Topic Maps and Entity Authority Records: an Effective Cyber Infrastructure for Digital Humanities

Jamie Norrish Alison Stevenson

20 February 2008

The implicit connections and cross-references between and within texts, which occur in all print collections, can be made explicit in a collection of electronic texts. Correctly encoded and exposed they create a framework to support resource discovery and navigation by following links between topics. This framework provides opportunities to visualise dense points of interconnection and, deployed across otherwise separate collections, can reveal unforeseen networks and associations. Thus approached, the creation and online delivery of digital texts moves from a digital library model with its goal as the provision of access, to a digital humanities model directed towards the innovative use of information technologies to derive new knowledge from our cultural inheritance.

Using this approach the New Zealand Electronic Text Centre (NZETC) has developed a delivery system for its collection of over 2500 New Zealand and Pacific Island texts using TEI XML, the ISO Topic Map technology[1] and innovative entity authority management. Like a simple back-of-book index but on a much grander scale, a topic map aggregates information to provide binding points from which everything that is known about a given subject can be reached. The ontology which structures the relationships between different types of topics is based on the CIDOC Conceptual Reference Model[2] and can therefore accommodate a wide range of types. To date the NZETC Topic Map has included only those topics and relationships which are simple, verifiable and object based. Topics currently represent authors and publishers, texts and images, as well as people and places mentioned or depicted in those texts and images. This has proved successful in presenting the collection as a resource for research, but work is now underway to expand the structured mark-up embedded in texts to encode scholarly thinking about a set of resources. Topic-based navigable linkages between texts will include 'allusions' and 'influence' (both of one text upon another and of an abstract idea upon a corpus, text, or fragment of text).[3]

[1] For further information see Conal Tuhoy's "Topic Maps and TEI — Using Topic Maps as a Tool for Presenting TEI Documents" (2006) `http://hdl.handle.net/10063/160`.

[2] The CIDOC CRM is an ISO standard (ISO 21127:2006) which provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. For more information see `http://cidoc.ics.forth.gr/`.

[3] The initial project work is being undertaken in collaboration with Dr Brian Opie from Victoria University of Wellington and is centred around the work and influences of William Golder, the author of the first volume of poetry printed and published in New Zealand.

Importantly, the topic map extends beyond the NZETC collection to incorporate relevant external resources which expose structured metadata about entities in their collection (see figure 1 on page 3).

Cross-collection linkages are particularly valuable where they reveal interdisciplinary connections which can provide fertile ground for analysis. For example the National Library of New Zealand hosts a full text archive of the Transactions and Proceedings of the Royal Society containing New Zealand science writing 1868–1961. By linking people topics in the NZETC collection to articles authored in the Royal Society collection it is possible to discern an interesting overlap between the 19th century community of New Zealand Pakeha artists and early colonial geologists and botanists.

In order to achieve this interlinking, between collections, and across institutional and disciplinary boundaries, every topic must be uniquely and correctly identified. In a large, full text collection the same name may refer to multiple entities,[4] while a single entity may be known by many names.[5] When working across collections it is necessary to be able to confidently identify an individual in a variety of contexts. Authority control is consequently of the utmost importance in preventing confusion and chaos.

The library world has of course long worked with authority control systems, but the model underlying most such systems is inadequate for a digital world. Often the identifier for an entity is neither persistent nor unique, and a single name or form of a name is unnecessarily privileged (indeed, stands in *as* the entity itself). In order to accommodate our goals for the site, the NZETC created the Entity Authority Tool Set (EATS),[6] an authority control system that provides unique, persistent, sharable[7] identifiers for any sort of entity. The system has two particular benefits in regards to the needs of digital humanities researchers for what the ACLS described as a robust cyber infrastructure.[8]

Firstly, EATS enables automatic processing of names within textual material. When dealing with a large collection, resource constraints typically do not permit manual processing — for example, marking up every name with a pointer to the correct record in the authority list, or simply recognising text strings as names to begin with. To make this process at least semi-automated, EATS stores names broken down (as much as possible) into component parts. By keeping track of language and script information associated with the names, the system is able to use multiple sets of rules to know how to properly glue these parts together into valid name forms. So, for example, William Herbert Ellery Gilbert might be referred to in a text by "William Gilbert", "W. H. E. Gilbert", "Gilbert, Wm.", or a number of other forms; all of these can be automatically recognised due to the language and script rules associated with the system. Similarly Chiang Kai-shek, being a Chinese name, should be presented

[4] For example, the name Te Heuheu is used in a number of texts to refer to multiple people who have it as part of their full name.

[5] For example the author Iris Guiver Wilkinson wrote under the pseudonym Robin Hyde.

[6] For more analysis of the weakness of current Library standards for authority control and for more detail information on EATS see Jamie Norrish's "EATS: an entity authority tool set" (2007) at `http://researcharchive.vuw.ac.nz/handle/10063/220`.

[7] Being a web-based application the identifiers are also dereferencable (ie resolve a web resource about the entity) and therefore can be used as a resource by any web project.

[8] "Our Cultural Commonwealth: The final report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences" (2006) `http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf`.

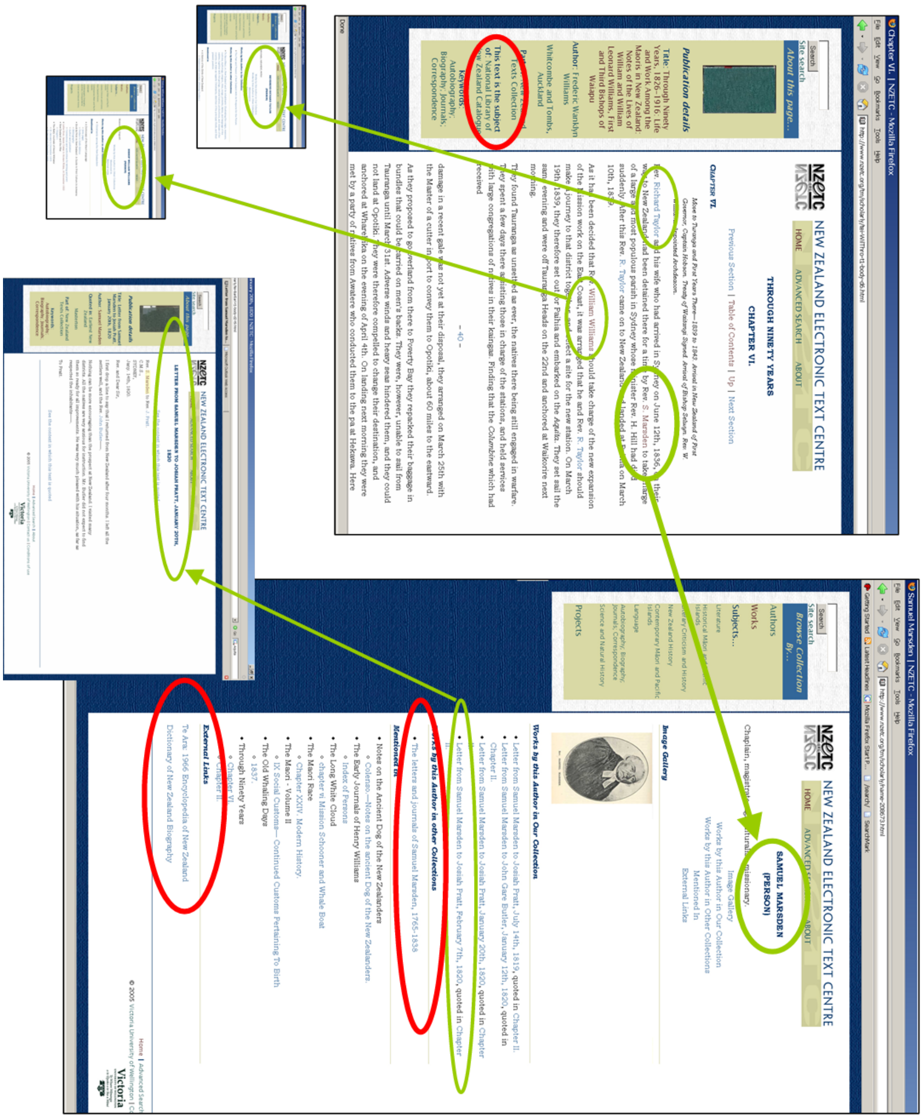

Figure 1: A mention of Samuel Marsden in a given text is linked to a topic page for Marsden which in turn provides links to other texts which mention him, external resources about him and to the full text of works that he has authored both in the NZETC collection and in other online collections entirely separate from the NZETC.

with the family name first, and, when written in Chinese script, without a space between the name parts (蒋介石).

The ability to identify entities within plain text and add structured, machine-readable mark-up contributes to the growth of electronic text corpora suitable for the types of computational analysis offered by projects such as the MONK environment.[9] This is, however, distinct from the problem of identifying substrings within a text that might be names, but that are not found within EATS. This problem, though significant, does not fall within the main scope of the EATS system.[10] Similarly, disambiguating multiple matches for the same name is generally best left to the determination of a human being: even date matches are too often problematic.[11]

Secondly, the system is built around the need to allow for an entity to carry sometimes conflicting, or merely different, information from multiple sources, and to reference those sources.[12] Having information from multiple sources aids in the process of disambiguating entities with the same names; just as important is being able to link out to other relevant resources. For example, our topic page for William Colenso links not only to works in the NZETC collection, but also to works in other collections, where the information on those other collections is part of the EATS record.

It is, however, barely sufficient to link in this way directly from one project to another. EATS, being a web application, can itself be exposed to the net and act as a central hub for information and resources pertaining to the entities within the system. Since all properties of an entity are made as assertions by an organisation, EATS allows multiple such organisations to use and modify records without touching anyone else's data; adding data harvesting to the mix allows for centralisation of information (and, particularly, pointers to further information) without requiring much organisational centralisation.

One benefit of this approach is handling entities about which there is substantial difference of view. With topics derived from research (such as ideas and events) there are likely to be differences of opinion as to both the identification of entities and the relationships between them. For example one organisation may see one event where another sees two. To be able to model this as three entities, with relationships between them asserted by the organisations, a potentially confusing situation becomes clear, without any group having to give up its own view of the world. The EATS system can achieve this because all information about an entity is in the form of a property assertion made by a particular authority in a particular record (see figure 2 on page 5).

The technologies developed and deployed by the NZETC including EATS are all based on open standards. The tools and frameworks that have been created are designed to provide durable resources to meet the needs of the academic and wider community in that they promote interlinking between digital collections and projects and are themselves interoperable with other standards-based pro-

[9] Metadata Offer New Knowledge `http://www.monkproject.org/`.

[10] EATS can be provided with name data on which to make various judgements (such as non-obvious abbreviations like Wm for William), and it would be trivial to get a list of individual parts of names from the system, for identification purposes, but there is no code for actually performing this process.

[11] That said, the NZETC has had some success with automated filtering of duplicate matches into different categories, based on name and date similarity (not equality); publication dates provide a useful cut-off point for matching.

[12] A source may be a primary text or an institution's authority record.

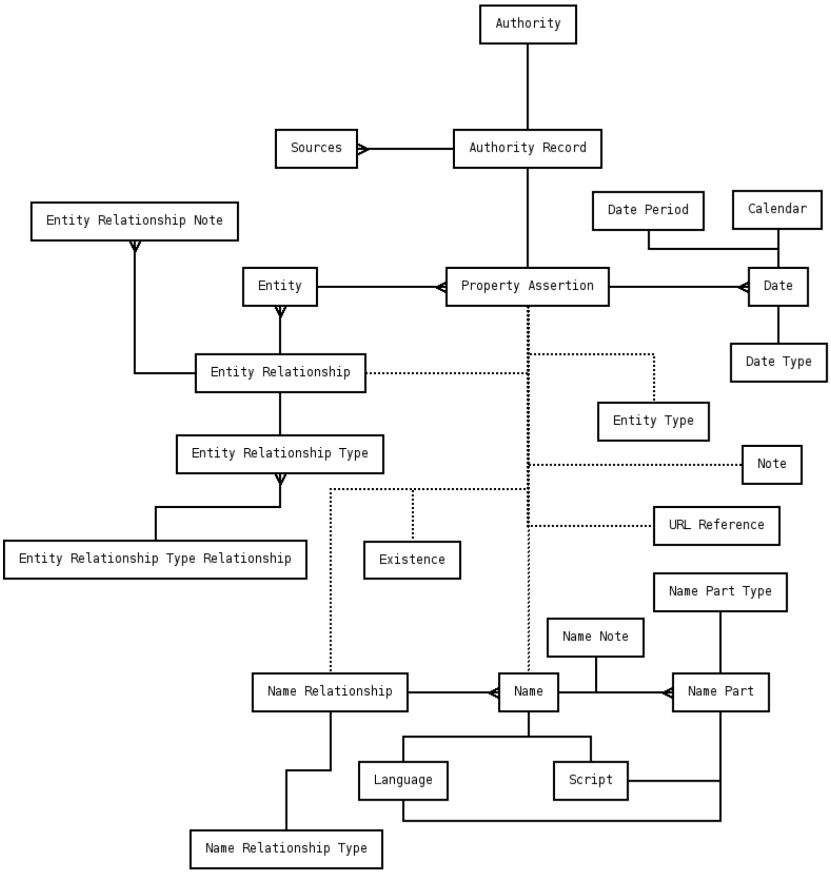

Figure 2: The EATS objects and basic relationships

grams and applications including web-based references tools, eResearch virtual spaces and institutional repositories.

Only once both cultural heritage institutions and digital humanities projects adopt suitable entity identifiers and participate in a shared mapping system such as EATS, can there exist unambiguous discovery of both individual resources and connections between them. The wider adoption of this type of entity authority system will contribute substantially to the creation of the robust cyber infrastructure that will, in the words of the ACLS “allow digital scholarship to be cumulative, collaborative, and synergistic.”