Interactivity, Distributed Workflows, and Thick Provenance: A Review of Challenges confronting Digital Humanities Research Objects

Katrina Fenlon (kfenlon@umd.edu; https://orcid.org/0000-0003-1483-5335)

Introduction

While Research Objects (ROs) are primarily oriented toward scientific research workflows, the RO model and parallel approaches have gained some uptake in the humanities, enough to suggest their potential to undergird sustainable, networked humanities research infrastructures. Digital scholarship in the humanities takes a great variety of forms that range widely beyond traditional publications, and which incorporate narratives, media, datasets and interactive components—any of which may be physically dispersed as well as dynamic and evolving over time. Despite the rapid growth of digital scholarship in the humanities, most existing research infrastructures lack support for the creation, management, sharing, maintenance, and preservation of complex, networked digital objects. ROs, and the community and tools that are growing around ROs, offer a potential, partial solution.

While the concept of the RO has seen significantly more uptake in the humanities than has the formal data model (Bechhofer, 2013; Belhajjame et al., 2015), several compelling applications of the concept that suggest the time is ripe for considering broader integration of the model into distributed infrastructures. These applications include platforms for data sharing and collaborative scholarship, platforms for digital and semantic publishing, and digital repositories in several domains.

This paper reviews existing applications of the ROs model to identify challenges confronting the application of ROs to humanities digital scholarship. This paper builds on Fenlon (2019), which investigated the application of the ROs model to digital humanities collections, and which identified three promising strengths of the model for the realm of digital humanities: (1) ROs readily perform the most essential function of a collection: to aggregate related resources in order to support scholarly objectives; (2) ROs have the capacity for explicit, semantic descriptions of interrelationships among components that are often hidden in digital humanities collections (and therefore vulnerable to dissolution); and (3) the RO model accommodates aggregations of linked data, offering researchers the opportunity to create and annotate virtual, fully referential collections.

Having identified some strengths and limitations of the RO model for digital humanities collections through one experimental application of model, this paper builds on that analysis by reviewing the literature on ROs in the humanities and examining a range of applications of the RO and similar models within humanities and cultural heritage domains. This paper frames the review around three main challenges and their implications for future implementations of ROs to support digital research in the humanities: First, digital humanities scholarship requires specialized interactive use, so realizing the advantages of ROs for the humanities will depend on implementations that create platforms for experimentation and development by communities. Second, the idiosyncratic workflows employed in the construction of networked humanities scholarship means that workflow-oriented ROs will not gain significant uptake in the humanities unless they can capture distributed, sociotechnical workflows in meaningful ways. Third, humanities ROs will require capturing provenance in ways and at a level of detail that may be unfamiliar to the ROs scientific origins; humanities scholarship requires “thick,” multilayered, context-rich provenance descriptions that can accommodate conflicting assertions and formalize uncertainty.

Challenge 1. Essential interactivity for specialized use

Much of humanities digital scholarship is *essentially* interactive. New modes of production and publication in the humanities are intended for user interaction or participation, and dynamic and responsive representation based on research context. Digital collections and archives, digital editions, maps, models, and simulations, and other modes of digital scholarship all rely on interactive components to express their interpretive contributions, or to enact their scholarly purposes. The interactive and dynamic components of digital scholarship include things like customized browsing and searching facilities that take advantage of extensive, rich scholarly encodings and annotations; platforms for collaborative annotation; dynamic maps and visualizations; etc. Such components are intended to do multiple things at once: to make arguments, to manifest interpretive stances, to enable knowledge transfer, and simultaneously to serve as active platforms for ongoing interpretation and research (Palmer, 2009; Fenlon, 2017; and others).

Prior empirical work on applying the RO model to digital humanities collections found the main limitation of the model for digital humanities collections to be that functional components, designed for ongoing end-user interaction, are not usefully captured in a basic RO model and instead fall to the implementations built on top of research-object management systems (Fenlon, 2019). ROs can, of course, accommodate as flat code objects that are intended to be interactive; and ROs have been employed for this purpose to support data migration and archiving (e.g., the RO BagIt profile). But the purpose of digital humanities scholarship is to be alive and functional, and for ROs to be useful in this domain will require implementations that support platforms for flexible, participatory development.

In a conceptual sense, the RO model has demonstrated value for this kind of platform approach in the humanities. The Perseids project offers a platform for sharing and peer-review of the transcriptions, annotations, and analyses that constitute research data in the Classics. The Perseids architecture is built around the concept of *data publications*, which are modeled as collections of related data objects. The Perseids team explicitly relates the *data publication* model to the RO model (Almas, 2017). Like ROs, Perseids *data publications* weave in several domain standards (including the TEI Epidoc schema, W3C Web Annotation, and others) to undergird an infrastructure that supports scholarly requirements specific to the Classics domain: transcription, fine-grained annotation, collaborative editing (with versioning), a research environment that facilitates data-type-specific extensions, and tailored workflows for peer review (Almas, 2017). Similarly, the CERES (Community Enhanced Repository for Engaged Scholarship) toolkit, created by the Northeastern University Libraries Digital Scholarship Group, explicitly draws on the concept of the RO in its system for supporting networked humanities scholarship and publishing. CERES allows digital humanities creators to build custom publications that pull objects from different repositories using APIs (including the Northeastern University Libraries' Digital Repository Service and the Digital Public Library of America) (Sweeney, Flanders & Levesque, 2017).

It is unclear how the RO model may fit into the broader, more diversified landscape of linked data and the Semantic Web in cultural institutions and in the humanities, but the conceptual fit within digital scholarship is established. ROs and similar models have substantial potential to underpin systems that support a variety of implementations. Realizing the advantages of ROs for the humanities will depend on implementations that create platforms for experimentation and collaborative development by distributed communities (Fenlon, 2019). Such platforms must accommodate dynamic interface-building, to allow scholarly communities with distinctive interests and needs to mobilize ROs in different ways. They must also accommodate participation and co-

creation through contributions of linked-data annotations and enrichments, including linking among ROs and the concepts and entities within ROs.

Challenge 2. Distributed and idiosyncratic workflows of networked humanities scholarship
Humanities digital scholarship is increasingly networked: heavily interconnected with and dependent on external resources for functionality and meaning. Many digital humanities publications in various forms—monographs, multimedia productions, exhibits, collections—draw on, reference, embed, and patch together distributed resources called from other collections, often via API. For example, a collection may center on a set of high-resolution images of primary sources, which are called from another digital library's IIIF image server. Some of the longest-running, large-scale cultural heritage digital libraries (including *Europeana* and the *Digital Public Library of America)* are aggregations of descriptive surrogates, which link to original content hosted externally. Externally maintained schemas, authorities, and utilities undergird digital editions. Visualization and mapping projects generate content using external services. And with the growth of linked data in cultural collections, projects increasingly leverage external data sources as primary content, to which scholars then add layers of interpretive narrative, annotations, context, and interconnection.

Humanities workflows rarely happen in self-contained or end-to-end research infrastructures, thwarting the possibility of sufficiently rich, *automatic* workflow capture. Indeed, efforts to build a workflow-oriented, unified cyberinfrastructure for supporting humanities scholarship tend to founder (e.g., Dombrowski, 2014). However, niche, task- or domain-specific infrastructures can capture constrained workflows. For example, in the domain of musicology, Page et al. (2017) observe how digital editions and annotations of encoded works are "manifestations of workflows deployed in musicological scholarship," and offer a compelling framework for representing musical ROs, which include images, text, audio, and encoded music (Page et al., 2017; De Roure et al., 2018). Computational workflows are readily captured within humanities research environments, and ROs have come into play for this purpose. For example, the HathiTrust Research Center Data Capsule environment is moving toward systematic provenance-capture for computational text analysis workflows. These workflows take as inputs *worksets* (Jett et al., 2017), which are conceptually and technically akin to ROs: aggregate digital objects that implement addressability for and relational expressivity among components using domain ontologies. Unlike ROs, *worksets* are envisioned as the inputs of workflows in the current model of the HathiTrust Data Capsule environment, rather than encompassing whole research workflows (Murdock et al., 2017). But workflow-oriented ROs will not gain significant uptake in humanities contexts unless they can also capture and make useful more complex, distributed, sociotechnical workflows in meaningful ways.

With their capacity for linked data using domain vocabularies, ROs readily accommodate many of the *artifacts* of networked digital scholarship in the humanities, along with their interrelationships (Fenlon, 2019). But can ROs accommodate humanities *workflows* in useful ways? In their effort to undergird DARIAH (pan-European infrastructure for digital arts and humanities research) through the systematic production of humanities ROs, Blanke and Hedges (2013) observed that humanities scholars employ sequential workflows, but "except in relatively specialised cases we rarely encountered workflows that could be automated, shared with and used by others, such as occur in many scientific disciplines." While auto-generated and computer-useable workflows may not apply to most humanities research processes, formally characterized, (semi-) manually captured workflows would be highly useful for review, validation, archiving, reproducibility, reuse, and other purposes. While the RO model has the capacity and flexibility for complex workflow representation, more research is needed to characterize humanities workflows;

to identify how such characterizations can be made useful; and to identify model extensions and unique implementation strategies workflows might require in different domains.

Challenge 3. Thick provenance

Drilling down on the problem of workflow capture, digital humanities scholarship places special demands on data provenance—not only on the provenance of digital resources (such as files, compound objects, datasets) or components thereof (such as passages of music, paragraphs of a text, or lines of a poem), but also the provenance of attached, contextual information. Archival artifacts—the evidence of the humanities—often possess simultaneous, multiple and parallel provenances (Gilliland, 2014; Hurley, 2005). Documenting the provenance of the evidence itself can be complicated, but beyond that, the provenance *of the provenance* must also be documented. Any assertion made about any artifact (in the form of metadata or annotation), or any contextual and secondary information attached to artifacts in the context of digital scholarship, require provenance. Annotations and metadata are often, in the humanities, products of scholarly, interpretive work. Therefore, each annotation or metadata proposition itself is subject to claims of authorship, competing perspectives, expression of uncertainty, and further annotation—all requiring provenance information.

Because provenance is a multilayered thing in humanities scholarship, different humanities disciplines and subdisciplines may require domain-specific provenance schemas and standards, which specialize existing standards for the expression of the provenance of different kinds of resources, ranging from digital media files to annotations. Humanities ROs will require thick, multilayered, context-rich provenance descriptions, which can accommodate conflicting assertions and formalize uncertainty. It is unclear whether existing implementations of the RO model can accommodate this level of description, though the model itself has the capacity.

The ResearchSpace environment (Oldman and Tanase, 2018) offers exemplary support for documentation of thick, multifaceted provenance of humanities ROs. ResearchSpace is an open-source platform created by the British Museum to facilitate scholarly data sharing, formal argumentation, and semantic publishing within communities of researchers. ResearchSpace does not directly employ the RO model, though its architecture does rely on aggregates of linked data, taking advantage of related standards including W3C Web Annotation and Linked Data Platform containers. In this environment, provenance and argumentation are expressed using the CIDOC-CRM specialization CRMInf (The Argumentation Model). Scholars can use this vocabulary to build narratives and thick descriptions around digital ROs through annotation and data-linking. These narratives of provenance allow and formalize the expression of uncertainty and competing perspectives, and the environment also serves to document the scholarly work that goes into building these narratives (ResearchSpace Team, 2018).

The reasons for highlighting the ResearchSpace approach to provenance in this review of humanities ROs are (1) to exemplify the unique demands of formalizing humanities provenance, and (2) to exemplify the highly distinctive, domain-specific implementation requirements that confront the RO and other domain-independent data models. Describing humanities provenance will require vocabularies to express argument and belief, as Oldman et al. (2015) observe. Beyond the RO model's use of Prov and Web Annotation, humanities provenance will demand domain-specific argumentation extensions such as CRMInf. It is clear that ROs can theoretically accommodate thick provenance description, just as they can theoretically accommodate the representation of highly complex workflows, but can they usefully undergird implementations that are centered in humanities research needs? The ResearchSpace interface is tailored toward knowledge work, toward the collaborative construction of multifaceted provenance descriptions,

without requiring users to code or gain expert-level knowledge of domain ontologies. Tools for the authorship of humanities ROs, or tools that implement ROs behind the scenes, may benefit from taking the same approach.

Conclusion

ROs make a great deal of sense for modeling cultural information; skeletons of a similar shape—the simple and powerful combination of aggregation and annotation to represent compound digital objects—already structures large-scale cultural data aggregations, e.g., through the *Europeana Data Model* and the *Digital Public Library of America Metadata Application Profile*, which are both founded on *ore:aggregations* plus *oa:annotations*. But the challenges confronting widespread application of the RO model to humanities digital scholarship are significant. This review of existing applications has identified three central challenges:

1. Digital humanities scholarship requires specialized interactive use, so realizing the advantages of ROs for the humanities will depend on implementations that create platforms for experimentation and development by communities.
2. The idiosyncratic workflows of networked humanities scholarship means that workflow-oriented ROs will not gain significant uptake in the humanities unless they can capture distributed, sociotechnical workflows in meaningful ways.
3. Humanities ROs will require thick, multilayered, context-rich provenance descriptions that can accommodate conflicting assertions and formalize uncertainty, along with implementations that support the documentation of such provenance.

In particular, the challenge of characterizing and formally expressing diverse humanities workflows, along with the provenance of data and contextual information within those workflows, presents the most urgent challenge and exciting opportunity for the future of humanities cyberinfrastructure. To many stakeholders in humanities cyberinfrastructure, “workflows are the new content” (Dempsey, 2016; Baynes et al., 2016; Schonfeld and Waters, 2018). While research on workflows is underway on multiple fronts (including Liu et al., 2017), it is clear already that there will be significant semantic differences between conceptual and technical elements in scientific workflows (and provenance) and those in the humanities; and these differences will affect the implementation of ROs for humanities research. Historically, attempts to implement scientific research infrastructures (including data models like the RO model) to support humanities scholarship have hit an obstacle in the form of semantic gulfs. For example, in the *Linking and Querying Ancient Texts* (LaQuAT) project, an effort to transfer eScience infrastructure in support of a humanities virtual research environment, Anderson and Blanke observed a fundamental challenge in integrating humanities data from different databases. They located the solution to that problem in humanities research communities: “integrating humanities research material...will require researchers to make the connections themselves, including decisions on how they are expressed and how to understand and explore the data more effectively” (Anderson and Blanke, 2012). Oldman et al. (2015), reviewing the state of linked data in the humanities, observed that basic linked data publication for many kinds of humanities sources can be counterproductive, “unless adapted to reflect specific methods and practices, and integrated into the epistemological processes they genuinely belong to.” This caution resonates with the challenges identified for the adoption of the RO model—or indeed for the importation of any data model, even domain-independent data models—into the humanities. The main challenges to implementing ROs for humanities research also present exciting opportunities for a more sustainable cross-disciplinary infrastructure (Fenlon, 2019), but implementation strategies must be centered in scholarly communities, and grow out from the practices, needs, and epistemologies of specific areas of study in the humanities and cultural institutions.

References

Almas, B. (2017). Perseids: Experimenting with Infrastructure for Creating and Sharing Research Data in the Digital Humanities. *Data Science Journal*, *16*(0). https://doi.org/10.5334/dsj-2017-019

Anderson, S., & Blanke, T. (2012). Taking the Long View: From e-Science Humanities to Humanities Digital Ecosystems. *Historical Social Research / Historische Sozialforschung*, *37*(3 (141)), 147–164.

Baynes, M. A., Sommer, D., Melley, D., & Lickiss, T. (2016, April). *Workflow is the new content: Expanding the scope of interaction between publishers and researchers*. Panel presentation presented at the Society for Scholarly Publishing. Retrieved from https://www.sspnet.org/events/past-events/workflow-is-the-new-content-expanding-the-scope-of-interaction-between-publishers-and-researchers/

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., … Goble, C. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, *29*(2), 599–611. https://doi.org/10.1016/j.future.2011.08.004

Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., … Goble, C. (2015). Using a suite of ontologies for preserving workflow-centric research objects. *Journal of Web Semantics*, *32*, 16–42. https://doi.org/10.1016/j.websem.2015.01.003

Blanke, T., & Hedges, M. (2013). Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems*, *29*(2), 654–661. https://doi.org/10.1016/j.future.2011.06.006

De Roure, D., Klyne, G., Page, K., Pybus, J., Weigl, D. M., & Willcox, P. (2018, July). *Digital Music Objects: Research Objects for Music*. Presented at the Research Object workshop (RO2018) at IEEE eScience Conference 2018. Retrieved from https://zenodo.org/record/1442453#.XB6Chc9KhhE

Dempsey, L. (2016, October). *The Library in the Life of the User: Two Collection Directions*. Education. Retrieved from https://www.slideshare.net/lisld/the-library-in-the-life-of-the-user-two-collection-directions

Dombrowski, Q. (2014). What Ever Happened to Project Bamboo? *Literary and Linguistic Computing*, *29*(3), 326–339. https://doi.org/10.1093/llc/fqu026

Fenlon, K. (2017). *Thematic research collections: Libraries and the evolution of alternative scholarly publishing in the humanities* (Doctoral dissertation, University of Illinois at Urbana-Champaign). Retrieved from http://hdl.handle.net/2142/99380

Fenlon, Katrina. (2019). *Modeling Digital Humanities Collections as Research Objects*. Presented at the ACM/IEEE Joint Conference on Digital Libraries 2019. Retrieved from https://hcommons.org/deposits/item/hc:24889/

Gilliland, A. J. (2014). *Conceptualizing 21st-Century Archives*. ALA Editions.

Hurley, C. (2005). Parallel provenance [Series of parts]: Part 1: What, if anything, is archival description?. [An earlier version of this article was presented at the Archives and Collective Memory: Challenges and Issues in a Pluralised Archival Role seminar (2004: Melbourne).]. *Archives and Manuscripts*, *33*(1), 110.

Jett, J., Cole, T. W., & Downie, J. S. (2017). Exploiting graph-based data to realize new functionalities for scholar-built worksets. *Proceedings of the Association for Information Science and Technology*, *54*(1), 716–717. https://doi.org/10.1002/pra2.2017.14505401128

Liu, A., Kleinman, S., Douglass, J., Thomas, L., Champagne, A., & Russell, J. (2017). *Open, Shareable, Reproducible Workflows for the Digital Humanities: The Case of the 4Humanities.org "WhatEvery1Says" Project*. Presented at the Digital Humanities (DH2017). Retrieved from https://dh2017.adho.org/abstracts/034/034.pdf

Murdock, J., Jett, J., Cole, T., Ma, Y., Downie, J. S., & Plale, B. (2017). Towards Publishing Secure Capsule-based Analysis. *Proceedings of the 17th ACM/IEEE Joint Conference on*

Digital Libraries, 261–264. Retrieved from http://dl.acm.org/citation.cfm?id=3200334.3200367

Oldman, D., Doerr, M., & Gradmann, S. (2015). Zen and the Art of Linked Data. In *A New Companion to Digital Humanities* (pp. 251–273). https://doi.org/10.1002/9781118680605.ch18

Oldman, D., & Tanase, D. (2018). Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. In D. Vrandečić, K. Bontcheva, Mari Carmen Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, … E. Simperl (Eds.), *The Semantic Web – ISWC 2018* (pp. 325–340). Retrieved from https://link.springer.com/chapter/10.1007%2F978-3-030-00668-6_20

Page, K., Lewis, D., & Weigl, D. (2017). *Contextual interpretation of digital music notation*. Presented at the Digital Humanities (DH2017), Montréal, Canada.

Palmer, C. L., Teffeau, L. C., & Pirmann, C. M. (2009). *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*. Retrieved from OCLC Research and Programs website: http://www.oclc.org/content/dam/research/publications/library/2009/2009-02.pdf

ResearchSpace Team, British Museum. (2018, December). *Moving from Documentation to Knowledge Building: ResearchSpace Principles and Practices*. Presented at the Stiftung Preußischer Kulturbesitz (Prussian Cultural Heritage Foundation) Berlin. Retrieved from https://www.researchspace.org/docs/Berlin.pdf

Schonfeld, R. C., & Waters, D. (2018, April). *The turn to research workflow and the strategic implications for the academy*. Presented at the Coalition for Networked Information (CNI) Spring Membership Meeting, San Diego, CA. Retrieved from https://vimeo.com/271130388

Sweeney, S. J., Flanders, J., & Levesque, A. (2017). Community-Enhanced Repository for Engaged Scholarship: A case study on supporting digital humanities research. *College & Undergraduate Libraries*, *24*(2–4), 322–336. https://doi.org/10.1080/10691316.2017.1336144