Implementations of ORE

Abstract

What are people doing with ORE in the real world? In this chapter we will explore eight different implementations of ORE that may be of interest to librarians. The Texas Digital Library created an implementation of ORE as a component of its digital library of electronic dissertations and theses. Microsoft External Research recently introduced the Zentity institutional repository and a plug-in for Word that generates Resource Maps. At Johns Hopkins University, librarians are participating in e-Science initiatives with the U. S. National Virtual Observatory to help astronomers manage massive data sets. In Australia, the LORE tool was created as an extension to the Mozilla Firefox web browser to enable literary scholars to encapsulate their digital resources and bibliographic metadata as ORE aggregations. Lastly, we speak with Patrick Hochstenbach about his thoughts on ORE and the Biblio institutional repository and academic bibliography at Ghent University in Belgium.

Vireo: An ORE Implementation for DSpace

Many academic libraries provide services to support students and faculty in the submission and archiving of electronic theses and dissertations (ETD). In the state of Texas, the Texas Digital Library (TDL) is a consortium of eighteen universities and has a mission of providing common infrastructure, services, and training to support the scholarly communication needs of its member institutions.[1] Among other services, TDL provides platforms for hosting open-access journals and wikis, and it supports a federation of institutional repositories.

These institutional repositories are running the DSpace software. Some members host their own DSpace repository; some share a repository at another member's institution; and others use a shared repository hosted by TDL. These institutional repositories provide a publishing and archiving platform, typically for born-digital documents such as journal article preprints, conference papers, and technical reports. Some members were using their institutional repository for the submission of theses and dissertations, too.

> *Vireo*
> http://www.tdl.org/etds

With support from the Institute of Museum and Library Services, TDL began a process of seeking input from its stakeholders to design a new system for managing and preserving ETDs. The project, named Vireo, sought to leverage existing infrastructure, implement new workflows, and scale up to a distributed, statewide ETD system. The Manakin software was used to create a customized user interface for DSpace to enable students to submit their dissertations.[2] The dissertation and its related files and metadata are then stored in the local DSpace repository.

At a high level, DSpace repositories are organized into communities, collections, and items. Items are made up of metadata and bundles, which contain one or more bitstreams.[3] For example, a university department may constitute a *community*, which has a *collection* of technical reports. Each individual technical report may be represented as an *item* that includes a *metadata* record that describes the report and a *bundle* of files, which can be
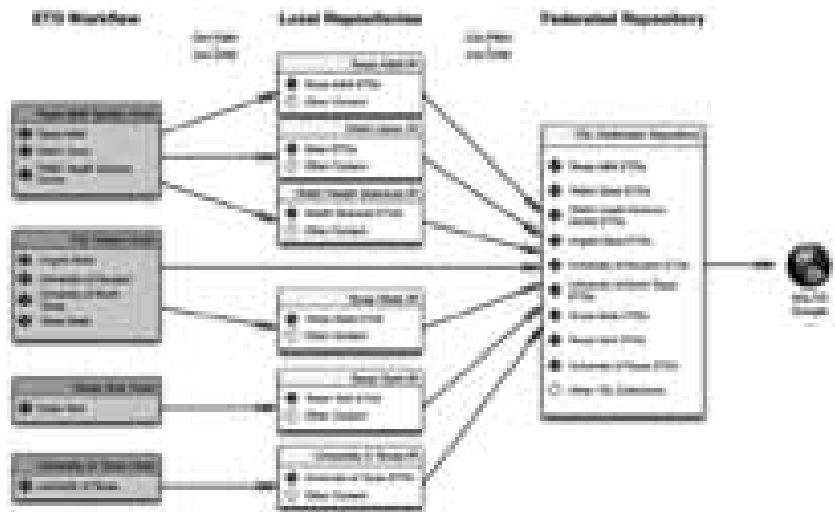
Figure 12
Federating DSpace repositories of electronic theses and dissertations at the Texas Digital Libraries

thought of as *bitstreams*. In this case, there may be a bitstream that represents an Adobe Acrobat file (PDF) of the report.

In mapping the DSpace data model to ORE, TDL decided to define aggregations for communities, collections, and items. It wrote some code to enable each DSpace repository to generate and interpret Resource Maps for these kinds of aggregations and to expose them as metadata records using DSpace's OAI-PMH interface. (Revisit chapter 1 for more information about the OAI-PMH.) The Resource Maps are serialized as Atom XML and can be harvested by an OAI-PMH service provider by specifying the proper metadata prefix.[4]

For its central DSpace repository, TDL employed an OAI-PMH harvester to harvest the Resource Maps and developed an ORE item importer. The ORE item importer resolves the URIs of the Aggregated Resources described in the harvested Resource Map, fetches them from the Remote DSpace repository, and rebuilds the item with its bitstreams and metadata in the central DSpace repository. TDL also built a custom scheduling system to automate harvesting.[5]

In this way, TDL is using ORE to harvest all of the dissertations and metadata from its member institutions into a central repository where they can be more easily preserved and made accessible in a single location. Future plans include public syndication of the Resource Maps so that anyone on the Internet can access and use the ETDs in semantic applications. In fact, interest has already been expressed by water quality researchers in Texas who want to automate the harvest of data from dissertations that relate to their field.

TDL invested a great deal of time in developing and testing its software because it is implementing the

software in a production environment with thousands of users. It is planning to release its ORE modifications to DSpace as open source software in February 2010.[6]

Foresite

Foresite began as a project funded by the Joint Information Systems Committee (JISC) in the United Kingdom to produce a demonstration of the ORE standard by creating Resource Maps of journals and their contents from the JSTOR archive of academic journals and delivering them as ATOM documents to deposit in a DSpace repository using SWORD. The Resource Maps were ingested into DSpace as items that reference the content residing in JSTOR.[7]

Foresite libraries
http://code.google.com/p/foresite-toolkit/

Foresite is probably more commonly known for producing open source Java and Python libraries for constructing, parsing, manipulating, and serializing

We wanted to operate at the object-level, with a protocol that is focused on objects and not just the metadata. The intent with Vireo was not to produce something that can harvest metadata but something that can harvest metadata and items inside of collections . . . and that's one of the things that was exciting to us about ORE. We want to share everything we can possible share in order to maximize the scholarship contained in the Texas Digital Library.

—Mark McFarland, Texas Digital Library

Resource Maps. Both sets of libraries support the parsing and serialization of Resource Maps that are suggested in the ORE specification: Atom XML, RDF/XML, and RDFa. Additionally, they support serialization in Notation3 (N3), N-Triples, and Terse RDF Triple Language (Turtle).

ORE is used for describing compound/complex digital objects such as aggregations of journals, issues, articles, and pages within JSTOR and enabling the digital preservation all of the copies of a resource. Of the two sets of libraries, Foresite's implementation of ORE is more complete in Python than Java. In the Python libraries, Foresite hides the ORE data model (in RDF) underneath an object-oriented layer and familiar "pythonic" style. It was used to create ORE descriptions of the complete holdings of JSTOR, making available the graph of interconnected journals, issues, and articles, through structure as well as citations.[8]

JSTOR is currently modifying the Foresite code to use its own internal formats rather than the information exported in the original project. It hopes to make the Resource Maps available to users at some point in the future.[9] The Foresite libraries are available for download from Google Code.

Microsoft External Research

Microsoft External Research partners with universities to support research, traditionally in computer science, but also in other areas such as library science and e-Science. Along with supporting research projects that are directed outside of Microsoft, External Research engages in activities such as sponsoring academic conferences, providing fellowships and internships, and producing software tools to foster and improve the research process. Microsoft provided early support for the development of the ORE specification along with the National Science Foundation, the Andrew W. Mellon Foundation, and the Coalition for Networked Information.[10]

Zentity

The creation of the Scholarly Communication program within Microsoft External Research by Tony Hey in 2007 has also yielded many valuable contributions. One example is the Zentity repository, which was launched at the Open Repositories conference in Atlanta in 2009. Microsoft sought to build a new repository platform from scratch on top of its product stack: Microsoft Windows, SQL Server, and the Microsoft Entity and .NET Frameworks. Zentity provides a turn-key repository solution with a default set of user interfaces, workflows, and a schema that defines typical repository entities and relationships.[11] They made an effort to incorporate as many open community protocols as possible, including SWORD,[12] the OAI-PMH,

The most important thing to know about ORE is not ORE at all, but the web architecture on which it is based. Understanding how the web works, and then how RDF works, is more important than knowing the details of ORE. Once the fundamentals are understood, ORE is very straightforward. Understanding that it is not just another compound object format like METS or MPEG21 DIDL is the most important thing for libraries. It could be used for Archives in place of Encoded Archival Description (EAD), for compound objects in place of METS, for collections in place of proprietary databases or description formats.

—Robert Sanderson,
Los Alamos National Labs

and ORE, to enable interoperability and integration with other tools and services. An included toolkit and code samples allow developers to present data in original ways, demonstrating, for example, the relationships between a published paper, authors, research data, associated lectures, presentation slides, or PDFs.[13]

Open Repositories 2009
https://or09.library.gatech.edu

While Zentity is one of the newest players in the institutional repository space, it may be the most mature and tightly integrated ORE implementation that is currently available as a part of a repository platform. Any time you have the URI for an entity, you can retrieve a Resource Map from the data store that describes all of the entities' relationships. For example, if you have the URI of a person, you can see an aggregation of the papers that person has authored, or the lectures that person has given, or the papers that person had reviewed. Resource Maps for these aggregations are defined automatically and updated dynamically, essentially by querying the store and then serializing them

as RDF/XML. Even though it is built atop SQL Server, Zentity is designed to behave more like an RDF triple-store than a relational database.[14]

Other well-known repository software such as DSpace, Fedora, and e-Prints are considered to be *open source*, which means that their source code is published and freely shared. Depending on the license being used, software developers may be free to create their own additions to or implementations of the software that can be proprietary. In fact, some companies have begun selling their own commercial implementations and extensions of these repositories. In this way, these repositories may be seen as *open core* software, because the base repository Remains in the open source while the additions to it can be proprietary. Microsoft is planning on releasing the source code for Zentity as *open edge* software, which is exactly the opposite of open core software. In other words, the core of Zentity (e.g., SQL Server) is a proprietary, commercial product, but the extensions to create a repository application will be open source and can be

---

ORE is a way of assigning semantics to the web. The more data that is described and linked extends its capability and opens up more data to a long tail of creative and unintended uses.

—Alex Wade, Microsoft External Research

---

freely shared and modified.

Zentity can be freely downloaded from Microsoft's website. There is also a discussion board for Zentity that is hosted by the Microsoft Research Community.

Zentity download website
http://research.microsoft.com/en-us/downloads/
48e60ac1-a95a-4163-a23d-28a914007743

Microsoft Research Community discussion board for Zentity
http://community.research.microsoft.com/forums/90.aspx

Article Authoring Add-In for Microsoft Word

At this point, most ORE implementations focus on platforms that store data or relate existing data to each other. ORE can also be tremendously useful when it is integrated into tools that create new data, such as the Article Authoring Add-in for Microsoft Word. The add-in was originally developed to help authors use Word to

write articles in a format required by the National Library of Medicine. It enables more metadata to be captured and stored at the authoring stage and enables semantic information to be preserved through the publishing process, which is essential for enabling search and semantic analysis once the articles are archived within repositories.[15] The author can also directly submit the article to PubMed Central or another repository from directly within Word, using its SWORD functionality.[16]

In the past, librarians have put too much emphasis on the container: the book, the journal, the article. And by doing so, we have pigeonholed our collections. By throwing away the container and embracing approaches like ORE and Linked Data, it opens up our data to a wider field of discovery and use for much richer applications. ORE broadens the impact of data by making it machine-accessible.

—Alex Wade, Microsoft External Research

A demonstration of the add-in and a description of its ties to OpenXML can be found on YouTube, as well as a more detailed account of how it uses ORE. In a nutshell, the add-in attempts to make it easier for researchers to write articles. Authors can insert properly formatted bibliographic citations by directly querying PubMed Central from Word, and the add-in can automatically populate metadata (e.g., grant information, author affiliation) that the author used to have to enter into a web form before submitting. As authors insert data into their articles, the add-in records some of its semantics. For example, an author may embed a data set, workflow, or image that has its own URI into a document. When the Word file is saved, a Resource Map describing these Aggregated Resources is serialized as RDF/XML and embedded into the article's .docx file. In this way, a downstream ORE application can later extract the Resource Map and handle the article as an Aggregation.[17]

YouTube video: Scientific & Technical Article Authoring Add-in Tour
http://www.youtube.com/openxml#p/u/11/
EuhAokemuH8

The Article Authoring Add-In for Microsoft Word is still being enhanced, but the current version can be freely downloaded from Microsoft's website.

U. S. National Virtual Observatory

The goal of the U. S. National Virtual Observatory (NVO) is to "enable a new way of doing astronomy" by to making it possible for researchers to find, retrieve, and analyze astronomical data from ground- and space-based telescopes worldwide.[18] The NVO is sponsored by the National Science Foundation and is based at Johns Hopkins University (JHU). Librarians at the university's Digital Research and Curation Center were early collaborators with astronomers in building solutions for submitting, publishing, and curating data sets for the NVO community, applying many of the principles of library science to the management of large, astronomical data collections. Tim DiLauro, Digital Library Architect at the JHU Sheridan Libraries, describes their project:

> The overall goal of the project is to capture the data that is associated with publications, deposit them into a data archive, and enable services over the data in the archive. One of the most fundamental aspects of scientific scholarly communication is the ability to access and examine cited data. Without this ability, the very essence of the scientific method, with its requirement of validating results, becomes compromised. The NVO is playing a leadership role in building services for the astronomy community to access and analyze astronomical data. However, thus far the scope of the NVO has deliberately not included long-term data curation, focusing instead on data location and data access standards and protocols. One of the goals of our project, which is a collaboration of astronomers, a scholarly society, its publishing production partner, and research libraries, is to capture data that is related to a
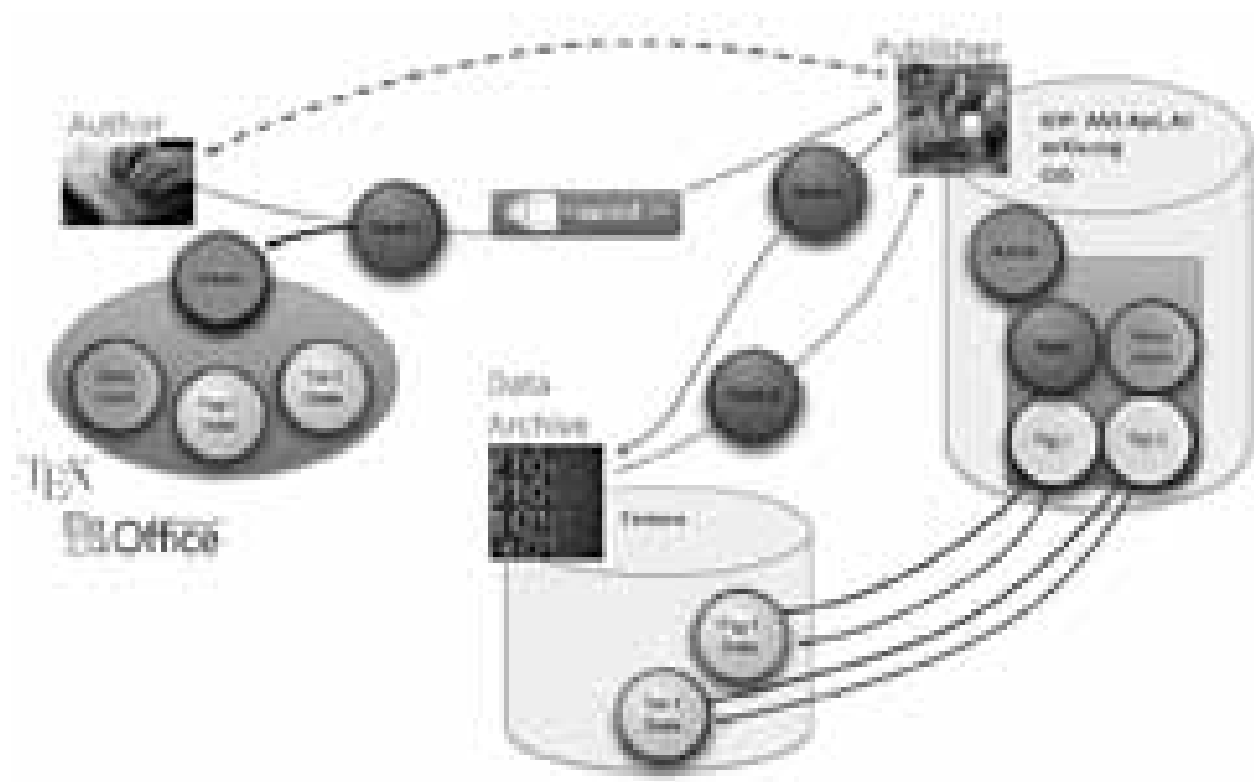

Figure 13
ORE data publishing workflow for the U. S. National Virtual Observatory

journal article when it is submitted [and archive it]. The challenges are several:

- To gather more metadata and datasets from authors without significantly increasing their workload,
- To simplify deposit process for authors and publishers, and
- To enable linking between articles and datasets without significant impact on publisher systems.

To accomplish these goals, we chose ORE as an enabling technology.[19] In the project, they are using ORE to support the description of the relationships between data and an article. For example, an article may include images, tables, and graphs that are embedded in it. Treating the article as an Aggregation, a Resource Map is generated that identifies and links the data behind these embedded objects and the article when it is submitted. If the article was written using Microsoft Word for Windows, the Resource Map can be created by the Article Authoring Add-In. JHU has also created a web-based application that can generate a Resource Map for other formats that are common in Astronomy, such as LaTeX. In either case, the document is submitted with its Resource Map using SWORD to the publishers.[20]

Unlike the other ORE implementations described in this chapter, JHU does not maintain the Resource Maps after they are generated. They are used by the publisher to link and ingest the article and its data when they are submitted. The publishers' systems can then track the relationships in their own way.[21]

JHU considered other options, such as METS, and specifically structMaps, but decided that ORE was a better fit because it was designed to express relationships among resources and to support the expression of complex objects. They also anticipate that the tools that will be developed for ORE will align more closely with their needs in the future.[22]

Along with DiLauro, who served on the ORE Technical Committee, Sayeed Choudhury from JHU's Sheridan Libraries helped develop the ORE specification by serving on the ORE Advisory Committee. The NVO is currently being operationalized by NASA as the U. S. Virtual Astronomical Observatory. JHU's implementation of ORE will rolled into the Data Conservancy,[23] one of two DataNet projects that was funded by the NSF in 2009.

LORE: A Compound Object Authoring and Publishing Tool for Literary Scholars

The Australian Literature Resource (also known as AustLit) is a collaboration between the National Library of Australia and twelve Australian universities to index and provide authoritative information on more than 100,000 Australian authors, going back to 1780.[24] Literature Object Re-use and Exchange (LORE) was created in this context as a lightweight tool to enable researchers to create and publish ORE-compliant literary objects that encapsulate their digital resources and bibliographic metadata.[25] LORE runs as a plug-in to the Mozilla Firefox web browser. It provides a graphical tool for drawing and labeling typed relationships between objects using terms from a bibliographic ontology. Metadata can be attached to the object, which can then be published as an RDF graph to a repository, where it can be searched, downloaded, edited, and reused by others.[26] It stores and queries Named Graphs that represent literary compound objects using web services on a Sesame 2 (RDF triple-store) or Fedora repository. The types for relationships among or between Aggregations and metadata that describe the Aggregated Resources are specified by an OWL ontology that was developed after examining the topic types and relationships present in AustLit. The ontology is based on FRBR but was extended to support additional relationships.

The LORE authoring interface displays a graphical visualization of the Resource Maps with their Aggregated Resources represented as nodes with arrows between them representing their typed relationships. A node presents a preview of its resource such as a thumbnail image, making it easy to locate and identify resources. Clicking on an identifier will load the resource in the web browser window. Along with the visualization, the Resource Maps are displayed as RDF/XML in a text window. New resources can be added as they are browsed in the browser window. A toolbar allows objects to be saved and loaded from the repository, and another panel enables the user to search and browse Resource Maps. Finally, metadata can be added or edited in the Properties panel.[27]

LORE was created as a part of the Aus-e-Lit project by the eResearch group at the University of Queensland[28] under the direction of Jane Hunter, who also served on the ORE Advisory Committee.

Notes

1. "About the Texas Digital Library," Texas Digital Library website, http://www.tdl.org/about-tdl (accessed March 6, 2010).
2. Scott Phillips, Cody Green, Alexey Maslov, Adam Mikeal, and John Leggett, "Manakin: A New Face for DSpace," *D-Lib Magazine* 13, no. 11/12 (Nov./Dec. 2007), http://www.dlib.org/dlib/november07/phillips/11phillips.html (accessed March 15, 2010).
3. MacKenzie Smith, *Lecture Notes on Computer Science*, vol. 2458, 2002, 543–549 http://dspace.mit.edu/handle/1721.1/26706 (accessed March 11, 2010).
4. Adam Mikeal, James Creel, Alexey Maslov, Scott Phillips, and John Leggett, "Large-Scale ETD Repositories: A

Interview with Patrick Hochstenbach, Ghent University

Please describe your overall project. In what context are you using ORE?

I am working with Ghent University's Academic Bibliography and institutional repository, which together we refer to as Biblio.[29]

Biblio is based on three subprojects:

1. Ghent University Library was asked by the University Department of Research Affairs to create a bibliography of all publications written by our university. The main purpose of the bibliography is to create bibliometric calculations. Faculty and department funding as well as doctoral promotions are based on the number of publications available in our bibliography. The bibliography needs to be complete, and the metadata quality needs to be very high to perform accurate bibliometrics. Ghent University researchers are required to register all their publications in the Biblio application in order to receive funding. Library staff enriches these registered publications by adding and enhancing their metadata.

2. The Biblio application is also an institutional repository. Our library is very open-access-minded. We are involved in many national and international projects to promote open access publishing and the archiving of scientific research output.

3. Ghent University is member of the European DRIVER project, which is a European Commission Seventh Framework program to optimize access to research data. Our group was involved in setting up a Belgian portal for institutional repositories, and we theorized about possible extensions to the OAI-PMH–based grid to allow for easier integration with new networks and exchange of complex objects.[30]

The Biblio project started four years ago as a DSpace implementation. However, new requirements became too difficult to implement in DSpace, and so we needed to migrate to a new solution. Two years ago, in a joint project between Lund University and Ghent University, we developed a new generation of the Biblio software from scratch that acts now as the combined bibliography and institutional repository of Ghent University.

Can you describe your particular implementation of ORE? How are you using it?

In our Driver Technology Watch Report, we theorized about several possible techniques to disseminate complex objects from institutional repositories. To test the theory in practice, we implemented several of these ideas in the Biblio application: ORE, MPEG-21/DIDL, METS, Sitemaps, microformats and simple CSV/Excel exports.[31]

The Biblio application uses a template system to generate webpages. Exporting ORE or any other format required adding a new template that creates RDF webpages instead of HTML webpages. We changed our code a bit to allow for HTTP Content Negotiation so that RDF web harvesters would be able to get easier access to the ORE exports.

What makes ORE well suited for this purpose? Did you consider any other options?

Yes, we considered all the options from our DRIVER report. All the options are very well suited in some way or another. Our partners inside the university tend to go for the simplest export possible: CSV/Excel. They have direct contact with our development team and can easily contact us to ask for new features. In national and international projects where this kind of direct interaction is not possible, ORE provides a very good way to join the Linked Data network[32] with all the available tools and semantics. However, big search engines like the Google, Yahoo, MSN tend to favor their own semantics, and here microformats would be better suited.

So, it depends a bit on which project or network partners you have and in which way (or better "if") you can negotiate on the what standards and protocols to use.

Do you make your aggregations publicly accessible, and if so, how? What do you envision people doing with them?

At the moment we export all the formats mentioned above, but we don't use them to import data (this is still on our to-do list). Our main task is to provide several alternative interfaces to our datasets. We encourage external developers to use these exports to provide extended services to our repository.

We aren't in direct contact with developers using our ORE datasets, but we see many crawlers accessing our site in search for these RDF exports.

Our experiments can give valuable input for our DRIVER partners to decide which format to promote in Europe. METS is popular in the UK, the Netherlands is using MPEG-21/DIDL, ORE is getting more and more popular linked also to the Linked Data bandwagon. Microformats are hugely popular outside the library world (big names supporting it) but doesn't get much traction in the digital repository world.

Case Study of a Digital Library Application," in *JCDL '09: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries,* 135–144 (New York: Association for Computing Machinery, 2009), http://doi.acm.org/10.1145/1555400.1555423 (accessed March 6, 2010).

5. Alexey Maslov, Adam Mikeal, Scott Phillips, John Leggett, and Mark McFarland, "Adding OAI-ORE Support to Repository Platforms," (presentation, Open Repositories Conference, Atlanta, GA, May 18–21, 2009), http://hdl.handle.net/1969.1/86479 (accessed March 6, 2010).

6. Mark McFarland, interview by the author, January 15, 2010.

7. Robert Sanderson, interview by the author, January 12, 2010.

8. Ibid.

9. Ibid.

10. Alex Wade, interview by the author, January 15, 2010.

11. Microsoft External Research Scholarly Communications program website, http://www.microsoft.com/mscorp/tc/scholarly_communication.mspx (accessed March 6, 2010).

12. Julie Allinson, Sebastien François and Stuart Lewis, "SWORD: Simple Web-Service Offering Repository Deposit," *Ariadne,* issue 54 (January 2008), http://www.ariadne.ac.uk/issue54/allinson-et-al (accessed March 6, 2010).

13. Microsoft External Research Scholarly Communications program website.

14. Wade interview.

15. "Article Authoring Add-in for Word 2007," Microsoft Research website, http://research.microsoft.com/authoring (accessed March 6, 2010).

16. Wade interview.

17. Wade interview.

18. "What Is the NVO?" U. S. National Virtual Observatory website, http://www.us-vo.org/what.cfm (accessed March 6, 2010).

19. Tim DiLauro, interview by the author, January 20, 2010.

20. Ibid.

21. Ibid.

22. Ibid.

23. Ibid.

24. "About AustLit," AustLit website, http://www.austlit.edu.au/about (accessed March 6, 2010).

25. Anna Gerber and Jane Hunter, "LORE: A Compound Object Authoring and Publishing Tool for Literary Scholars Based on the FRBR," (presentation, Open Repositories Conference, Atlanta, GA, May 18–21, 2009), http://hdl.handle.net/1853/28466 (accessed March 6, 2010).

26. Ibid.

27. Anna Gerber and Jane Hunter, 2008. "LORE: A Compound Object Authoring and Publishing Tool for the Australian Literature Studies Community," In *Digital Libraries: Universal and Ubiquitous Access to Information: 11th International Conference on Asian Digital Libraries,*

ICADL 2008, Bali, Indonesia, December 2008, Proceedings, ed. George Buchanan, Masood Masoodian, and Sally Jo Cunningham, 246–255 (Berlin: Springer-Verlag, 2008), DOI: 10.1007/978-3-540-89533-6_25.

28. "Compound Object Authoring and Publishing," University of Queensland eResearch, http://www.itee.uq.edu.au/~eresearch/projects/aus-e-lit/#compoundobjects (accessed March 6, 2010).

29. Ghent University, Academic Bibliography and Institutional Repository of Ghent University, http://biblio.ugent.be (accessed March 15, 2010).

30. Patrick Hochstenbach, Karen Van Godtsenhoven, Maurice Vanderfeesten, Rosemary Russell, Gerd Schmelz Pedersen, and Mikael Karstens Elbaek, *Driver Technology Watch Report* (Driver Project, 2008), http://hdl.handle.net/1854/LU-723558 (accessed March 15, 2010).

31. See also Patrick Hochstenbach, "Linked-Data in the Academic Bibliography," TekTok—Digital Library Technology Blog, Oct. 7, 2009, http://lib.ugent.be/tektok/2009/10/test.html (acessed March 15, 2010).

32. Linked Data website, http://linkeddata.org (accessed March 15, 2010).