Worksets Expand the Scholarly Utility of Digital Libraries

Kevin R. Page[1], Jacob Jett[2], Timothy W. Cole[2], Deren Kudeki[2], David Bainbridge[3],
Peter Organisciak[4], J. Stephen Downie[2]

[1]Oxford e-Research Centre, Dept. of Engineering Science, University of Oxford, UK, kevin.page@oerc.ox.ac.uk
[2]School of Information Science, University of Illinois at UC, USA, {jjett2,t-cole3,dkudeki, jdownie}@illinois.edu
[3]Dept. of Computer Science, University of Waikato, Hamilton, New Zealand, davidb@waikato.ac.nz
[4]Research Methods and Information Science, University of Denver, Denver, CO USA, Peter.Organisciak@du.edu

ABSTRACT

Scholars using digital libraries and archives routinely create worksets—aggregations of digital objects—as a way to segregate resources of interest for in-depth scrutiny. To illustrate how worksets can enhance the scholarly utility of digital library content, we distill from prior user studies three key objectives for worksets (extra-digital library manipulation, intra-item properties, and robust representations), and discuss how they motivated the workset model being developed at the HathiTrust Research Center (HTRC). We describe how HTRC's implementation of its RDF-compliant workset model helps to satisfy these objectives.

CCS CONCEPTS

• **Applied computing~Digital libraries and archives**   • Information systems~Retrieval tasks and goals

KEYWORDS

digital libraries, worksets, HathiTrust, semantic web, RDF

1 INTRODUCTION

Worksets have been introduced as aggregations of digital data that undergo analysis for scholarly study [1-3]. In this context we draw a distinction between archives, which provide evidence for the existence of digital objects in support of an investigation or argument; and digital libraries (DL), where the digital objects and their intellectual content are collected studied, analyzed, and synthesized in support of a scholarly hypothesis. Worksets take this a step further, as an additional mechanism for creating and

---

[1] With no intended implication of comprehensive coverage.

refining knowledge within a layered DL (as described in [4]), whereby the act of aggregation they facilitate is in of itself an intellectual product. In Section 2 we selectively describe three motivated objectives[1], which serve as specific illustrations of where worksets move beyond traditional DL functionality. In Section 3 we present key elements of an implemented workset model, and relate its utility to the requirements following from the three objectives.

2 KEY WORKSET OBJECTIVES

2.1 Objective 1: Extra-DL Resources

When working within a single DL, options to aggregate and manipulate digital objects into a workset are limited to the features and objects exposed by each DL's API or user interface. However, Fenlon et al. [5] have shown that scholars want the ability to incorporate external metadata and external digital objects within their worksets. Thus, there is a need for resource manipulation at an "extra-DL" level, to remove these single-DL constraints and to fashion worksets drawing from multiple sources, leveraging external metadata and including external resources. Realizing this objective enables inclusion of features produced by the computational analysis of workset items or tailored to domain-specific investigation of the items within a DL. In terms of [4], this illustrates the distinction between the 'collection layer' provided by the underlying DL(s), and the 'computational analysis' and 'exploratory analysis' layers realized through worksets.

2.2 Objective 2: Intra-Item Properties

As well as being a vessel for holding and manipulating extra-DL criteria, worksets must be a mechanism for processing features and/or properties *within* digital objects. These properties might be identified through computational analysis of item metadata and content, creating further metadata exposed, in turn, as criteria for forming and refining subsequent worksets. Interviews and surveys of HTRC's scholarly users [5] suggest that scholars desire a greater choice in the granularity of objects and object-metadata than is offered natively by a particular DL. For example, the default unit of granularity in the HathiTrust DL is the digitized volume; users express a preference for examining resources at the level of chapters, articles, individual pages or other textual divisions. Thus there is a need for worksets to support augmentation of item descriptions using intra-item metadata and to incorporate intra-item content.

2.3 Objective 3: Robust Representations

The representation of worksets must be adaptable into forms appropriate for different research stages and applications, while still conforming to community expectations for interoperability. For example, a scholarly analysis might require the inclusion of digital objects protected by copyright which would be required to remain within a DL's access controls. This predicates that any processing or manipulation of item content must be undertaken within a trusted infrastructure, i.e., following a non-consumptive research model. While a workset representation in a secure space can directly encapsulate restricted content, publically shared versions can only include objects by reference. Thus a workset model must be amenable to different levels of object representation according to circumstances, and offer consistency through the deterministic translation from one representation to another.

3 WORKSET IMPLEMENTATION

The implementation of worksets at the HTRC is built using a semantic web stack via an RDF-compliant ontology [2] serving as the data model for a Virtuoso triple store. This implementation realizes the aforementioned workset objectives in a number of ways, which we describe here.

Achieving Objective 1. Workset descriptions, including manifests of workset items, are maintained in a Virtuoso triple store. Each workset is stored as an individual RDF graph. Bibliographic metadata for each item in the HathiTrust DL is transformed into RDF and stored in the default graph of the triple store, with changes synchronized daily; in other respects the triple store exists as a resource decoupled from the core HathiTrust infrastructure. Items (content and metadata) referenced in a workset manifest, including any non-HathiTrust items, are linked using persistent URLs. e.g., HathiTrust volume handles. Benefits include the ability to add or subtract workset items without making tangible alterations to item metadata, and linking items and context enriching resources, e.g., granular extracted features and digital annotations, regardless of where stored (through predicates like "annotatedBy" and "hasFeaturesDataSet"). Domain-specific, computed, and extended workset-specific properties can be included simply by extending the workset graph. This affordance is inherent in the extensible and self-describing nature of RDF.

Achieving Objective 2. Our data model makes no assumptions regarding the objects which can be added to a workset. Everything is aggregated using an agnostic "gathers" predicate, with 'gatherable' resources typed accordingly (additionally/multiply typed). The use of an RDF ontology enables the easy overlay of these new information resources while retaining their domain semantics; from the semantic web perspective, as long as a source can be named (i.e., identified by a persistent global identifier, e.g., a URL) then it can be aggregated into a workset. This allows reuse of existing ontologies to describe the granularity required, e.g., schema:Article (bibliographic), fabio:Page (structural). This, in effect, reduces the effort required from scholars who wish to expand the descriptive fidelity within intra-item aggregations. Similarly, metadata for finer-grained, intra-item objects can be added by extending the graph with appropriate semantics.

Achieving Objective 3. Our workset implementation uses a 'manifest' to associate items with a workset, whereby the RDF encoding does not embed any actual item content. Depending on who, or what, is consuming the workset manifest, different localized workset formations may be realized. For example, the HTRC workset builder is a user interface for creating and refining worksets. It operates at the manifest level, fetching item metadata dynamically as needed, and never accessing item content (which is potentially in copyright). By contrast, the HTRC's Data API uses the manifest of the RDF-workset representation to fetch item content which is computationally analyzed within a secure, copyright respecting, Data Capsule [3]. As a further illustration of flexibility, since a manifest-only RDF workset representation contains no copyrighted content, it can be shared with other scholars, allowing broad reuse and recombination of worksets.

4 SUMMARY

As DLs proliferate and grow in scale, scholars require a way to both span DL boundaries and segment DLs in order to create the coherent collections of digital objects required to support their research. Appropriately modeled worksets provide a means to this end and thereby enhance DL utility. RDF is well-suited for describing worksets because of its inherent extensibility, linking, and accommodation for semantics from multiple ontologies. This latter feature facilitates workset descriptions supporting multiple domains—those operationally required for a DL alongside those specialized to the discipline of the scholar. Furthermore, standards for retrieving RDF workset descriptions from triple stores include features which facilitate on-the-fly transformation of RDF graphs, e.g., SPARQL Construct, and support operations spanning multiple DL triple stores, e.g., SPARQL Federated Query.

REFERENCES

[1] J. S. Downie, T. W. Cole, and B. Plale. 2013. Workset creation for scholarly analysis: Prototyping project. Retrieved Jan. 31, 2018 from: http://hdl.handle.net/2142/99011

[2] J. Jett, T. W. Cole, C. Maden, and J. S. Downie. 2016. The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. Journal of Open Humanities Data 2:e1. DOI: http://doi.org/10.5334/johd.3

[3] J. S. Downie, B. Plale, & T. W. Cole. 2017. Workset creation for scholarly analysis and data capsules: Laying the foundations for secure computation with copyrighted data in the HTRC. Retrieved Jan. 31, 2018 from: http://hdl.handle.net/2142/99010

[4] K. R. Page, S. Bechhofer, G. Fazekas, D. M. Weigl, & T. Wilmering. 2017. Realising a layered digital library: Exploration and analysis of the Live Music Archive through linked data. Proc. ACM/IEEE Joint Conference on Digital Libraries (JCDL). DOI: http://doi.org/10.1109/JCDL.2017.7991563

[5] K. Fenlon, M. Senseney, H. Green, S. Bhattacharyya, C. Willis, and J. S. Downie. 2014. Scholar-built collections: A study of user requirements for research in large-scale digital libraries. Proc. Am. Soc. Info. Sci. Tech., 51: 1–10. DOI: http://doi.org/10.1002/meet.2014.14505101047