AMBIENT FINDABILITY AND STRUCTURED SERENDIPITY: ENHANCED RESOURCE DISCOVERY FOR FULL TEXT COLLECTIONS

ALISON STEVENSON, CONAL TUOHY, JAMIE NORRISH
New Zealand Electronic Text Centre
New Zealand
alison.stevenson@vuw.ac.nz
conal.tuhoy@vuw.ac.nz
jamie.norrish@vuw.ac.nz

University Libraries manage increasingly large collections of full text digital resources. These might be repositories of born digital research outputs, e-reserves collections or online libraries of material digitised to provide open access to significant texts. Whatever the content of the material, the structured data of full text resources can be exploited to enhance research discovery. The implicit connections and cross-references between books and papers, 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 both within and between texts by following links between topics. Using this approach the New Zealand Electronic Text Centre (NZETC) at Victoria University of Wellington has developed a delivery system for its growing online digital library using the ISO Topic Map technology. Like a simple back-of-book index or a library classification system, a topic map aggregates information to provide binding points from which everything that is known about a given subject can be reached. Topics in the NZETC digital library represent authors and publishers, texts, and images, as well as people and places mentioned or depicted in those texts and images. Importantly, the Topic Map extends beyond the NZETC collection to incorporate relevant external resources which expose structured metadata about their collection. Innovative entity authority records management enables, for example, the topic page for William Colenso to automatically provide access not only to the full text of his works in the NZETC collection but out to another book-length work in the Auckland University's "Early NZ Books Collection" and to several essays in the National Library's archive of the Royal Society Journals. It also enables links to externally provided services providing information on Library holdings of print copies of the text. The NZETC system is based on international standards for the representation and interchange of knowledge including TEI XML, XTM, XSL and the CIDOC CRM. The NZETC collection currently includes over 2500 texts covering 110,000 topics.

An ever larger proportion of the textual resources used by university staff and students in their research and learning are electronic and a significant number of those resources are being produced by the universities themselves. In addition to the collections of subscription datasets, ejournals and ebooks that are purchased, university libraries in New Zealand are actively building substantial, freely accessible, full text collections of both historical material and new research. Each of the eight universities and seven of the polytechnics now maintain institutional repositories providing public access to the full text of research outputs. Three of the universities have undertaken the creation of online libraries of digitised documentary heritage material. At Victoria University of Wellington the New Zealand Electronic Text Centre provides access to over 2500 texts. At the University of Auckland Library there is the New Zealand Electronic Poetry Centre, the "Early New Zealand Books" collection and a partial archive of the Journal of the Polynesian Society. The University of Waikato has produced an online collection of "Niupepa Maori" and another of selected New Zealand content from the Illustrated London News. All of this material, including TIFF or JPEG[1] page images and image or text-based PDFs[2], could at a stretch be

[1] TIFF and JPEG are both commonly used image file formats. For more information see http://www.w3.org/Graphics/JPEG/itu-t81.pdf (JPEG) and http://partners.adobe.com/public/developer/tiff/index.html (TIFF)

described as “full-text” in the sense that a reader has electronic read access to all of the words, but only some resources can be called “well-structured full-text”. A well-structured digital resource employs some form of standards-based, machine-readable, text mark-up to describe the structure of the text. Although some of the general comments made in this paper about the enhanced resource discovery opportunities offered by full text digital collections apply equally well to TIFF/JPEG/PDF formatted resources as to well-structured material, it is to the latter that is referred when the term ‘digital text’ is employed from here on in and which is the focus of discussion.

This paper does not seek to argue that such well-structured, full-text digitised editions of significant parts of New Zealand’s documentary heritage can for all purposes replace the original print edition or manuscript. As Esther, a character in Orhan Pamuk’s “My Name is Red” (2002), observes

> *“A letter doesn’t communicate by words alone. A letter, just like a book, can be read by smelling it, touching it and fondling it. Thereby, intelligent folk will say, ‘Go on then, and read what the letter tells you!’ whereas the dull witted will say, ‘Go on then, read what he’s written!”*

For academic researchers, particularly those in the humanities, there are occasions when the paper, ink and glue of the physical artefact can convey information not captured in a digitised version. To acknowledge as much does not detract from an argument that there are a number of characteristics of digital text that offer prospects for enhanced resource discovery. It is those qualities and opportunities that this paper explores. It is in respect of one key characteristic of digital text, hypertextuality, that the possibilities of ambient findability (wayfinding, navigation and retrieval) and structured serendipity (the value of unsought finding) will be discussed. Prior to that, four other characteristics will be briefly examined: capacity, accessibility, flexibility, and manipulability.[3]

As streams of binary data which can be compressed and stored on devices which get cheaper each year, very large collections of digital text can be managed by libraries which could not accommodate a similar number of titles in physical form. At a very simple level the increased capacity of University Libraries to provide access to a greater number of titles enhances resource discovery since the large collection are more likely to contain resources relevant to the researcher's interest[4].

In similarly simplistic terms the fact that, once created and made available through an open access online collection, digital texts can be accessed around the world, twenty-fours hours a day, three hundred and sixty five days a year, means that opportunities for users to find these resources are increased when compared to their ability to find and use resources on metres of shelving in a specific geographic location which probably closes at 5pm. Figure 1 shows the hourly usage statistics from July 2007 for the New Zealand Electronic Text Centre. The level of usage is fairly constant – the number of hits between midnight and 1am (New Zealand time) is not significantly different to that between midday and 1pm. This is because the texts are globally accessible – users in Europe and North America search and browse and read while the South Pacific sleeps.

[2] PDF is a commonly use document format. For more information see http://www.adobe.com/devnet/pdf/pdf_reference.html

[3] These characteristics are drawn from Daniel J. Cohen and Roy Rosenzweig’s “Digital History” http://chnm.gmu.edu/digitalhistory (accessed 30th November 2007)

[4] Note however that large collections can have their own resource discovery issues directly related to their size. Without advanced browsing, search and result filtering options users can be overwhelmed by sheer volume of content.

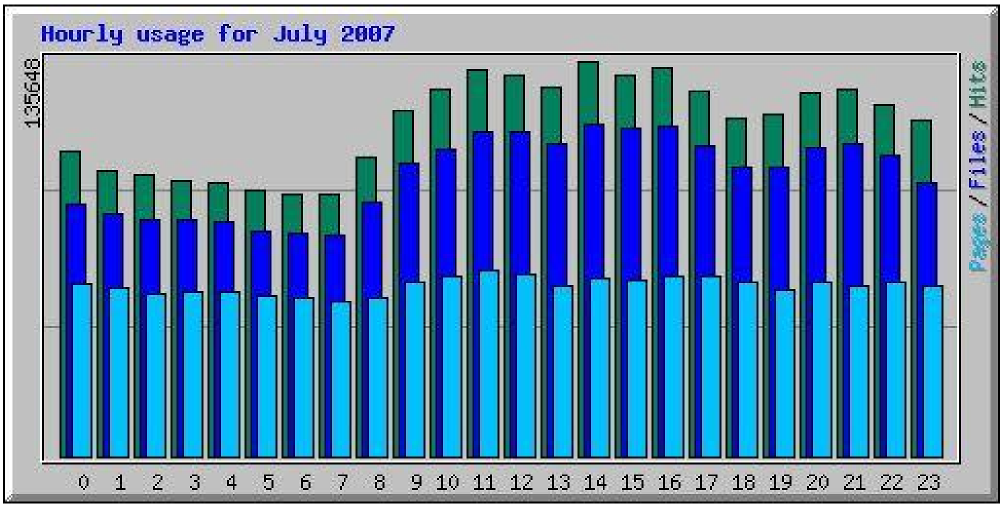


Figure 1 Hourly Usage Statistics, July 2007, for the NZETC Collection

The flexibility of digital text allows it to be presented in a variety of forms: as HTML for web delivery for example; as PDF, RTF[5], plain text[6], or eBook format for download and offline access; as DAISY book with synthetic audio for print disabled readers. Each format can be automatically derived from a single, well-structured source such as TEI XML[7]. Thus the digital texts can be delivered in the forms most useful to a range of users and placed in a range of environments for discovery e.g. the Royal New Zealand Foundation of the Blind library of talking books (DAISY texts[8]) or Second Life InfoIsland (eBook format not yet defined[9]).

To shift formats and place resources where the users are (rather than expecting the users to come to a particular place to access the material) is a step towards "ambient findability". It is phrase coined and a concept much discussed by Peter Morville:

> *"Ambient findability describes a world at the crossroads of ubiquitous computing and the Internet in which we can find anyone or anything from anywhere at anytime. It's not necessarily a goal, and we'll never quite arrive, but we're sure as heck headed in the right direction."*[10]

Although Morville's vision is substantially broader than web-based discovery of texts and includes ambient interfaces, sensors and nanotechnology, his definitions of findability are highly relevant to any discussion of resource discovery.

> *"The quality of being locatable or navigable.*
> *The degree to which a particular object is easy to discover or locate.*
> *The degree to which a system or environment supports navigation and retrieval."*
> (*Ambient Findability* O'Reilly Media, 2005, by Peter Morville)

[5] Rich Text Format – a Microsoft document file format for document exchange.

[6] Plain text is unformatted. Character encoding is commonly Unicode.

[7] TEI stands for Text Encoding Initiative. The TEI *Guidelines for Electronic Text Encoding and Interchange* define and document a markup language for representing the structural, renditional, and conceptual features of texts. These guidelines are expressed as a modular, extensible XML schema – TEI XML.

[8] DAISY denotes the Digital Accessible Information System. For more information see http://www.daisy.org/

[9] Although often oversized eBooks do exists in Second Life and can be made by members it is not yet possible to "borrow" titles directly from InfoIsland Library.

[10] http://www.boxesandarrows.com/view/ambient_findability_talking_with_peter_morville [accessed 14th November 2007]

The manipulability of well structured digital texts, the fact that they can be used as the input to computational processes, is another characteristic which offers several opportunities to increase findability. The ability to programmatically to extract parts of the text as metadata and reconfigure for use in other systems such as reference management tools, archival finding aids, metadata aggregators, or library catalogues[11] means that a single resource can be discovered from any number of environments. The entire text forms the basis for search engine indexes providing a potent discovery tool particularly for collections which do not have carefully collated indexes.

While powerful, full-text searching and the automatic creation of metadata records do not always take advantage of the richness of data available to computational processes. 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 through the creation of hypertext links. Correctly encoded and exposed the cross-references 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. The system is able to provide a digital variation, perhaps advancement, of the sort of structured serendipity that users of print collections enjoy when browsing the shelves.

Using this approach the New Zealand Electronic Text Centre (NZETC) has developed a delivery system for its collection of New Zealand and Pacific Island texts using TEI XML, the ISO Topic Map technology[12] 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 NZETC delivery framework is a dynamically generated semantic framework – a metadata repository implemented using the Topic Maps instead of the more usual implementation based directly on a relational database. The topic map metadata repository provides the system with an unusually flexible and open-ended conceptual structure. This has a number of benefits, including greatly simplifying the integration of disparate information systems and facilitating the presentation of contextually rich web pages. Users are able to move around the resources on the site tracking topics of interest rather than merely browsing the material linearly or through text searching. In a topic map, web based resources are grouped around items called "topics", each of which represents some subject of interest. In the NZETC topic map, the topics represent books, chapters, and illustrations, and also people and places mentioned in those books. 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).

Topics in a topic map are linked together with hyperlinks called "associations". There can be different types of association in a topic map, representing the different kinds of relationship in the real world. For instance, in the NZETC topic map, the topic which represents a particular person may be linked to a topic which represents a chapter of a book which mentions that person. This association would be labelled to indicate that it represents a "mention". Similarly, the same person's topic might be linked to a particular photograph topic, via a "depiction" association. This identification and codification of topics and associations is essentially the act of creating an ontology. Modelling domain relationships requires a sophisticated analysis of real work entities, a difficult and time consuming task. We have therefore taken advantage of the ten year effort by the International Committee for Museum Documentation (CIDOC) group to create a high level

[11] Through the automatic generation of MARC records or EAD data from well-structured TEI XML for example. See Mimmo, Jones and Crane "Generating Analytical Catalog Records from Well Structured Digital Texts" (JCDL'05)

[12] Topic Maps is an ISO standard for the representation and interchange of knowledge, with an emphasis on the findability of information. The standard is formally known as ISO/IEC 13250:2003.

ontology known as the CIDOC CRM[13]. This ontology was designed to enable information integration for cultural heritage data and their correlation with library and archive information. The NZETC has based the semantics of the topic map ontology on the event based model of the CIDOC CRM as illustrated below.

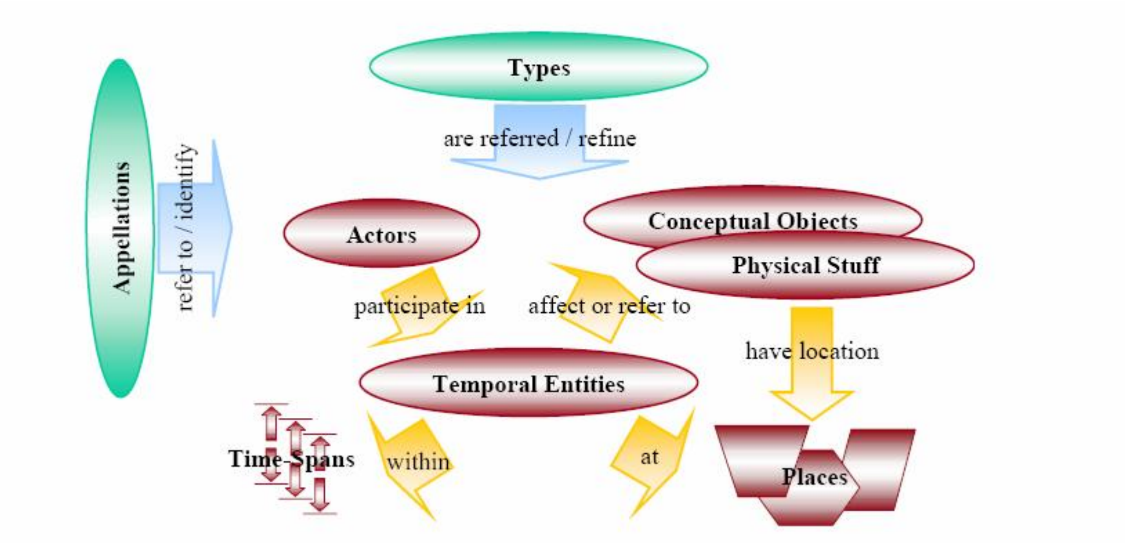


Figure 2A qualitative metaschema of the CIDOC CRM taken from Martin Doerr "The CIDOC CRM – An Ontological Approach to Semantic Interoperability of metadata AI Magazine", Volume 24, Number 3 (2003)

When populated with data harvested from the texts the topic map provides a graph of the interconnections between people, places and texts (Figure 2) which can then be rendered as hyperlinks creating a browsable navigations framework (Figure 3)

[13] Since 2006 this has been a recognised ISO standard (ISO 21127:2006). CIDOC CRM documentation is available at http://cidoc.ics.forth.gr/

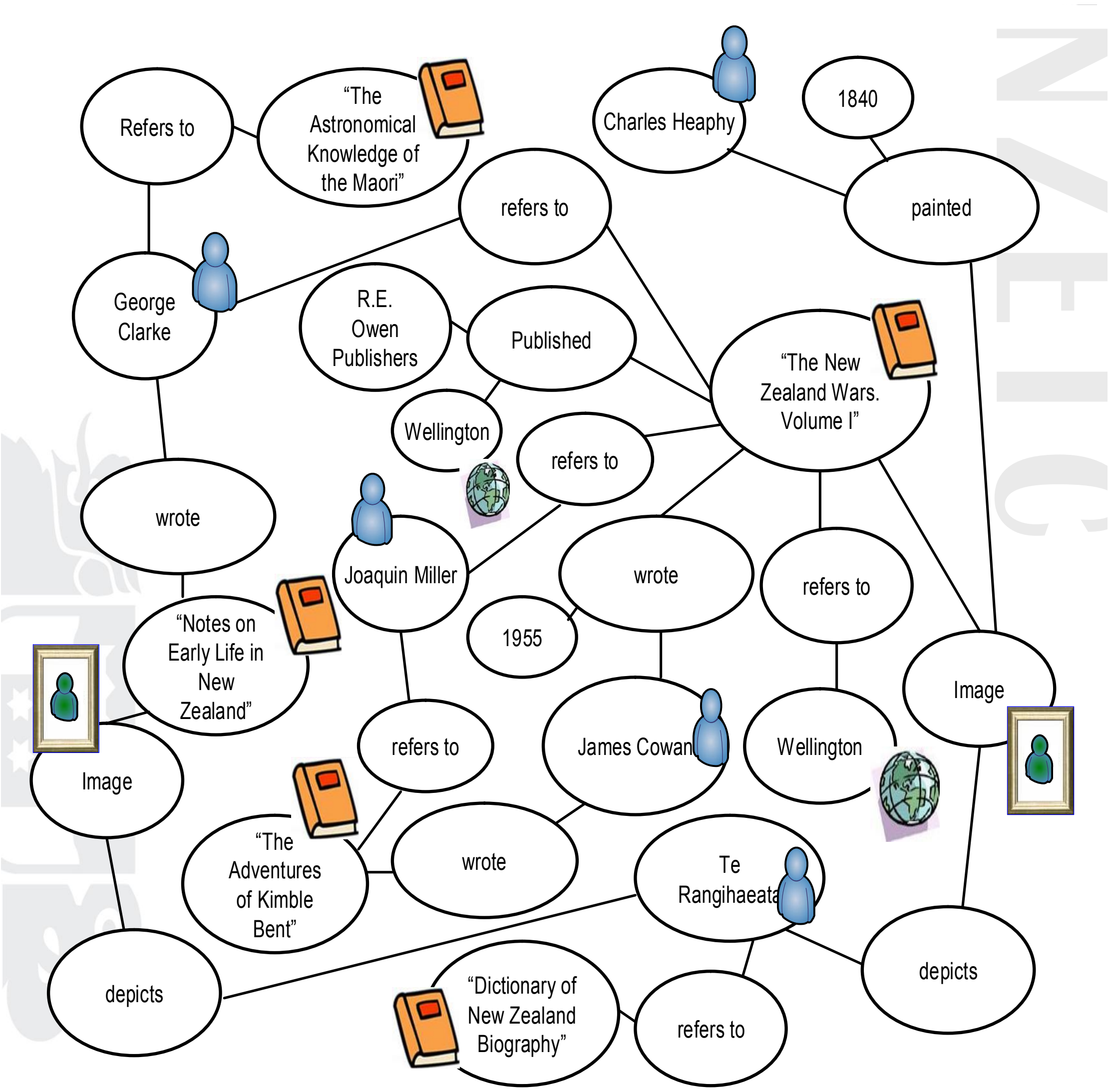
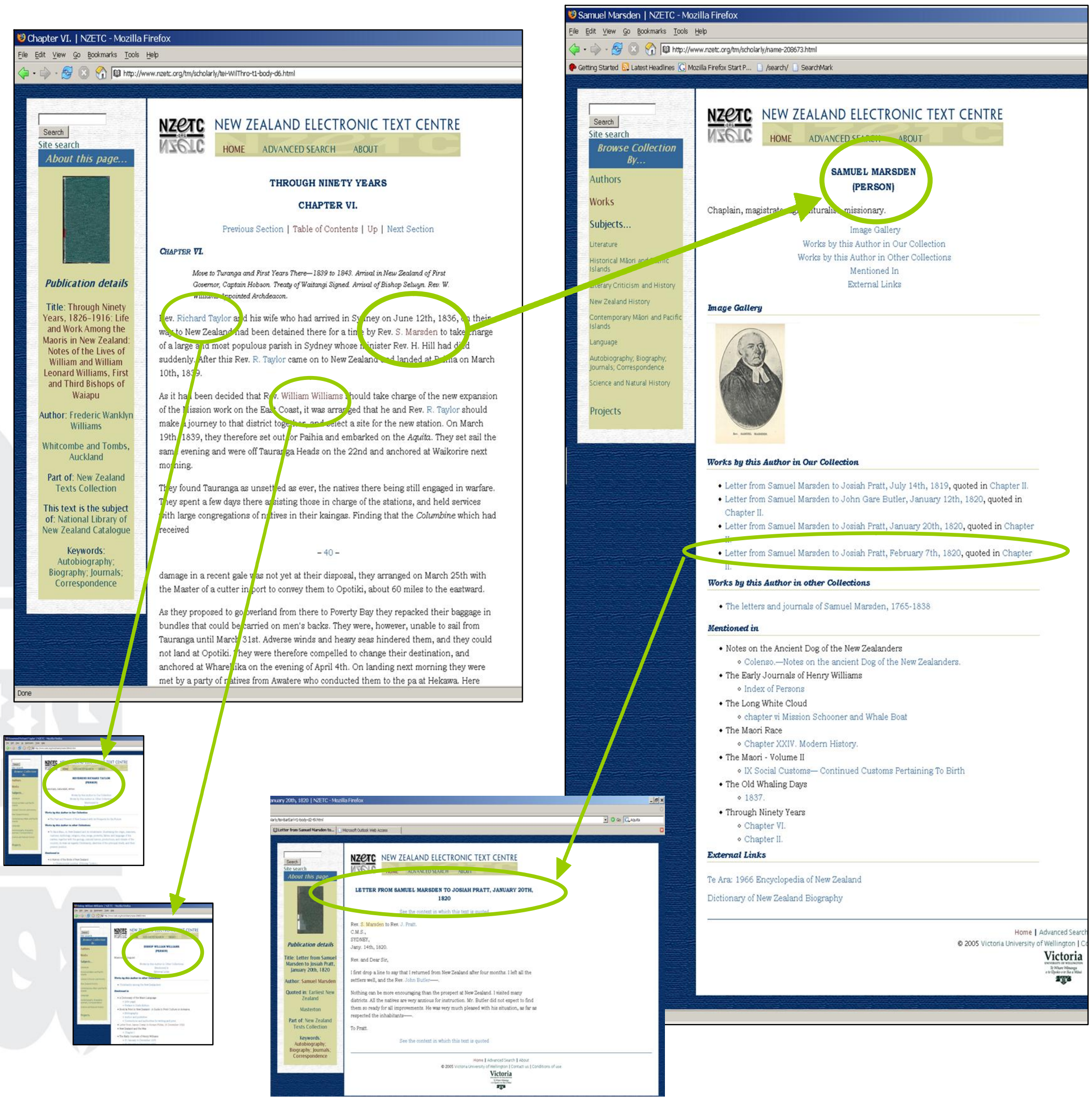


Figure 3 Example set of concrete topics and relationships from NZETC collection



Chapter VI. | NZETC - Mozilla Firefox

http://www.nzetc.org/tm/scholarly/tei-WilThro-t1-body-d6.html

NEW ZEALAND ELECTRONIC TEXT CENTRE

THROUGH NINETY YEARS

CHAPTER VI.

Previous Section | Table of Contents | Up | Next Section

CHAPTER VI.

Move to Turanga and First Years There—1839 to 1843. Arrival in New Zealand of First Governor, Captain Hobson. Treaty of Waitangi Signed. Arrival of Bishop Selwyn. Rev. W. Williams Appointed Archdeacon.

Rev. Richard Taylor and his wife who had arrived in Sydney on June 12th, 1836, on their way to New Zealand had been detained there for a time by Rev. S. Marsden to take charge of a large and most populous parish in Sydney whose minister Rev. H. Hill had died suddenly. After this Rev. R. Taylor came on to New Zealand and landed at Paihia on March 10th, 1839.

As it had been decided that Rev. William Williams should take charge of the new expansion of the Mission work on the East Coast, it was arranged that he and Rev. R. Taylor should make a journey to that district together and select a site for the new station. On March 19th, 1839, they therefore set out for Paihia and embarked on the *Aquita*. They set sail the same evening and were off Tauranga Heads on the 22nd and anchored at Waikorire next morning.

They found Tauranga as unsettled as ever, the natives there being still engaged in warfare. They spent a few days there assisting those in charge of the stations, and held services with large congregations of natives in their kaingas. Finding that the *Columbine* which had received

– 40 –

damage in a recent gale was not yet at their disposal, they arranged on March 25th with the Master of a cutter in port to convey them to Opotiki, about 60 miles to the eastward.

As they proposed to go overland from there to Poverty Bay they repacked their baggage in bundles that could be carried on men's backs. They were, however, unable to sail from Tauranga until March 31st. Adverse winds and heavy seas hindered them, and they could not land at Opotiki. They were therefore compelled to change their destination, and anchored at Whareikika on the evening of April 4th. On landing next morning they were met by a party of natives from Awatere who conducted them to the pa at Hekawa. Here

Publication details

Title: Through Ninety Years, 1826-1916: Life and Work Among the Maoris in New Zealand: Notes of the Lives of William and William Leonard Williams, First and Third Bishops of Waiapu

Author: Frederic Wanklyn Williams

Whitcombe and Tombs, Auckland

Part of: New Zealand Texts Collection

This text is the subject of: National Library of New Zealand Catalogue

Keywords: Autobiography; Biography; Journals; Correspondence

Samuel Marsden | NZETC - Mozilla Firefox

http://www.nzetc.org/tm/scholarly/name-208673.html

SAMUEL MARSDEN (PERSON)

Chaplain, magistrate, agriculturalist, missionary.

Image Gallery
Works by this Author in Our Collection
Works by this Author in Other Collections
Mentioned In
External Links

Image Gallery

Works by this Author in Our Collection

- Letter from Samuel Marsden to Josiah Pratt, July 14th, 1819, quoted in Chapter II.
- Letter from Samuel Marsden to John Gare Butler, January 12th, 1820, quoted in Chapter II.
- Letter from Samuel Marsden to Josiah Pratt, January 20th, 1820, quoted in Chapter II.
- Letter from Samuel Marsden to Josiah Pratt, February 7th, 1820, quoted in Chapter II.

Works by this Author in other Collections

- The letters and journals of Samuel Marsden, 1765-1838

Mentioned in

- Notes on the Ancient Dog of the New Zealanders
  - Colenso.—Notes on the ancient Dog of the New Zealanders.
- The Early Journals of Henry Williams
  - Index of Persons
- The Long White Cloud
  - chapter vi Mission Schooner and Whale Boat
- The Maori Race
  - Chapter XXIV. Modern History.
- The Maori - Volume II
  - IX Social Customs— Continued Customs Pertaining To Birth
- The Old Whaling Days
  - 1837.
- Through Ninety Years
  - Chapter VI.
  - Chapter II.

External Links

Te Ara: 1966 Encyclopedia of New Zealand

Dictionary of New Zealand Biography

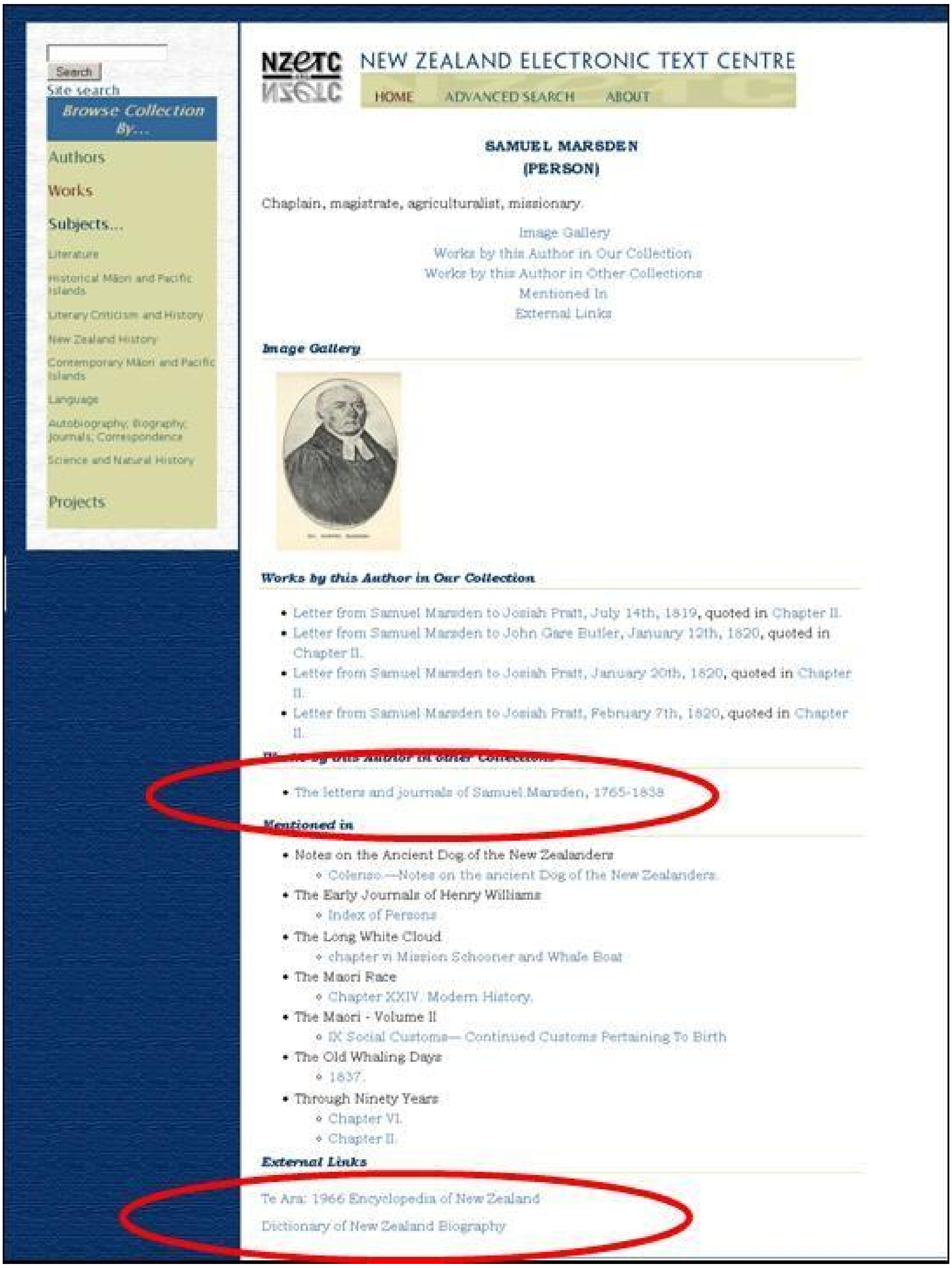
© 2005 Victoria University of Wellington

LETTER FROM SAMUEL MARSDEN TO JOSIAH PRATT, JANUARY 20TH, 1820

File Edit View Go Bookmarks Tools Help
Search
Site search
About this page...
Browse Collection By...
Authors
Works
Subjects...
Literature
Historical Māori and Pacific Islands
Literary Criticism and History
New Zealand History
Contemporary Māori and Pacific Islands
Language
Autobiography; Biography; Journals; Correspondence
Science and Natural History
Projects
HOME
ADVANCED SEARCH
ABOUT
Done
Home | Advanced Search

Figure 4 Sample pages from the NZETC collection showing hyperlinks embedded in HTML presentation of resources providing access to topic pages

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 5 below)


Search
Site search
Browse Collection By...
Authors
Works
Subjects...
Literature
Historical Māori and Pacific Islands
Literary Criticism and History
New Zealand History
Contemporary Māori and Pacific Islands
Language
Autobiography, Biography, Journals, Correspondence
Science and Natural History
Projects
NZETC
NEW ZEALAND ELECTRONIC TEXT CENTRE
HOME
ADVANCED SEARCH
ABOUT

SAMUEL MARSDEN (PERSON)

Chaplain, magistrate, agriculturalist, missionary.

Image Gallery
Works by this Author in Our Collection
Works by this Author in Other Collections
Mentioned In
External Links

Image Gallery

Works by this Author in Our Collection

- Letter from Samuel Marsden to Josiah Pratt, July 14th, 1819, quoted in Chapter II.
- Letter from Samuel Marsden to John Gare Butler, January 12th, 1820, quoted in Chapter II.
- Letter from Samuel Marsden to Josiah Pratt, January 20th, 1820, quoted in Chapter II.
- Letter from Samuel Marsden to Josiah Pratt, February 7th, 1820, quoted in Chapter II.

Works by this Author in Other Collections

- The letters and journals of Samuel Marsden, 1765-1838

Mentioned in

- Notes on the Ancient Dog of the New Zealanders
  - Colenso.—Notes on the ancient Dog of the New Zealanders.
- The Early Journals of Henry Williams
  - Index of Persons
- The Long White Cloud
  - chapter vi Mission Schooner and Whale Boat
- The Maori Race
  - Chapter XXIV. Modern History.
- The Maori - Volume II
  - IX Social Customs— Continued Customs Pertaining To Birth
- The Old Whaling Days
  - 1837.
- Through Ninety Years
  - Chapter VI.
  - Chapter II.

External Links

Te Ara: 1966 Encyclopedia of New Zealand

Dictionary of New Zealand Biography

Figure 5 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 collection entirely separate from the NZETC

The topic pages thus act as way markers pointing out possible routes for resource discovery.

Cross-collection linkages are particularly valuable fertile ground for examples of structured serendipity where automatically created navigational hyperlinks can create opportunities for “accidental” knowledge discover in the interdisciplinary connections they reveal. 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, while a single entity may be known by many names. 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), an authority control system that provides unique, persistent, sharable identifiers for any sort of entity.

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 with the family name first, and, when written in Chinese script, without a space between the name parts. The ability to identify entities within plain text and add structured, machine-readable mark-up contributes to the growth of well-structured electronic text corpora which can be delivered within the type of navigation framework described above.

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. 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.

The technologies developed and deployed by the NZETC including 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 programs and applications including web-based references tools, eResearch virtual spaces and institutional repositories.

Only with wider adoption of suitable entity identifiers and participation in a shared mapping systems 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 is required to support evermore interesting modes and methods of resource discovery.

References

Ambient Findability O'Reilly Media, 2005, by Peter Morville

EATS: an Entity Authority Tool Set, 2007, by Jamie Norrish (http://hdl.handle.net/10063/220)

Topic Maps and TEI - Using Topic Maps as a Tool for Presenting TEI Documents, 2005 by Conal Tuhoy (http://hdl.handle.net/10063/160)