Note: We used "Matchmaker" in our original proposal as a code name, but now call the project RMap. In addition, our approach to working with RDA has finalized into presenting RMap at two group meetings in Amsterdam and a follow-up workshop including various individuals involved in RDA groups.

Publishing and Preserving Data as Primary Research Objects: A Model for Future Scholarly Communication

1. What is the research question and why is it important?

The Data Conservancy (DC), IEEE, and Portico propose a two-year project to design and prototype data curation infrastructure that connects publications and data in innovative and persistent ways to advance scholarship and improve data management across a range of disciplines. Not only does the community require preservation of publications and data (either separately or together), but also it requires preservation of the *relationships* among them. As scholars continue to explore the possibilities presented by these relationships, it is incumbent on their colleagues in digital librarianship, data curation, and preservation to develop a creative vision and infrastructure to support this work. We believe that the models developed as a result of the project proposed here will enable new forms of scholarly communication, and thus set the stage for the future of research and digital publishing. Furthermore, a partnership that includes a data curation organization, a scholarly publisher with a vested interest in data management, and a proven and respected preservation service will provide the broad perspective and multifaceted experience that will allow this project to create meaningful solutions for the community.

2. What is the state of research on this question?

In February 2013, the White House Office of Science and Technology Policy (OSTP) charged federal funding agencies with research and development budgets over $100 million to develop policies and plans to support public access to publications and data resulting from federal funding. These memos have had the desired effect of focusing and galvanizing the government,

educational and private sectors toward building infrastructure for publications and data. As recent examples, the NIH has outlined a major program known as Big Data to Knowledge (BD2K) and additional funding agencies have joined NIH and NSF in requiring data management plans as part of the proposal submission process.

Even at this early stage, it has become clear that infrastructure needs to support both publications and data in a cohesive manner. One of the most promising modeling concepts is this regard is the OAI-ORE (Object Reuse and Exchange) that features the concept of resource maps (ReMs) or information graphs that describe aggregations of publications and data and—perhaps more importantly –the relationships between them. In a fundamental sense, these resource maps represent the modeling framework to ensure seamless, sustained (in the sense of preservation) connections between publications and data. Private sector companies such as Google, Amazon or Facebook use their own proprietary information graphs to describe and access content and services. OAI-ORE resource maps represent an open complement to such proprietary approaches. One could argue that OAI-ORE resource maps (that are often serialized in RDF) are "heavyweight" compared to approaches such as Schema.org.

However, it is worth noting that while web-based search and discovery is an important use case, it is not the only use case in the scholarly environment. In addition to modes of access not readily supported through web browsers (e.g., visualization and simulation), preservation needs mandate models and information graphs that account for provenance. In order to support a range of diverse content and services in an open, sustained manner, our community needs to develop its own set of information graphs that complement government and private sector approaches.

It is also worth noting Herbert van de Sompel's recent comments about the evolution of the Open

Archives Initiative. Van de Sompel has asserted that the first phase (OAI-PHM) focused on a lightweight protocol for discovery of collections. OAI-ORE represents an attempt to account for the compound nature of diverse, distributed scholarly objects. Van de Sompel has cited the ORE related work of Johns Hopkins as an exemplar in this regard. Van de Sompel has also pointed out that web technologies have matured sufficiently to take full advantage of information graphs in the form of OAI-ORE resource maps.

Our proposed work represents the culmination of years of design, architecture and prototyping of models and systems that integrate publications and data into scholarly compound objects for greater discovery, access and preservation. DC, IEEE, and Portico possess unique and significant expertise, capacity and relationships that will bolster the prospects for success and widespread adoption of the results of this work to support new directions in scholarship. DC's research has demonstrated that data management is most effective when it occurs early in the data or project life cycle. However, a vast amount of scholarly output exists in the form of publications and cited data that represent a later snapshot of the research process. While the prospects for future data management further upstream in the research process should be explored, there exists an immediate and pressing need to make the most effective use of publications and data connected during the important formal publication stage of research.

To advance research and improve data management during the publication phase of the scholarly process, DC, IEEE, and Portico will design a framework that connects publications and data—both cited and uncited—and preserve that connection, while also providing a pathway to richer forms of publication that include other elements of the research process (e.g., provenance) or types of content (e.g., software, visualizations). The proposed research and prototyping represent fundamental advances in the state of the art, yet also build on existing capabilities, thereby

resulting in work that is simultaneously ambitious and tractable. Data Conservancy's existing pilot project with arXiv.org represents one of these existing capabilities that the proposed work will leverage (researchers may simultaneously upload papers and datasets to arXiv, which in turn transfers the datasets to Data Conservancy—the relationship between the papers and datasets is maintained at both locations).

This arXiv.org pilot represented an early exploration of the OAI-ORE capabilities of DC and integration of data submission as part of an existing submission workflow. The pilot demonstrated DC's capability of supporting multiple persistent identifier schemes by confirming its ability to attach arXiv.org native identifiers within the information graph generated by DC. Even with the caveats of a pilot, several individuals chose to use this data deposit mechanism. The experience gained from examining file formats, metadata, etc. represent another important source of knowledge that will guide this proposed work.

There are existing examples of publication connected to cited data. Researchers currently can upload datasets to Figshare (http://figshare.com), which results in the creation of a DataCite (http://www.datacite.org) Digital Object Identifier (DOI). DataCite itself provides other services (all, as of the time of writing, described as being in beta), including a metadata search for datasets registered with DataCite, an OAI provider that exposes DataCite metadata for OAI-PMH, a content resolver that can expose metadata stored in the DataCite metadata store and can redirect queries to repositories containing the data set referenced by the metadata and a DOI citation formatter. Additionally, as part of the EU-funded ORCID and DataCite Interoperability Network (ODIN) project (http://odin-project.eu), a beta service has been developed to enable researchers to add research datasets and other content with DataCite DOIs to their Open Researcher and Contributor ID (ORCID) (http://orcid.org) profile by integrating with the

DataCite Metadata Store. OpenAIREPlus, the successor project to the EU-funded OpenAIRE project (http://www.openaire.eu/en/home) is aimed at linking the aggregated research publications aggregated via the OpenAIRE portal to the accompanying research and project information, datasets and author information. The efforts of OpenAIREPlus are part of the focus of Research Data Alliance (RDA) Europe Open Data & Publications Interoperability Working Group. (Please see section 4 below for proposed collaboration between this project, OpenAIRE and RDA).

The use of information graphs to capture and make operational links among data is increasingly a focus of the Linked Open Data in Libraries, Archives and Museums (LODLAM). As part of its research into ways to make "the execution context, within which data is processed, analyzed, transformed and rendered, accessible over long periods", the EU-co-funded TIMBUS project (http://timbusproject.net/events/events/206-from-preserving-data-to-preserving-researchcuration-of-process-and-context) and the Workflow 4Ever team (http://www.wf4ever-project.org) have explored the use of resource descriptions graphs.

The proposed work will advance the state-of-the-art in the following ways:

- The resulting framework and prototype will represent the connections among cited and uncited data and publications with a graph-based view that captures many-to-many relationships rather than the point-to-point viewpoint of current systems.
- The framework will include preservation of the connection between the data and publications.
- This infrastructure will be designed and prototyped with a multidisciplinary approach from the onset, thus reducing the dependencies or idiosyncrasies that often arise from disciplinary specific approaches.

- With this generalized approach, the results of this effort will be well suited for cross-disciplinary connections between publications and data (including types of data typically unaccounted for, such as software, visualizations, etc.). Together, the three organizations have a deep understanding of digital scholarly publishing and preservation, including the related issues of long-term access and reusability of content. This nuanced and long-term viewpoint is essential for designing and implementing holistic, modular systems that will persist over time.

3. Why is the proposer qualified to address the research question for which funds are sought?

Data Conservancy, IEEE, and Portico bring a wealth of essential experience and capacity to the proposed work through existing infrastructure, expertise, access to publications, data from different domains, and relationships with publishers, professional societies, scientific communities, and global data infrastructure projects.

Data Conservancy brings expertise in management of a large data archive with data from multiple disciplines, and participates actively with the Federation of Earth Sciences Information Partners (ESIP), NSF's Earthcube initiative, collaborates with the Virtual Astronomical Observatory (VAO), and held partnership discussions with the Global Names Architecture (GNA) and the InterUniversity Consortium for Political and Social Science Research (ICPSR).

IEEE publishes nearly 30% of the peer-reviewed scholarly output in electrical engineering and computer science through its 120 journals, 1000 annual conferences and its IEEE standards. It brings expertise in management of complex, data-intensive scholarly journal publishing, close ties with a distinguished professional society, understanding of author requirements, and engagement with the scholarly user community.

Portico brings deep expertise in digital preservation and publisher workflow through its work with content from more than 17,000 journals and 200,000 books across the spectrum of scholarly disciplines. Portico has strong and long-standing relationships with more than 200 publishers from 18 countries around the globe, including large and small, commercial, not-for-profit, university presses, and open access publishers. Portico will provide experience with publisher requirements, preservation standards, engagement with the publishing and preservation communities, and sustainability planning.

4. What is the research methodology?

The proposed work will consist of two components: a vetting phase and a prototype phase. For the vetting phase that will occur during the first year, IT architects and infrastructure leads from DC, IEEE, and Portico will build on the prior work of the arXiv.org pilot project to extend the modeling framework in the broader context discussed here.

- Gather requirements from partners (listed above) and other interested parties. [We plan to include the broader community here and will host a workshop in year one with key stakeholders in order to integrate their needs and perspectives into our planning process.] Refine use cases, scenarios, and/or stories to inform subsequent work. For example:
    - Develop abstract set of relationship assertions between data and publications.
    - Determine properties/attributes of relationship assertions (linkage, relationship, source, provenance, required attributes).
    - Refine models of some typical workflows for creation of relationship assertions.
    - Support discovery of relationships.
    - Consider mechanisms for disseminating of relationship assertions.

- Consider broader issues:
    - These assertions will not exist in a vacuum. How will they be combined with other assertions both within and beyond our own systems?
    - How would these integrate with the broader LOD world?
    - Reputation: The fact that the asserter is one of the creators of both the data and the publication might be important.
- Develop and document plan for prototype phase.

In addition, DC PI Choudhury has discussed the proposed work with the EU-funded OpenAIRE project and the Australian National Data Service (ANDS), both of which expressed support and interest in participating in the design process. By working with these global efforts, the project team can tap into additional expertise and ensure that the results are applied and amplified in a global context.

During the design phase, Choudhury, IEEE PI Grenier, and Portico PI Wittenberg will lead the effort to leverage existing relationships and identify data providers and possibly additional publishers that will work on prototypes. In order to demonstrate the most robust platform, it would be ideal to identify partners from the physical sciences, social sciences, and humanities.

During the second year, the project team will implement a prototype system that instantiates the design framework and demonstrates the capabilities of many-to-many connections represented through a graph-based view across disciplines.

- Determine how abstract information will be stored within our existing system implementations.
- Create concrete serializations for relationship assertions and any other information

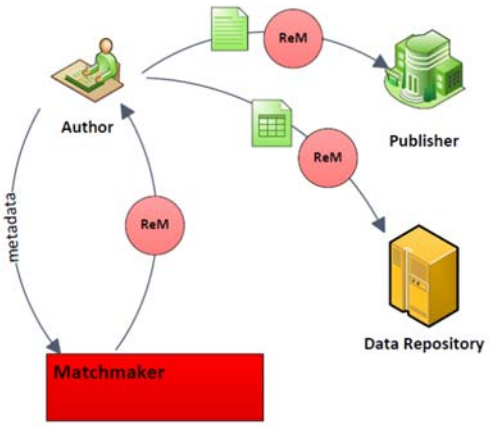
abstractions that they must move among.

- Develop APIs and implement in our reference systems (presumably Portico/JSTOR, IEEE publications, and DC).
- Create documentation of specifications.
- Prototype and demonstrate the key functionalities determined during the design phase.

Use Cases

Our proposed solution is a service, tentatively named Matchmaker, which builds relationships, stores relationships, updates relationships and allows those relationships to be retrieved. We describe some use cases below, prioritized in order of their importance for the user community.

1. An author is about to submit a paper to a publisher and a dataset to a repository and would like to send a ReM defining the relationship with both research outputs.
    - The author would submit the article and dataset metadata to the Matchmaker (perhaps through a widget) and request a relationship.
    - The Matchmaker would mint identifiers for the article & dataset and store the identifiers, metadata, and relationship in its database.
    - The Matchmaker would return a ReM to the Author.
    - The author could submit this ReM to the publisher and data repository along with the article and dataset.



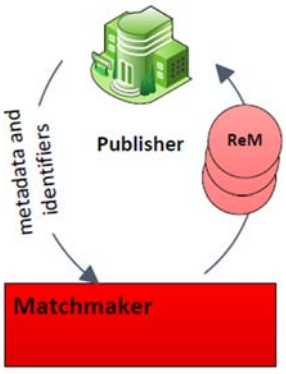
2. An author submits a paper and dataset to a publisher for publication and the publisher

creates a ReM defining the relationship between the research outputs:

- The publisher assigns DOIs to both the dataset and the article.
- The publisher could choose to create a ReM and submit it for preservation with the Matchmaker, or
- The publisher could submit the metadata and identifiers for both objects to Matchmaker and ask Matchmaker to create the ReM.

3. An author submits a paper to a publisher and a dataset supporting the research to a repository and the repository creates an ReM defining the relationship between the research outputs:
   - The publisher assigns a DOI to the article.
   - The repository assigns an identifier to the dataset.
   - The author links the paper to the dataset within the repository interface.
   - The repository creates a ReM and submits it for preservation with the Matchmaker.

4. At the time of publication, a publisher would like to determine if there are any existing relationships to an article, its author, its federal grant, etc. that it can include as reference links.
   - The publisher submits the article metadata and identifiers to the Matchmaker and requests any ReMs.
   - The Matchmaker looks them up in its relationship database and returns any ReM(s) it finds.

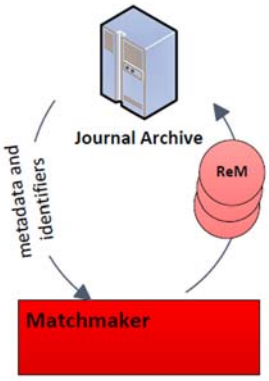


Note: the publisher may make this request as many times as it likes, perhaps once a year to update the article references on the website.

5. A journal archive is about to trigger content and would like to identify all relationships the articles in the triggered journal have with other resources.

   Note: this use case looks like the publisher use case above.

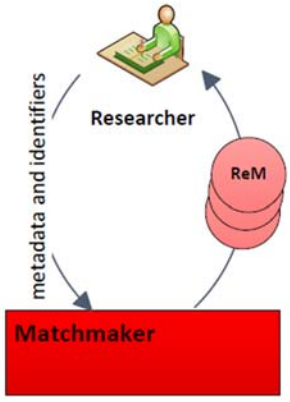
   - The journal archive submits the article metadata and identifiers to the Matchmaker and requests any ReMs.
   - The Matchmaker looks them up in its relationship database and returns any ReM(s) it finds.



6. A researcher has found an article describing a methodology and results she would like to replicate and before doing so would like to see the dataset produced by the author of the article.

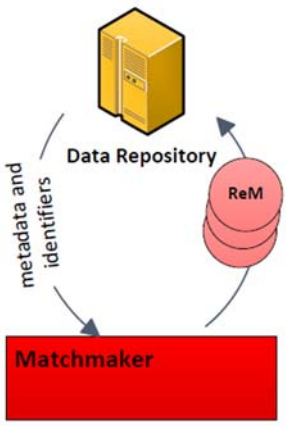
   Note: this use case looks like the previous two use cases.

   - The researcher submits the article metadata and identifiers to the Matchmaker and requests any ReMs (perhaps through a widget).
   - The Matchmaker looks them up in its relationship database and returns any ReM(s) it finds.
   - The researcher can then use the identifiers in the ReM to retrieve relevant datasets.



7. A data repository would like to keep its page of articles that reference this dataset current and queries the Matchmaker monthly for active relationships.

   - The data repository submits the dataset metadata and identifiers to the Matchmaker and requests any ReMs.
   - The Matchmaker looks them up in its relationship database



and returns any ReM(s) it finds.

The key fact is that any party can query for relationships or assert relationships at any time. The relationships are ever changing and growing and the relationships the Matchmaker provides on day one may be different from those on day ten.

What will be the output from the research project?

The corpus of scholarly digital publications and datasets, like all content on the web, consists of distinct building blocks or resources. Conceptually, these resources (article PDFs, figure graphics, tables derived from datasets, datasets, etc.) create a graph of clearly demarked relationships. For example, an article may contain figure graphics that are derived from a dataset.

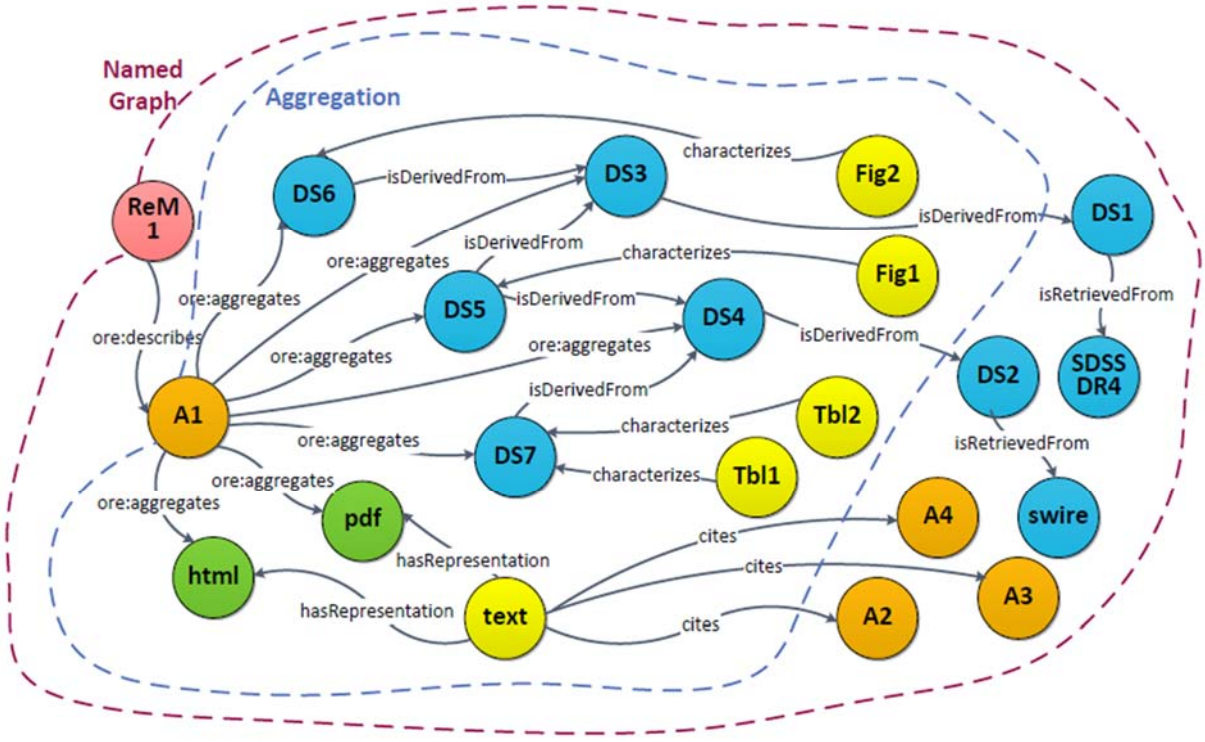


Figure 1: Named Graph representation of an article and its relationships

Within the current scholarly publishing environment, many relationships are not formally captured in a machine-readable manner. OAI-ORE has proposed that these relationships can be

formalized by defining aggregations of content in Resource Maps (ReMs). There remains, however, the problem of how to "make" these relationships. Some are known at the time of article or dataset creation or article publication and, if the author or publisher has the capacity they could create ReMs at those moments. Many publishers and authors, however, will not have this capacity. In addition, new relationships are created over time and thus will not be present in any initial ReM. DC, IEEE, and Portico propose a tool that will build, store, return and preserve ReMs to solve this problem in a manner that builds upon and extends existing work, generalizes to a variety of systems and emphasizes preservation of the resulting ReMs and associated objects. In a fundamental sense, the proposed work not only treats data as primary research objects but also designs and develops a key piece of infrastructure that builds persistent connections between data and publications in an ongoing manner.

The proposed work will be assessed and propagated through the Research Data Alliance (RDA; http://rd-alliance.org). As stated on its website:

> "The Research Data Alliance is an organisation that aims to accelerate and facilitate research data sharing and exchange. The work of the Research Data Alliance will primarily be undertaken through its working groups. Participation in working groups, starting new working groups, and attendance at twice-yearly annual meetings is open to all."

DC is already a key player within the RDA, with Choudhury and Sheridan Libraries IT Architect Tim DiLauro involved in or co-leading multiple working groups. Choudhury is also a member of the US Coordinating Committee. RDA has established a candidate working group focused on data and publications. Choudhury has spoken with several individuals and organizations involved in this working group. By hosting an RDA affiliated workshop at Johns Hopkins, the proposed work can be reviewed by an international group of stakeholders, many of whom have already developed relevant standards or prototypes. The proposed work would then become part

of this working group's final recommendations or outputs that are expected for completion by end of 2014. Correspondingly, DC, IEEE and Portico will complete relevant design work prior to the workshop, discuss and refine the approach during the workshop, and build a prototype system using DC and publishing systems after completion of the RDA working group's efforts.

Publisher & Scholar Engagement

DC has considerable expertise preserving and managing data and datasets and diverse partnerships with data archives, while Portico has extensive experience working with publishers to establish and maintain procedures for delivery, receipt, and transformation of scholarly content. Portico has a staff member dedicated to fostering relationships with scholarly publishers and is currently receiving, normalizing, and preserving content from 180 e-journal publishers. Portico's publisher relations director has become deeply familiar with the issues and challenges faced by its publishing partners as well as the strategies that are most effective in working with them productively. Portico has learned that it can be difficult to engage publishers in new projects if such engagement requires that they make changes in their workflows to accommodate data requirements. Publishers are often dealing with product/platform launches and associated content migrations, and focus on meeting their primary business needs before engaging with new work. Most publishers are already operating at capacity with the staff they have, and without a pressing internal business need, often have the interest but not the resources to explore new directions with a third party. By entering this space, IEEE, Portico and DC can ease the burden on scholars and publishers, by leveraging existing relationships and existing data stores of content and prototyping a model that can be used by the broader community

The End Products

At its heart the matchmaker is a workflow engine into which new modules may be added to manage or build relationships. Requests to the matchmaker will be made through web services, isolating the community from the technical details of implementation.

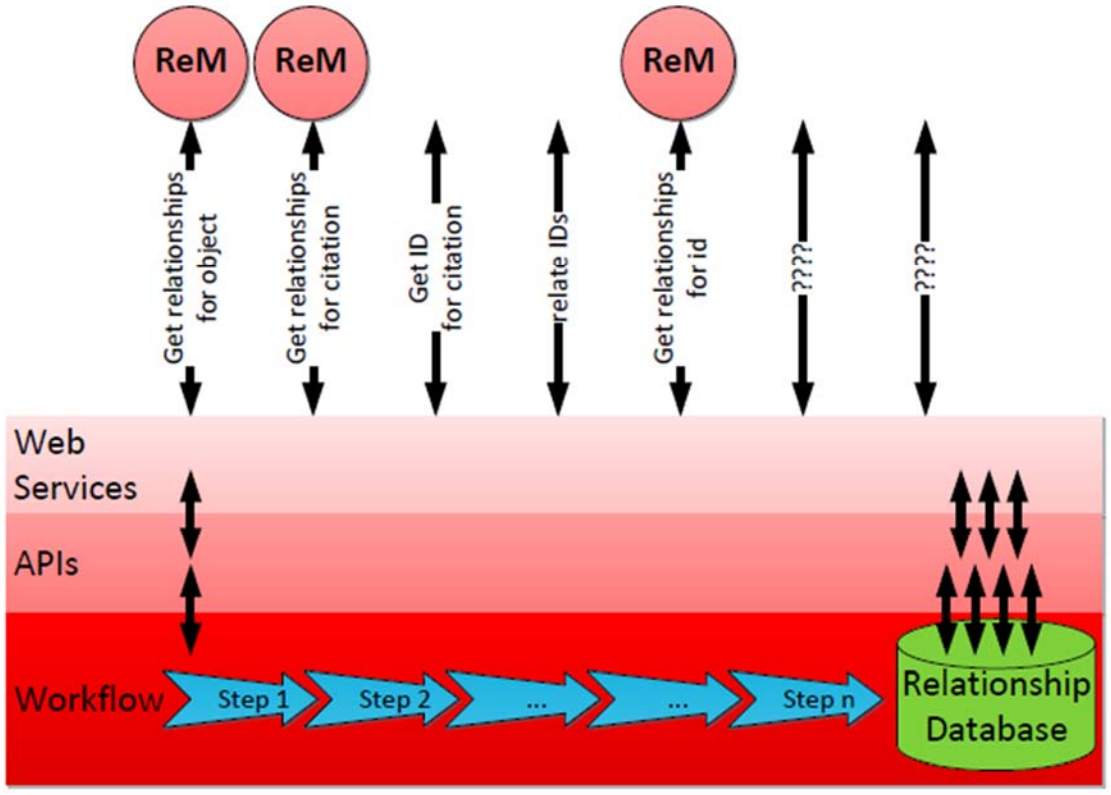
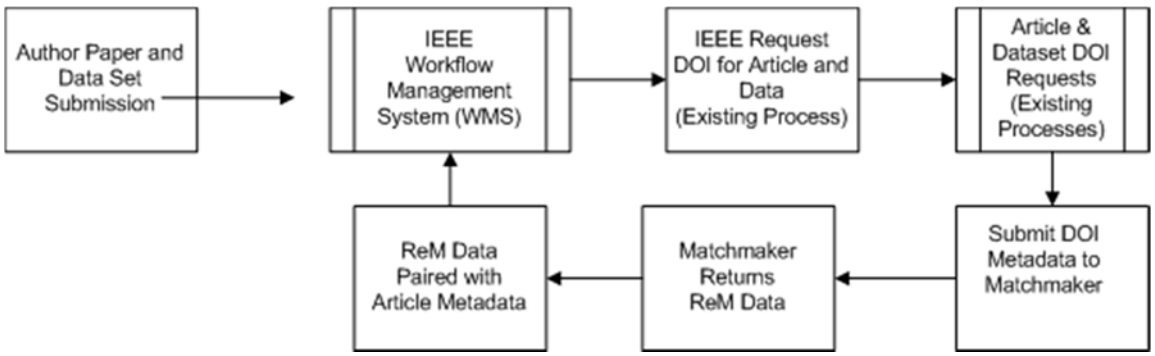


Figure 2: Matchmaker details

The team will develop requirements and a specification for Matchmaker during the initial phase of development. This will be followed by an abstraction and reference implementation to improve the design, requirements and specifications. The abstraction and reference implementation will be shared with the broader community for comment. Based on our experience, it is likely we will develop Matchmaker on top of existing workflow software using Java and web services APIs, however we will vet the development plan and reference implementation through our workshop and RDA.

IEEE Prototype Matchmaker Workflow

IEEE's publishing workflow system currently interacts with CrossRef for the purpose of creating unique identifiers for articles as well as data sets; additionally IEEE can register data set DOIs with DataCite. Via a current automated process, article and dataset records are updated with DOIs within the IEEE publishing workflow management system. Extending this process, IEEE will call the Matchmaker service and generate a ReM for both the article and its associated data.



IEEE has a history of supporting industry-wide data standards and interchange. Most recently IEEE played a key role in the collaboration between U.S. research funding agencies and publishers to develop a service to relate funding grant numbers to published articles (a service named FundRef). Many of the technical principles behind FundRef will be re-used to build the Matchmaker service into the IEEE journal publishing workflow.

By the end of year two, the matchmaker will:

1. Create relationships

   Inputs
   - Identifiers from a select set of allowed identifier types
   - Metadata records from a select set of allowed formats
   - Relationship indicators

Results
- Create a unique, URI identifier for each relationship
- Output an XML representation of each relationship
- Store each relationship in the database for later retrieval

2. Add relationships to the matchmaker database
3. Retrieve relationships

Inputs
- One or more identifiers

Results
- An XML representation of the direct and indirect relationships of the item associated with this identifier to other items.

The matchmaker framework will be designed to simplify the addition of modules and by the end of year two we will implement modules that will parse full articles (such as those within the Portico archive) to create new relationships.

Outputs of this project will include:

- An expandable list of named vocabulary of relationship types
- An XML schema to support relationships (ReMs) that is published and widely available for use by any party
- The matchmaker service, including the workflow engine, APIs, web services and a limited set of widget GUIs (other organizations will be able to build on top of the web services, as well).

Sustainability Planning

We understand that planning for the long-term sustainability of the Matchmaker tool and the

related services is an important part of our work during the grant period so we have considered a number of issues related to this question. For software licensing, it is Portico's preference to make the Matchmaker code open source and to involve the broader community in its long-term sustainability. Having said that, we are open to a number of possible business models, and would hope to be able to explore this question with participants at our workshop. We believe that in essence, the value of the project is the connectivity of the tool—if the community finds it useful, there are various possibilities for ways in which some or all of those sectors could help to support it, including contributed resources from publishers, membership fees for participating organizations, and/or a distributed system for maintaining and developing the tool and the related services that provide value. At this point it is difficult to define exactly what level of work will be required and the staffing and dollar amounts involved, but Portico is open to a model in which it supports this ongoing work as part of its portfolio of services. Part of PI Wittenberg's contribution to the project in year one will be to help guide the exploration of possible long-term sustainability models and to write up and recommend some potentially viable models to be considered by the project team and community participants in year two. Wittenberg has had experience with sustainability planning and business modeling in her work at ITHAKA but also in her prior roles in digital publishing.