**BROWN — Entity-Based**

<0 figures>

**An Entity-Based Approach to Interoperability in the Canadian Writing Research Collaboratory**

Brown, S., Antoniuk, J., Brundin, M., Simpson, J. and Ilovan, M.

A lack of interoperability between digital resources for humanists, whether big or small, plagues our research community. The challenge of the ‘Million Books’ problem identified by Gregory Crane (2006) and the opportunity of examining aspects of humanistic phenomena that appear, as Alan Liu (2012) puts it, ‘only at scale’ are lures that drive the digital humanities community forward in developing new tools and exploring emerging modes of enquiry for large datasets. Yet only a small handful of well-funded projects are actually in a position to investigate such phenomena due in part to restrictions on the use of commercially held large data collections but at least as much on the lack of interoperability and barriers to aggregation between open collections, many of which originate in the digital humanities community. This ‘multivarious isolation’, identified by Rebecca Frost Davis and Quinn Dombrowski (2011), stems from a number of intellectual, institutional, and social factors, but a 2011 Research Information Network report found that the biggest obstacle for researchers in our century was incompatible data structures and disparate and scattered data sources. If digital research in the humanities is to be significantly advanced, this problem must be addressed.

This paper describes the approach to this challenge taken by the Canadian Writing Research Collaboratory (CWRC) (Brown, 2011), a virtual research environment hosting and supporting a range of projects, including ‘seed’ projects with existing data in a wide range of formats. CWRC aims to bridge the gap between the digital humanities community and humanities scholars who lack an understanding of best practices and standards, as well as bridging diverse datasets (Brown, 2011; 2013). We describe the approach to interoperability built into the Collaboratory’s systems and argue that such an approach based on linked data principles supports scaling up beyond the Collaboratory to form a basis for interoperability across disparate and distributed projects with overlapping content. The resulting interoperability is in some senses quite rudimentary, but at the same time constitutes a step towards a more robust ecology of interlinked resources.

Core to the approach of developing the Collaboratory such that it facilitates interoperability is the entity-based nature of the architecture, as expressed in entity records, entity schemas, and an entity management system. The approach is similar to that employed by the Humanities Networked Infrastructure (HuNI) project, which brings together a large collection of diverse datasets from various Australian providers (Burrows, 2013), and that of the New Zealand Electronic Text Collection’s Entity Authority Tool Set (EATS) (Norrish, 2007). Entity records in the context of CWRC have two purposes: to enable Semantic Web–type linked open data processes and to facilitate traditional authority control and information retrieval in a digital repository. In order to achieve these two purposes, CWRC entities are structured as elemental, atomic units of information, and, specifically, they are modeled as proper noun–named entities in four classes: persons, organizations, places, and titles (works). The core characteristics that are common to all four types of entities are an indication of the preferred or authoritative name, variant names if any, and important administrative data such as record creation date, contributing project or projects, and an access condition statement. Entity schemas are used to model and govern the creation of entity records to ensure consistency among the entities; this is achieved by defining required elements and attributes, content types associated with each element, and the nesting hierarchy of elements. The CWRC XML entity schemas consist of a custom CWRC omnibus entity schema that models the person, organization, and place entities, and the Library of Congress’ MODS schema, which models the title entities; the omnibus entity schema was influenced by concepts from several existing schema standards, including EAC-CPF, FOAF, GeoNames, ISAAR (CPF), MADS, and MODS. Following the codified approach of authority records, the CWRC entity types parallel the Library of Congress person, organization, geographic name, and title name authority files, and the entities’ core structure contains the standard three components of an authority record: authoritative form, cross-variant forms, and administrative notes. These entity records provide the basis for a linked data entity within the Resource Description Framework (RDF), with a dereferencable uniform resource identifier (URI) and associated RDF triples that translate the information about that entity from the schema, as well as some of the information about that entity from elsewhere in CWRC, into an open Semantic Web dataset for use within and beyond the Collaboratory (Simpson and Brown, 2013).

With the exception of the use of MODS records for bibliographical entities, the CWRC entity schema is deliberately minimalist, more so than the EATS data model for instance, serving as an authority list that brokers between different data formats used by various projects about entities, rather than requiring all projects within CWRC to adopt the same standard for recording detailed information about those entities. For instance, the Lesbian and Gay Liberation in Canada (2015) project led by Constance Crompton and Michelle Schwartz is recording prosopographical information using the Text Encoding Initiative, while the Orlando Project (2015) has a rich bespoke XML schema for encoding prosopographical information, while bibliographical or editing projects may use MODS or METS or EAD. The CWRC approach allows the same person as described by documents using a range of schemas to refer to a common, minimalist entity, while permitting divergent representational practices and individual project autonomy. It also allows us to preserve and make somewhat interoperable existing digital datasets that we can only partially remediate, and to bring together XML materials with materials in other media. The CWRC entities thus facilitate the collocation of a wide range of resources that reference that entity. Rather than acting as receptacles for prosopographical information, they act as authority records across CWRC, but connect using linked data principles to other authoritative and often more detailed records from individual projects, facilitating information retrieval and aggregation. The approach facilitates linking into the Semantic Web more generally; indeed CWRC entities beyond those harvested from our seed projects will only be minted if users cannot locate a linked data identifier for that entity amongst trusted linked data entity collections such as the Virtual International Authority File.

This pragmatic approach to entities is complementary with, and indeed promotes, a number of existing standards, bridging datasets through a set of linked data that will support interoperability and leverage that bridging into conformity with Semantic Web standards. As Joris van Zundert has argued, the idea of uniformly behaving humanist scholars is a pipe dream; furthermore, he argues, ‘any meaningful humanities research . . . cannot be entirely governed by standards’ (2012, 173). CWRC’s strategy focuses on building interoperability from how scholars with a range of technological expertise and different degrees of willingness to engage with complex standards actually work, not how we think they should work in the best of all possible worlds.

The paper will demonstrate the user experience that stitches together the entity architecture with other components of the Collaboratory to enable users to incorporate relatively technical tasks into their workflow. The entity management system provides a user interface to create and edit entities, as well as mechanisms to perform entity extraction on source documents and global merging of these extracted entities into the CWRC common pool of core named entities. The centerpiece of CWRC’s interoperability strategy is a web-based XML and RDF Editor called CWRC-Writer (Brown et al., 2014; Rockwell et al., 2012). It aims to integrate the activities of correcting, writing, formatting, and editing a text with managing references to entities so as to minimize the additional work required for disambiguation, merging, and management. Tagging an entity within CWRC-Writer creates XML tags such as persname, orgname, or title, as well as equivalent RDF annotations employing the Open Annotation data model (2013), a practice that makes the project importantly about links rather than simple tagging (Hendler, 2011). Having the user provide the entity linking up front means that the act is informed by their knowledge of the material, and integrating it into the writing interface minimizes the additional time required to enhance the information. The immediate payoff to users will be the creation of links between their materials. This section of the paper demonstrates how entity tagging works within CWRC-Writer and shows our progress towards the cross-project aggregation of materials that entity tagging will support.

To further support the essential role of entities within the Collaboratory, we are exploring implementing, where possible, supplementary technologies and practices. These include support for the construction of a semiautomated named entity recognition (NER) tool (Barbosa et al., 2012), developing Semantic Web capability through ontology development, the exposure of data through a publicly accessibly triplestore (Brown and Simpson, 2013), and pursuing synergies with other linked data projects in the humanities.

The Canadian Writing Research Collaboratory employs entities to showcase the benefits of interoperability while providing immediate benefits to researchers. Our conclusion will touch on the ways in which this approach provides a foundation on which to build more advanced services in the future.

**References**

**Barbosa, D., Brown, S., Home, M. and Moroz, A.** (2012). On Semi-Automatic Tools for Annotating Scholarly Material. *Canadian Society for Digital Humanities Annual Meeting at the Congress of the Humanities and Social Sciences*, University of Waterloo / Wilfrid Laurier University, Waterloo, Ontario, 26 May–2 June 2012.

**Brown, S.** (2011). Don’t Mind the Gap: Evolving Digital Modes of Scholarly Production across the Digital-Humanities Divide. In Coleman, D. and KamboureliS. (eds), *Retooling the Humanities: The Culture of Research in Canadian Universities*. Edmonton: University of Alberta Press, pp. 203–31, http://hdl.handle.net/10402/era.25382 (open access preprint version).

**Brown, S.** (2014). Scaling Up Collaboration Online: Towards a Collaboratory for Research on Canadian Writing. *International Journal of Canadian Studies,* **48**: 233–51.

**Brown, S., Brundin, M., Chartrand, J., Knechtel, R., MacDonald, A., Rockwell, G. and Sellmer, M.** (2014). The CWRC-Writer Bridge: From Coder to Writer, XML to RDF, DH to Mainstream. *Digital Humanities Conference Abstracts 2014*, University of Lausanne, Lausanne, Switzerland, 7–12 July 2014, http://dharchive.org/paper/DH2014/Poster-782.xml.

**Brown, S., and Simpson, J.** (2013). The Curious Identity of Michael Field and Its Implications for Humanities Research with the Semantic Web. *IEEE, Big Data 2013, Humanities and Big Data Workshop*, Santa Clara, CA, 6–9 October 2013, pp. 77–85, doi:10.1109/BigData.2013.6691674.

**Burrows, T.** (2013). A Data-Centred ‘Virtual Laboratory’ for the Humanities: Designing the Australian Humanities Networked Infrastructure (HuNI) Service. *Literary and Linguistic Computing,* **28**(4): 576–81.

**Crane, G.** (2006). What Do You Do with a Million Books? *D-Lib Magazine,* **12**(3): 1.

**Davis, R. F. and Dombrowski, Q.** (2011). Divided and Conquered: How Multivarious Isolation Is Suppressing Digital Humanities Scholarship. *National Institute for Technology in Liberal Education*, http://www.nitle.org/live/files/36-divided-and-conquered.

**Hendler, J.** (2011). Why the Semantic Web Will Never Work. *8th Extended Semantic Web Conference (ESWC)*,Heraklion, Crete, 29–30 May 2011, http://videolectures.net/eswc2011\_hendler\_work/.

**Lesbian and Gay Liberation in Canada.** (2015). http://www.lglc.ca/.

**Liu, A.** (2012). The State of the Digital Humanities: A Report and a Critique. *Arts and Humanities in Higher Education,* **11(**1): 1–34, http://ahh.sagepub.com/content/early/2011/11/30/1474022211427364.full.pdf+html.

**Norrish, J.** (2007). EATS: An Entity Authority Tool Set. *Australia New Zealand Digital Encyclopedias Group Meeting*, Sydney, 7–8 December 2007.

**Open Annotation Data Model.** (2013). http://www.openannotation.org/spec/core/.

**Orlando Project.** (2015). http://www.ualberta.ca/orlando.

**Research Information Network.** (2011). *Reinventing Research? Information Practices* *in the Humanities.* Research Information Network Report, April, http://www.rin.ac.uk/system/files/attachments/Humanities\_Case\_Studies\_for\_screen\_2\_0.pdf.

**Rockwell, G., Brown, S., Chartrand, J., and Hesemeier, S.** (2012). CWRC-Writer: An In-Browser XML Editor. *Digital Humanities Conference Abstracts 2012*, University of Hamburg, Hamburg, Germany, 16–22 July 2012,<http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/cwrc-writer-an-in-browser-xml-editor/>.

**Simpson, J. and Brown, S.** (2013). From XML to RDF in the Orlando Project. *International Conference on Culture and Computing 2013*, Kyoto, Japan, 16–18 September 2013, pp. 194–95, doi:10.1109/CultureComputing.2013.61.

**Zundert, J. van.** (2012). ‘If You Build It, Will We Come?’ Large-Scale Digital Infrastructures as a Dead End for Digital Humanities.’ *Historical Social Research–Historische Sozialforschung,* **37**(3): 165–86.