Changing Definition of “Article”

• Primary unit of scholarly communication is becoming a multi-part, distributed object that often includes data and software

• Elements of a publication may reside in different repositories, maintained by different institutions, employing different technologies

• Goal of the project is to maintain and preserve the connections among these various components
Research Partnership

- Data Conservancy: Expertise in management of large data archives from multiple disciplines

- IEEE: Expertise in management of data-intensive scholarly journal publications

- Portico: Expertise in digital preservation, publisher workflow requirements, and existing relationships with 275 publishers
Work Plan

• Year One—Planning Phase: Gather requirements, create use cases, hold workshop with stakeholders, refine use scenarios based on community feedback

• Year Two—Prototype Development: Create system to identify, store, update, and retrieve relationships among publications and new forms of scholarly output, including data and software
Outcomes and Deliverables

• RMap tool working prototype

• Collaborative partnerships with the community

• System that supports emerging forms of digital scholarship and publishing

• Plan for sustainability of the project
TECHNOLOGY

The RMap Project

Data Conservancy

IEEE

PORTICO
Key Objectives

• Support assertions from broad set of contributors
• Integration with Linked Data
• Leverage existing data from other scholarly publishing stakeholders (publishers, identifier providers, data and software repositories)
• Some support for resources without identifiers
Data Model (simplified)
Data Model - Resource

- Things (abstract or concrete) that can have an identifier
- Basic building block of the WWW
- Key entity for description and retrieval within RMap
- Other core entities in the data model are also Resources
Data Model – RDF Statement (triple)

- Building blocks of the semantic web
- Conceptually of the form: `<subject> <predicate> <object>`
- Like subject-verb-object in English
- Distributed Scholarly Compound Object
- Primary unit of registration within RMap
- Basically a set of resources and related RDF description.
- Similar to OAI-ORE
Data Model - DiSCO
Data Model - DiSCO
- A person or thing (or group of these) responsible for some action
- Distinction between scholarly (e.g., author, funder, publisher, data processing program) and system (RMap component, user, etc.)
Data Model - Event

- An action or activity involving System Agents and other resources
- “Capture” provenance within RMap system
- Provenance of Scholarly Resources can be captured separately by registering it in RMap via DiSCOs.
REST API

• Benefits
  – Programming language independent
  – Abstraction away from underlying implementations and models

• Decisions
  – Resource request paths include API version.
  – Stick closely to web architecture metaphor, but stray when necessary.
## REST APIs (subset)

<table>
<thead>
<tr>
<th>Function</th>
<th>HTTP verb</th>
<th>API rel path (base=\api/{version})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieve related triples</td>
<td>GET</td>
<td>{resourceURI}/stmts</td>
</tr>
<tr>
<td>Retrieve related events</td>
<td>GET</td>
<td>{resourceURI}/events</td>
</tr>
<tr>
<td>Retrieve related DiSCO(s)</td>
<td>GET</td>
<td>{resourceURI}/discos</td>
</tr>
<tr>
<td>Create DiSCO</td>
<td>POST</td>
<td>/disco</td>
</tr>
<tr>
<td>Retrieve DiSCO</td>
<td>GET</td>
<td>/disco/{discoId}</td>
</tr>
<tr>
<td>Update DiSCO</td>
<td>POST</td>
<td>/disco/{discoId}/update</td>
</tr>
<tr>
<td>Delete a DiSCO</td>
<td>DELETE</td>
<td>/disco/{discoId}/delete</td>
</tr>
<tr>
<td>Retrieve an Event</td>
<td>GET</td>
<td>/event/{eventId}</td>
</tr>
<tr>
<td>Get DiSCO(s) related to event</td>
<td>GET</td>
<td>/event/{eventId}/discos</td>
</tr>
<tr>
<td>Perform SPARQL query</td>
<td>POST</td>
<td>/sparql</td>
</tr>
</tbody>
</table>
API Specification and Documentation

- Behaviors
  - Should come first
  - Just now clarifying
- API paths
- Data Models
- Serializations (content negotiation)
- Implementations
API Description (simplified simplified)

Function: Update DiSCO
• Behavior within RMap
  – Failed requests will be rolled back, so as not to require manual cleanup (transaction)
  – Insufficient authorization will result in failed transaction and offer to authenticate with other credentials.
  – A new DiSCO will be instantiated; the previous (old) DiSCO will be marked “inactive”
  – Add triple <new-DISCO-URI> <hasVersion> <old-DISCO-URI>
  – Resources will be instantiated for objects without identifiers (e.g., citation as string)
  – Scholarly Agents will be instantiated for agents lacking URIs (e.g., creator as string)
  – Event(s) created capture activity
• Request
  – Verb/relative path: POST /disco/{id}/update
  – Path parameters: {id} - URI of existing (old) DiSCO
  – Model: Resources + relationships (like OAI-ORE)
  – Serializations: RDF/XML, Turtle, or JSON-LD
• Response
  – Model: (custom)
  – Serializations: JSON, HTML
  – New DISCO URI in header: Location: <new-DISCO-URI>
  – Old DISCO URI in header: Link <old-DISCO-URI>;rel="predecessor-version"
  – Event URI(s) in header: Link <event-URI>;rel="http://www.w3.org/ns/prov#wasGeneratedBy"
  – [Enumerate response codes, labels, and their meanings]
Implementation Workflow (example)
Implementation Workflow (zoomed)
API Coverage

- Current focus on APIs to populate and access the graph
- Future focus
  - Authentication
  - Administrative
  - Composition & normalization
  - Inference engine
  - Operability
Technical Team Activity

- Developed initial data model
- Currently specifying and prototyping APIs
- Participation in RDA Data Publishing groups
- Prototype platform implementation planned for March 2015
Why is RMap Needed?

• The scholarly article is still the primary unit of scholarly communication
• But, articles are rapidly evolving into complex objects with many-to-many relationships
• These include multiple connections between article text, data, agents and their properties
• Describing and maintaining these relationships gives essential context to scholarly works
• Envisioning and making sense of this complexity in a cohesive way requires a new approach: RMap
An RMap Article Visualization
What Role Can RMap Play In the STM Community?

• It can provide a single unified and understandable view of complex scholarly articles
  – While several organizations provide various links and identifiers within and around STM articles, such as CrossRef and FundRef, there is no good single reliable view of all of these relationships currently
  – Comprehensive information about scholarly articles is siloed and difficult to assemble
• A clearer understanding of these articles, their constituent objects and of the relationships between them will enhance scholarly communication
• Research trends, author trends and funding trends will become more transparent
The RMap Community

- Publishers
- Authors
- Funders
- Data Repositories
- Librarians
- Others, for example:
  - CrossRef for citations and funding information (FundRef)
  - DataCite for relationships between articles and data
  - Dryad for relationships between articles and data
  - Pangaea for relationships between articles and data
  - Wiley and ACM
RMap Needs to Engage the STM Community to be Successful

- RMap is dependent on reaching out to multiple communities within STM to get them involved
  - Publishers, authors, funders, data repositories, identification providers, link providers and the like

- RMap needs data from data providers—data is crucial
  - Authors might be aware of data related to their articles. Authors, among others are very important.

- RMap needs people and institutions to use the provided data and find value in it
  - All of the above communities, along with others, such as librarians, will likely also be consumers of this information
Feedback from RMap Workshop

• RMap project should be a clearinghouse or meta-service that captures information about various data-linking services
• Initial use cases from original proposal were too broad and/or too ambiguous
• Need for clearer definition of the project, particularly regarding its specific goals
• Important not to replicate others’ work (particularly in the area of connecting publications and data) but to add value beyond what has already been done
Workshop Feedback Continued

- The challenge of “secondary data”, such as the inferred connections between publications and data or software remains unaddressed and important.
- The fact that the RMap Project has an established publishing partner is a comparative advantage.
- One approach would be to focus on the “input” side of the process (going after software and research workflows) in order to create a generalizable approach to gathering content.
Team Members and Acknowledgements

• Sayeed Choudhury, Tim DiLauro: Data Conservancy, Johns Hopkins

• Mark Donoghue, Gerry Grenier, Renny Guida, Ken Rawson: IEEE

• Vinay Cheruku, Karen Hanson, Amy Kirchhoff, John Meyer, Sheila Morrissey, Stephanie Orphan, Jabin White, Kate Wittenberg: Portico

This research project is made possible through generous support from the Alfred P. Sloan Foundation

We thank our workshop participants for their valuable feedback
Q&A

• Questions?

• For more information, please see:

• http://rmap-project.info/rmap/