more specific, an Aggregation is an RDF Resource of type `ore:Aggregation` that is a collection of other Resources. An Aggregation must be associated with at least one Resource Map. An Aggregation is identified by a URI (such as a URL), and this URI should not be assigned to any other Resource or be used for any other purpose than referencing the Aggregation. For example, if someone used an Aggregation as a whole and wanted to cite it, they should use the URI of the Aggregation and not any other URI, such as the link to a PDF that is contained in the Aggregation or the "splash page" that represents the Aggregation on a website. Because an Aggregation is a conceptual construct, it cannot be downloaded. Instead, the webserver uses the Aggregation's URI to provide access to a Resource Map, which is tangible and references the Aggregation and provides more information about it and the Resource Map itself.[7] In the next chapter we'll investigate how Resource Maps are discovered and served to client applications.

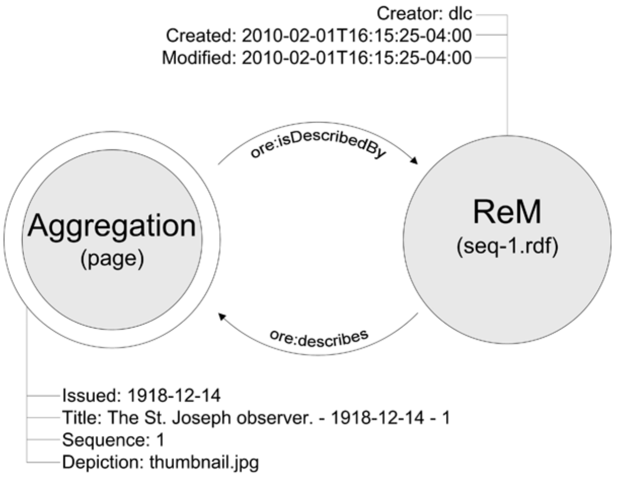


Figure 5
Adding metadata

In figure 4, we see an Aggregation that is a page of a newspaper from the NDNP and its Resource Map, which has been serialized into an RDF/XML file (seq-1.rdf) that can be downloaded from a webserver. ORE requires that a Resource Map express its relationship to the Aggregation (`ore:describes`), and the subject of this triple must be the URI of the Resource Map. The pattern of these triples (e.g., "The Resource Map describes the Aggregation") will begin to sound familiar as you read these graphs and they are repeated over and over again.

Adding Metadata

A Resource Map is required to express two basic metadata elements about itself. Minimally, the Resource Map must include who created the Resource Map by using `dcterms:creator` in a triple whose object is a Resource referenced in `dcterms:agent`.[8] This can then be the subject of additional optional triples that express descriptive text about the creator or the creator's e-mail address, for example, using the *foaf* (Friend Of A Friend) ontology.[9] Secondly, it must express and maintain the date-timestamp (using `dcterms:modified`) to reflect the last time the Resource Map was changed.

Besides these two required elements, a Resource Map can express additional metadata about either itself or about the Aggregation. For example, figure 5 shows the date the Resource Map was created. Some descriptive metadata about the Aggregation provides the date the newspaper page was issued, its title, its sequence in the newspaper (it is the first page), and a thumbnail image. Other common metadata include rights information (`dcterms:rights`) and the creator of the Aggregation, who may be different from the creator of the Resource Map.[10]

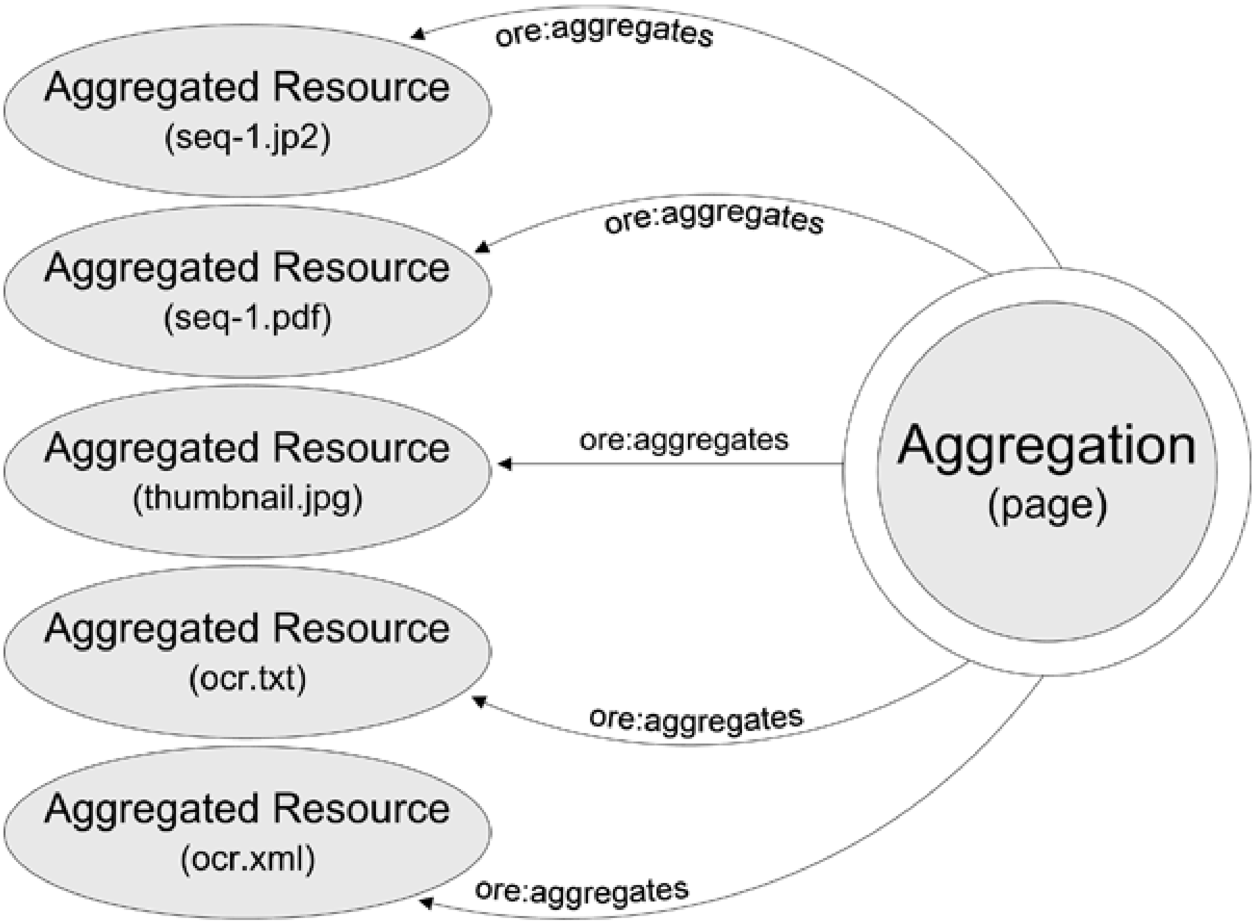


Figure 6
Aggregated Resources

Aggregated Resources

A Resource Map can include one or more triples with a predicate of `ore:aggregates` to denote the Aggregated Resources that make up the Aggregation. Each Aggregated Resource must have its own URI, and the Resource Map must use this URI to reference the Aggregated Resource. A Resource Map may also include triples with the `ore:isAggregatedBy` predicate to assert that one or more Aggregated Resources belong to other Aggregations. A graph that displays an Aggregation with one or more `ore:aggregates` relationships to Aggregated Resources is known collectively as the *Aggregation Graph*. All authoritative Resource Maps are required to assert the same Aggregation Graph.[11]

In the example of the National Digital Newspaper Program, a "page" Aggregation has at least five Aggregated Resources (see figure 6):

- a JPEG 2000 image derived from scanning the newspaper page (seq-1.jp2)
- an Adobe Acrobat file of the page (seq-1.pdf)
- a small thumbnail image of the page (thumbnail.jpg)
- the raw text output from performing OCR on the page (ocr.txt)
- structured XML resulting from the OCR scan (ocr.xml)

Each of these files represents the page in a different way for different uses. The JPEG 2000 and PDF files are mainly intended for people to use as digital surrogates in place of the original newspaper page. The thumbnail image is small and can be downloaded quickly to enable browsing through newspaper pages. Finally, the OCR files are useful for indexing and searching, among other things. Regardless of their purposes, they can be thought of together as a "page" and as a compound digital object. By defining an Aggregation for a page and these files as Aggregated Resources, their semantics can be maintained and leveraged by ORE and other Semantic Web applications.

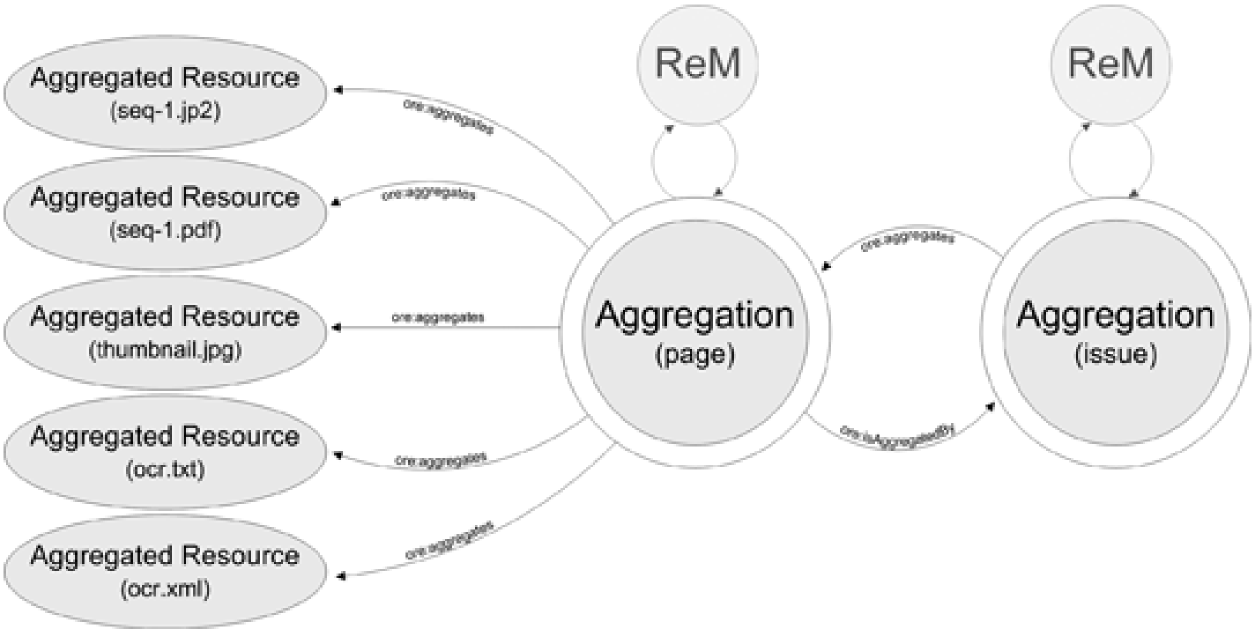

ReM
ReM
Aggregated Resource (seq-1.jp2)
Aggregated Resource (seq-1.pdf)
Aggregated Resource (thumbnail.jpg)
Aggregated Resource (ocr.txt)
Aggregated Resource (ocr.xml)
ore:aggregates
ore:aggregates
ore:aggregates
ore:aggregates
ore:aggregates
Aggregation (page)
Aggregation (issue)
ore:aggregates
ore:isAggregatedBy

Figure 7
Nested Aggregations

Nested Aggregations

An Aggregation can contain other Aggregations. The result is a recursive nesting of Aggregations. When this nesting occurs, the Aggregation being nested can be thought of and treated as an Aggregated Resource. This nesting must be expressed in multiple Resource Maps because a Resource Map is limited (by definition) to describing only one Aggregation. A Resource Map can (but is not required to) assert that the Aggregation that is being nested as an Aggregated Resource is described by another Resource Map using the `ore:isDescribedBy` predicate. This informs clients of the first Resource Map that a nested Aggregation is described by its own Resource Map and points to it. Otherwise, all of the same semantics (e.g., `ore:aggregates, ore:isAggregatedBy`) apply to nested Aggregations.[13]

Just as a newspaper page is described in the NDNP as an Aggregation of various files, pages are aggregated into issues: Aggregations of Aggregations (see figure 7). This nesting becomes more complex as issues are aggregated both by newspaper titles and also by batches. The Library of Congress makes it easy to browse and understand these different Aggregations by instantiating them on the Chronicling America website, for example:

- Batch: http://chroniclingamerica.loc.gov/batches/batch_mohi_carver_ver01#
- Title: http://chroniclingamerica.loc.gov/lccn/sn90061457#
- Issue: http://chroniclingamerica.loc.gov/lccn/sn90061457/1918-12-14/ed-1#
- Page: http://chroniclingamerica.loc.gov/lccn/sn90061457/1918-12-14/ed-1/seq-1#

Authoritative and Non-Authoritative Resource Maps

While a Resource Map is limited to describing only one Aggregation, an Aggregation may include more than one Resource Map. In other words, a Resource Map can only have one `ore:describes` assertion and cannot describe more than one Aggregation. For example, a nested Aggregation may include a Resource Map describing the Aggregation along with Resource Maps describing the Aggregations that are included in it as Aggregated Resources. When a client provides a webserver with the URI of the Aggregation, the Resource Map it references in its response is called the *authoritative* Resource Map. ORE requires that there be at least one authoritative Resource Map, although there may be more than one, each having its own unique URI. For example, an Aggregation may have multiple serializations (e.g., RDF/XML, RDFa, Atom XML) of its Resource Maps. In any case, all authoritative Resource Maps that describe the same Aggregation are required to assert the same Aggregation Graph. A *non-authoritative* Resource Map may describe an Aggregation; however, the web server will not reference it in responding to a request for the URI of the Aggregation.[12]

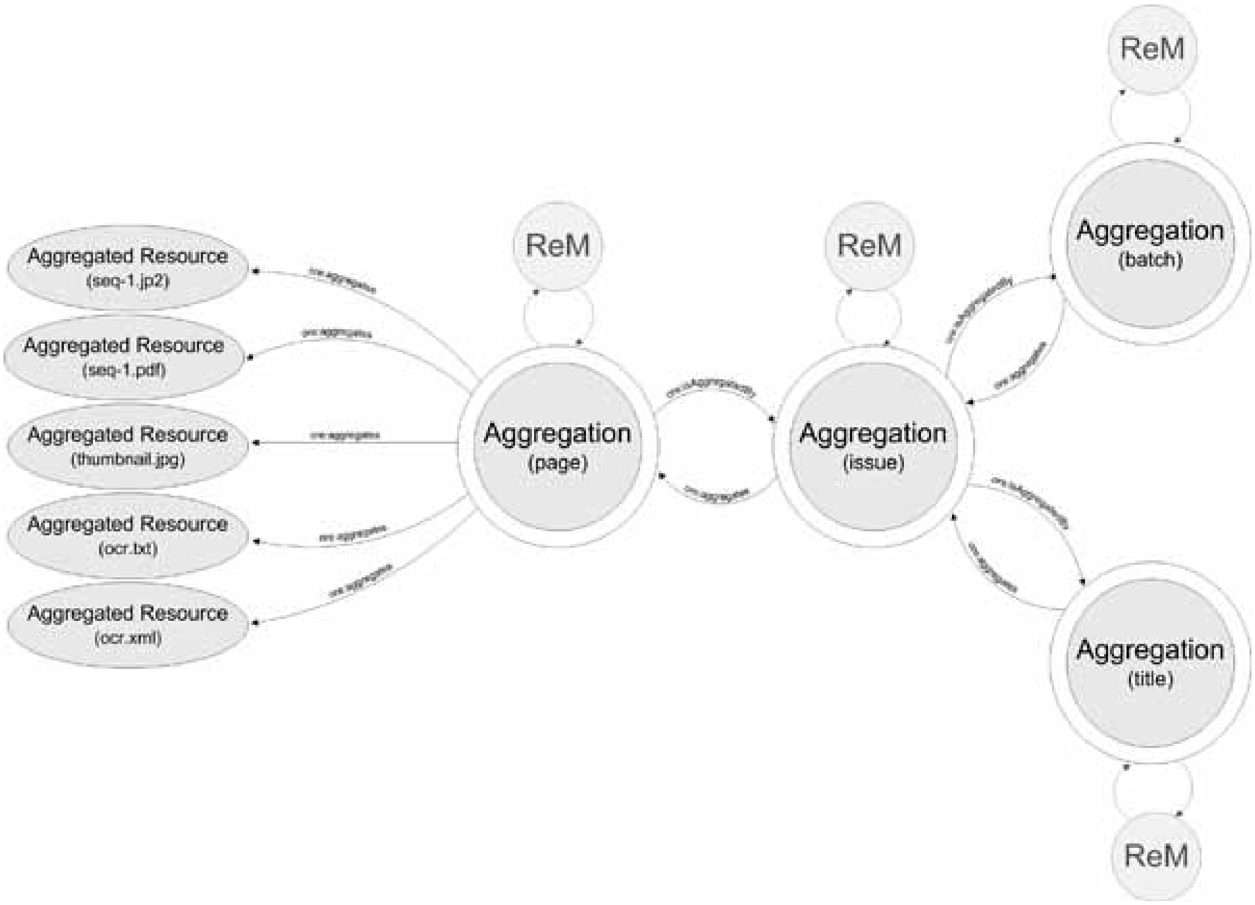


Figure 8
Putting it all together

Putting It All Together

Using and graphing RDF triples effectively demonstrates the ORE data model. In this chapter, we have explored the relationship between a Resource Map and its Aggregation, metadata, Aggregated Resources, and nested Aggregations. We've used an example, the National Digital Newspaper Program and the Library of Congress's Chronicling America website, to illustrate these concepts in an implementation (see figure 8). Keep in mind that this is not a comprehensive account of the ORE specification (for example, Proxies and other advanced concepts have not been presented), so you are encouraged to reference the documents in the last chapter for more information. In the next chapter, we will look at one way to serialize a Resource Map, RDF/XML, and investigate how Resource Maps can be exposed to clients by a web server and discovered.

Notes

1. Frank Manola and Eric Miller, "RDF Primer," World Wide Web Consortium website, Feb. 10, 2004, http://www.w3.org/TR/rdf-primer (accessed March 6, 2010).
2. Ibid.
3. National Digital Newspaper Program website, http://www.neh.gov/projects/ndnp.html (accessed March 6, 2010).
4. Ibid.
5. Ed Summers, interview by the author, January 21, 2010.
6. Ibid.
7. Carl Lagoze et al., "ORE Specification: Abstract Data Model," 2008, Open Archives Initiative Object Reuse and Exchange website, http://www.openarchives.org/ore/1.0/datamodel (accessed March 10, 2010).
8. Ibid.
9. "The Friend of a Friend (FOAF) Project," FOAF Project website, http://www.foaf-project.org (accessed March 6, 2010).
10. Lagoze et al., "ORE Specification: Abstract Data Model."
11. Ibid.
12. Ibid.
13. Ibid.