Topic Maps and TEI – using Topic Maps as a tool for presenting TEI documents

Conal Tuohy, New Zealand Electronic Text Centre, Victoria University of Wellington

Abstract

This paper describes a method used by the website of the New Zealand Electronic Text Centre (NZETC), in which Topic Maps are used as a tool for presenting TEI-encoded texts in HTML form.

Many electronic text archives transform their TEI texts into HTML for publishing their texts on the World Wide Web. Typically each chapter or page is transformed from TEI into a separate web page. Such a method produces websites that have the same structure as a physical book.

However, TEI is more expressive than HTML and can encode many other features of interest than just chapters, pages, and paragraphs. For example, TEI is also used to encode information about people and places and events, as well as literary criticism, and linguistic analysis. Indeed, TEI is designed to be extended to suit all kinds of scholarly needs.

These more complex aspects of text encoding are more difficult to transform into HTML. Because TEI is designed to be convenient for scholars to encode complex information, rather than for readers to understand it, it is necessary to transform the TEI into another form suitable for display. For instance, where a TEI corpus includes references to people, these references might be collated together to produce an index. For practical purposes, it is often necessary to extract information from TEI into a database, so that it can be queried conveniently and transformed into a web site.

The new "Topic Map" standard of the International Standards Organisation is identified as a suitable technology for solving this problem. A topic map is a kind of Web database with an extremely flexible structure. This paper describes a framework for using TEI in conjunction with Topic Maps to produce a large website which can be navigated easily in many directions.

What is TEI?

TEI (Text Encoding for Interchange) is an XML-based markup language (or a family of markup languages) for encoding texts. Although TEI is sometimes used to encode "born-digital" materials such as websites, more typically it is used to digitally encode the contents of printed books and manuscripts.

Multiple perspectives

A TEI document consists of a structured metadata header, followed by text, which may be broken up into paragraphs, and organised into parts or chapters.

However, going beyond its simplest form, TEI also offers a wide variety of specialized vocabularies for:

- linguistic analysis
- bibliographic metadata
- commentary
- dictionaries and thesauri
- biography and history
- … to mention a few.

Each of these vocabularies is designed to conform to some scholarly perspective or serve some scientific, literary, or other purpose. A TEI document which uses a combination of these vocabularies might therefore encode a multi-disciplinary, multi-dimensional description of a text.

TEI is not a presentation format

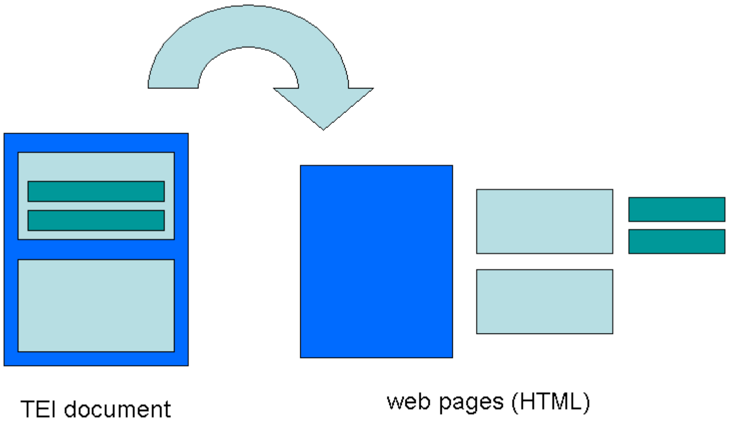
HTML is a simple presentation format in which text is described in terms of sections, headings, lists, and cross-references; similar to the concepts of word-processing. HTML is not generally concerned with the meaning of the text which it encodes. Because of this, generic web browsers can present HTML pages adequately no matter what their subject matter and content.

TEI, by contrast, is a format designed for text encoding in general, rather than just for presentation. TEI is much more complex than HTML, and can be extended even further to permit the description of texts according to whatever scheme or perspective a scholar wishes to use. The broadness and extensibility of the scope of TEI means that no single generic presentation mechanism can adequately present all possible TEI documents. Instead, any project which uses TEI must select, or develop, presentation mechanisms which are appropriate to the specific purposes of the project, and the specific types of encoding used in the project.

Data models and tools

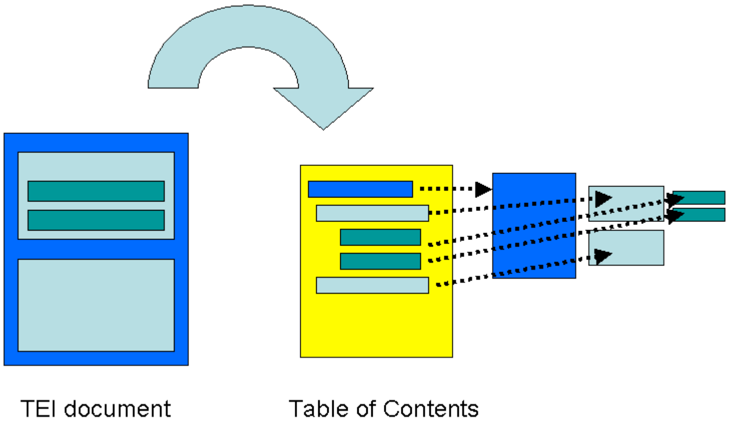
The currently standard practice for presenting TEI texts is to transform them into HTML, using Extensible Stylesheet Language Transformations (XSLT). XSLT is just one of a large number of XML-based technologies which we can use to model TEI texts as XML documents, with a tree-shaped structure.

Simple TEI documents have a structure which is comparable to that of HTML documents. These simple TEI documents can be transformed quite easy into HTML. Individual sections can be extracted and transformed into individual web pages.



Transforming TEI text into web pages

The tree structure of the TEI document can also be used to generate a web page containing a table of contents, with hyperlinks pointing to web pages which individually represent sections of the TEI. Similarly, tables of figures, lists of names, and other index pages can be generated by simple transformations of TEI into HTML.



Transforming TEI structure into a table of contents

However, the complex data structures encoded in many TEI documents are often encoded using references which cut across the tree-structure of the XML. Generic XML tools are therefore not always ideal for processing these data structures, because they operate at a lower level of abstraction. Many of the more specialised concepts of TEI have no simple equivalents in HTML. Information encoded in this way can still be presented in HTML form, but it requires a more sophisticated transformation.

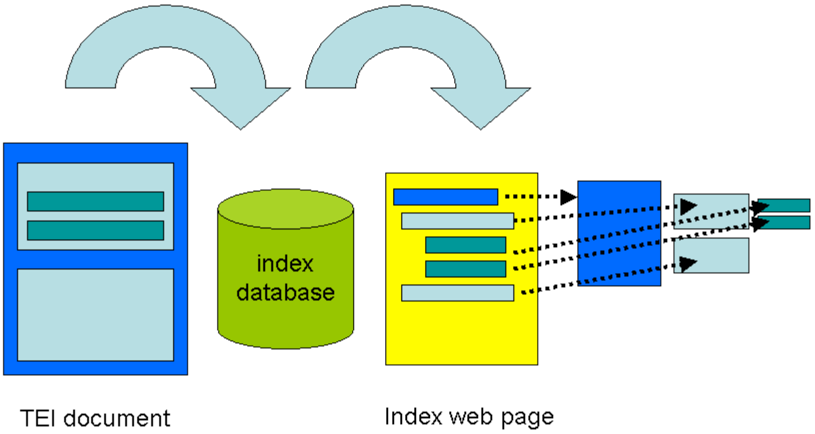
The markup in a TEI document is an expression of some theoretical model. Ideally, the software used to present the TEI-encoded text should also reflect that same model. However, because TEI embraces so many possible uses, it is a challenge to find software whose data model is really adequate to express all the different models which can be found in TEI documents.

Harvesting from TEI into a database

Some TEI markup can be hard to directly convert to an index, and often index pages will need to include material from a large number of TEI files

It is convenient to first extract (or "harvest") metadata from the TEI into a database, and then generate each HTML index from the database

For instance, where a TEI corpus includes references to people, these references might be collated together to produce an index. For practical purposes, it is often necessary to extract information from TEI into a database, so that it can be queried conveniently and transformed into a web site.



Harvesting TEI content into a database

Topic Maps are an appropriate technology

At the NZETC we identified Topic Maps as a technology which could be used to model all the rich data structures which we had encoded in TEI.

A topic map is a kind of hyper-textual metadata. If a web site is a hyper-*text*, then a topic map is a hyper-*index*.

The Topic Maps standard is an evolution of an earlier technology called "Topic Navigation Maps", designed to represent and to merge the indexes of printed books. Topic Maps were adopted by the International Standards Organisation and became an international standard for knowledge representation and integration. In 2000, Topic Maps acquired an XML serialisation format: a markup language called "XML Topic Maps" or XTM. This XML formalism has greatly simplified the implementation of Topic Map solutions and facilitated the integration of Topic Maps with other XML technologies. Topic Maps are supported by a growing number of software implementations, both proprietary and open source.

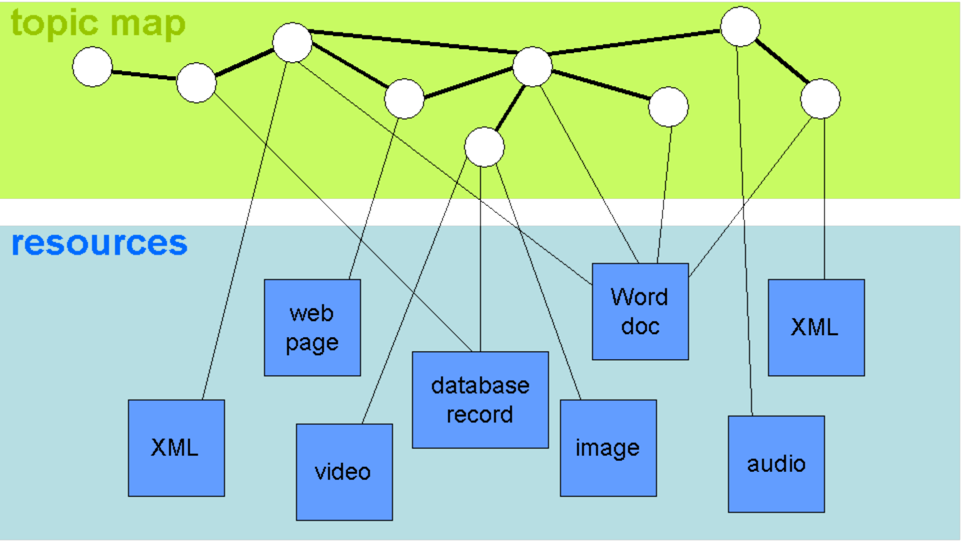
Topic maps are very free-form, able to manage all kinds of metadata structures: catalogue records, indexes, tables of contents, controlled vocabularies, multiple hierarchies, glossaries, thesauri, and taxonomies, all can be linked together within a single topic map. The structural flexibility of topic maps is very important to us since this is necessary to model the features which are found in TEI documents.

Another valuable feature of Topic Maps is that they are based on standard web technology. They have an XML syntax and they use URIs for hyperlinking. This makes them a natural fit with XML-based TEI, and with the World Wide Web.

Topic Map key concepts

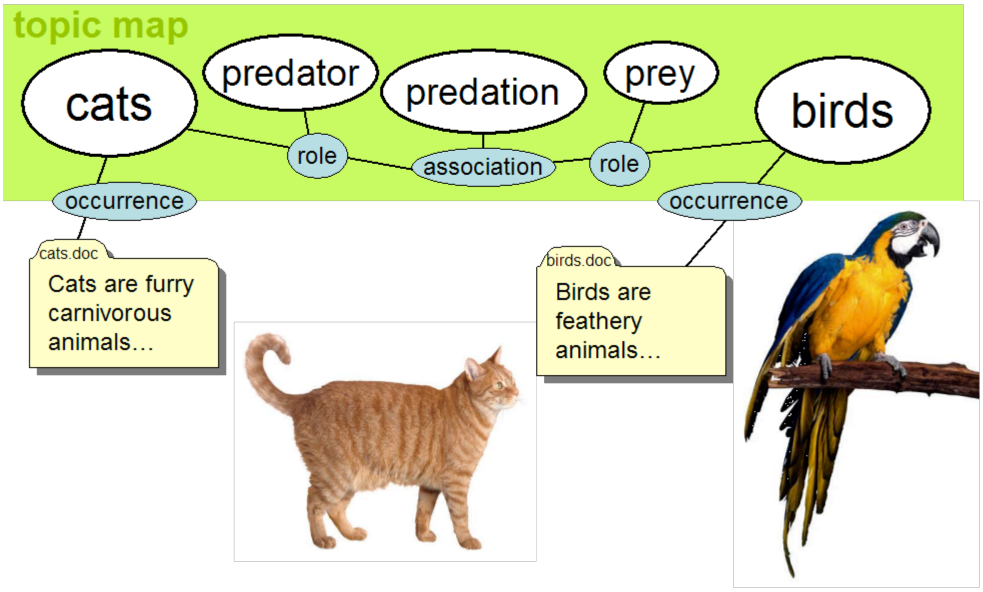
The following diagram shows how a topic map overlays a set of information resources, annotating them and defining relationships between them.

A topic map consists of a set of "topics" (the white circles in the diagram below). Each topic represents some subject of interest, which may represent a person, a document, a date, a page, a word, or whatever else is desired. Topics may be linked together by hyperlinks called "associations" (thick black lines in the diagram), and may also have associated information resources (i.e. documents of some sort) called "occurrences" (thin black lines in the diagram).



A topic map is a highly-structured repository of metadata which is distinct from the resources which it links to.

The diagram below shows a topic map containing topics called "cats" and "birds", which are intended to represent real cats and birds, respectively. The "cats" topic in the diagram has an *occurrence* called "cats.doc"; a resource which contains information about cats. Looking more closely at *associations*, we see that an association has a *type*, which is itself defined by a topic. Each member in the association plays a particular *role,* also defined by a topic. In the example, to model the fact that cats prey on birds, a *predation* association is defined in which cats play the *predator* role and birds play the role of *prey*.



Cats prey on birds

Building a topic map

We began our topic mapping by elaborating a conceptual model which would be adequate for the NZETC's digital library website.

First, we needed to decide which features of our TEI we were going to model in our Topic Map. Anything which we wanted to present on a page of its own had to be modelled as a topic.

One obvious set of elements were the TEI elements *group, text, front, body, back, div*, and *figure*. These elements, and the relationships between them, define the hierarchical structure of a book. We decided to create a topic in our topic map for every such element, which we called "TEI structural" topics. Where one such element contained another, we would also create a "containment" association between the corresponding topics in the topic map. This information would be used to generate tables of contents, and to provide next and previous links on each page.

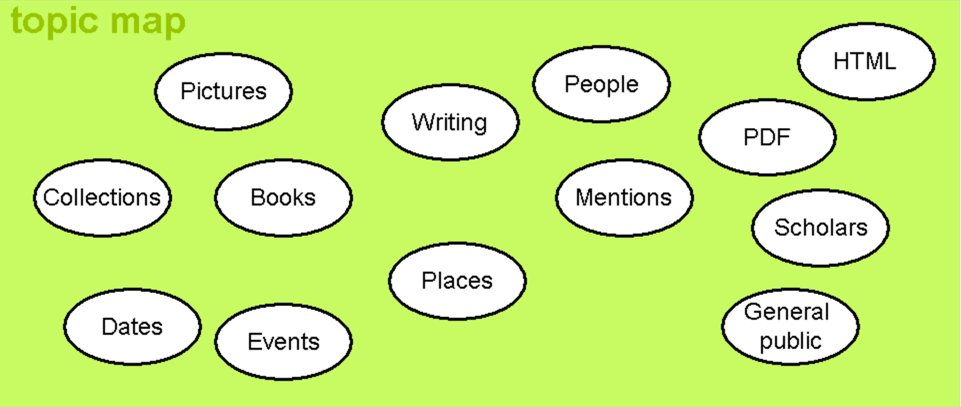
We also decided that we would create topics to represent people and places named in the texts using TEI *name* and *rs* markup. Initially this name markup was used for authors and publishers, but we gradually extended this to people mentioned in the body of the text as well.

Our XSLT harvester would create associations linking these named entities with the structural topics which mentioned them. This information would be used generate a web page for a person, containing links to all the places where the person is mentioned.

The other topics of interest were mainly derived from bibliographic data taken from the *teiHeader* metadata element. Such things as author, editor, funder, publisher, publication places and dates, subject classification, revisions, etc, could all be extracted from the TEI, represented as topics, and linked by associations to the topics representing the texts.

This analysis was the starting point for building our conceptual framework or ontology.

Ontology



Ontology

To map your information, you first need to identify what kinds of subjects are of interest, what kinds of relationships these subjects may have, and what kinds of people would be interested (or what kinds of interest would they have).

For example, in our topic map the kinds of things we are interested in include books, pictures, people, and places.

The kinds of relationships we're interested in include: books mention people, people write books, books contain pictures, and people and places are depicted in pictures.

People might be simply interested in the content of texts, or they might have a scholarly interest in the physical details of the texts.

These concepts are what our site is all about. They form the conceptual foundation of a topic map, sometimes called an "ontology".

The term "ontology" as it is used here is borrowed from philosophy, where it refers to the branch of metaphysics which deals with the nature of being and existence. Computer scientists have dragged the word down from its philosophical heights, and it can now even be used simply to mean a computerised representation of a conceptual framework.

We could have just built an ontology of our own, from these concepts, but we felt we might discover conceptual inadequacies in our model later, so to play safe we linked it up to an existing ontology.

This ontology was produced by the International Committee for Documentation (CIDOC) within the International Council of Museums (ICOM), known as the CIDOC "Conceptual Reference Model".

The CIDOC ontology is quite large, and we only needed part of it, so we just selected the concepts we needed, used those, and ignored the rest. If and when we decide we need to expand our ontology, we'll just add some more of their concepts into our ontology.

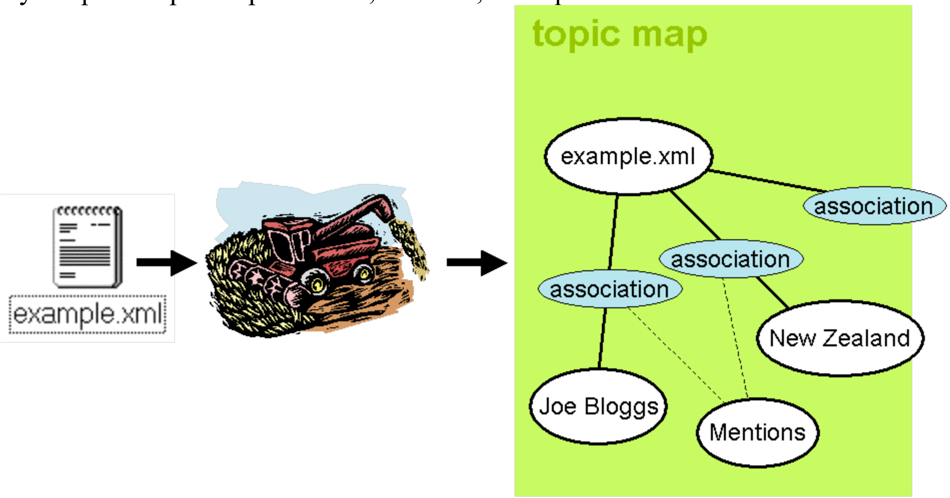
Harvesting topic maps

In our project we already had considerable experience with XSLT, which we'd been using to convert TEI documents into HTML, so we were able to put those skills to use writing XSLT to transform TEI documents into Topic Maps. These transformations identify the topics of interest in a TEI document and the relationships between those topics.

To produce the final Topic Map, all these harvested topic maps will be merged together. This produces a map of our entire website, currently containing about 50 thousand topics in total.

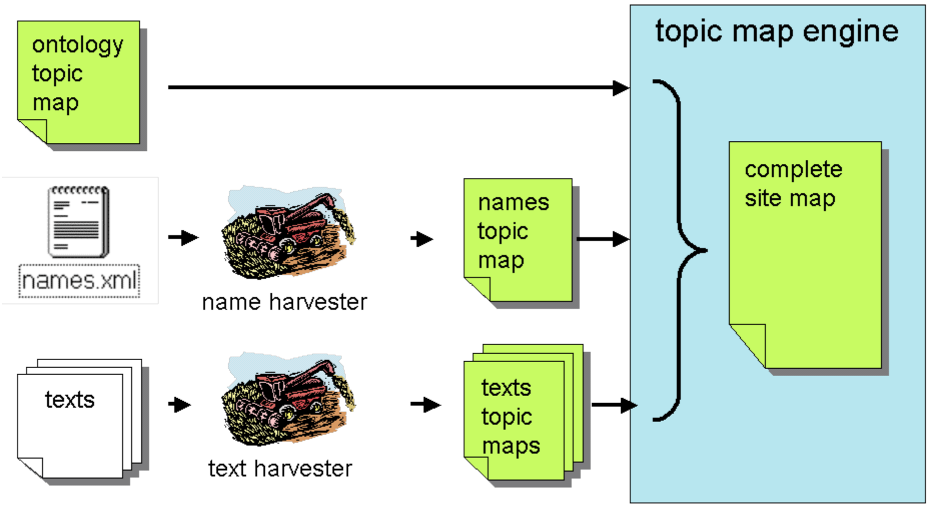
At the NZETC, our source materials are all in XML, which makes it particularly easy for a harvester to find the things it's interested in – names of people and places, pictures, authors, etc. Our harvester takes only minutes to read through hundreds of megabytes of XML files, and harvest tens of thousands of topics and hundreds of thousands of associations from out of them.

By encoding our metadata in TEI we are able to edit it freely while retaining the benefits of schema validation. By automatically harvesting metadata from the TEI to maintain the topic map we can reliably keep the topic map accurate, detailed, and up-to-date.



Harvesting a topic map from a TEI XML document

Our TEI harvester is an XSL transformation which transforms the TEI document into an XML Topic Map document. The XSLT creates an XTM topic to represent each TEI document (and each major section of the document). The XSLT also creates XTM topic for each TEI name it finds, and creates a "mentions" association between the topic which represents the text and the topic which represents the thing named. In the example the topic map asserts that the TEI document mentions both Joe Bloggs and New Zealand.



Topic maps harvested and merged to form a single unified map

The final topic map is compiled by a topic map engine, by merging the topic maps from our three sources:

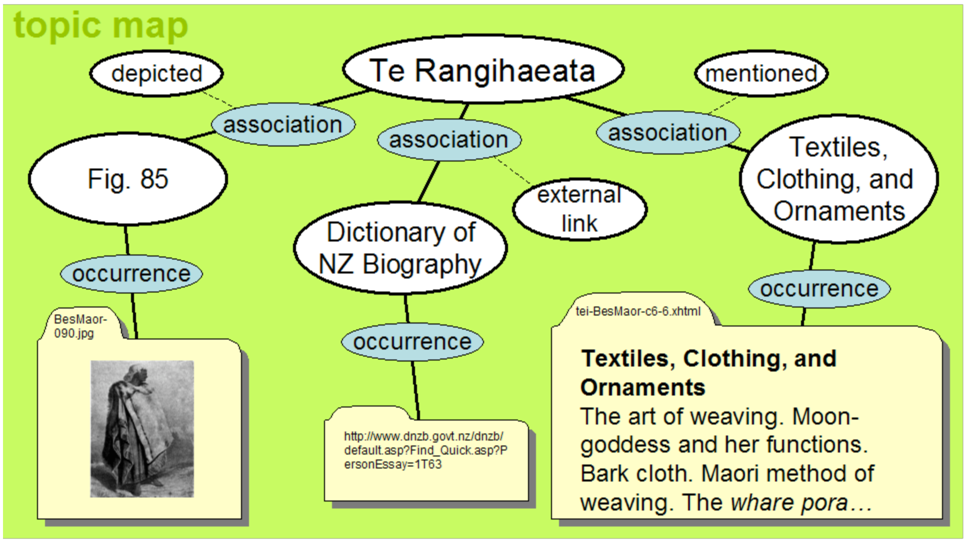
1. Our ontology topic map (containing the small set of our basic concepts).
2. The names map harvested from our name list (quite a large topic map)
3. Topic maps harvested from our XML texts (there are a couple of hundred of these, some big, some small)

As it reads each topic map, the topic map engine tries to find existing topics that match the topics in the map being imported. Where it finds an existing topic which represents the same subject as the new topic, the topic map engine automatically merges the two topics together. Topics can be made to merge automatically simply by sharing a unique identifier of some sort. This means that the data import process is declarative rather than procedural; it is simply enough to assert that two topics represent the same subject, and the Topic Map engine will merge them into a single topic, combining all the characteristics of the two original topics.

Creating a website from the topic map

Finally it only remains to display the topic map as a website.

To do this, we programmed our web server to generate a web page for each topic in the map. To do this, the web server asks the topic map engine for a topic, and creates a web page by copying information from the topic, as well as from topics which are associated with it, and from occurrences of those topics.


topic map
depicted
Te Rangihaeata
mentioned
association
association
association
Fig. 85
external link
Textiles, Clothing, and Ornaments
Dictionary of NZ Biography
occurrence
occurrence
occurrence
BesMaor-090.jpg
tei-BesMaor-c6-6.xhtml
http://www.dnzb.govt.nz/dnzb/default.asp?Find_Quick.asp?PersonEssay=1T63
Textiles, Clothing, and Ornaments
The art of weaving. Moon-goddess and her functions. Bark cloth. Maori method of weaving. The whare pora…

Fragment of NZETC website topic map

The diagram above represents part of our topic map. The central topic, Te Rangihaeata, was a chief of a Maori tribe, the Ngati Toa. In the topic map, the topic which represents him is associated with three other topics, each of which has an occurrence.

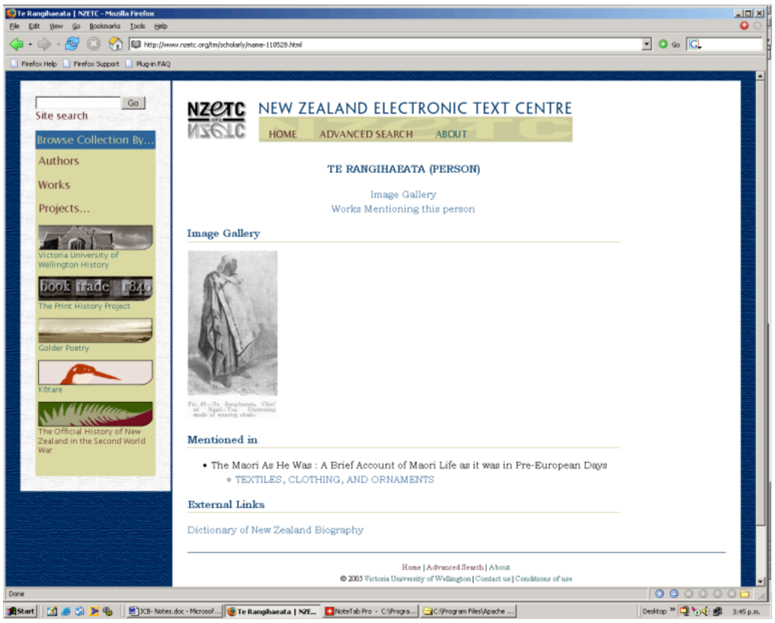
The association on the left represents a depiction of Te Rangihaeata. The picture which depicts him is "Figure 85" from a book by Elsdon Best called *The Maori as he was*.

On the right, "Textiles, Clothing and Ornaments" is a chapter from the same text, which mentions him. Both of these associations were of course harvested from TEI name markup in the TEI file in which *The Maori as he was* is encoded.

In the centre of the diagram, Te Rangihaeata is associated with a web page on the website of the Dictionary of New Zealand Biography. This last piece of information was harvested from our name list.

Note that the central topic "Te Rangihaeata" was harvested twice – once from the Elsdon Best book, and once from the names list. After harvesting, these two topics merged together automatically, leaving us with just one topic with 3 associations.

The figure below shows how the "Te Rangihaeata" topic is displayed as a web page. The page shows his name, the depiction, and the mention, as well as the external link.


Te Rangihaeata | NZETC - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.nzetc.org/tm/scholarly/name-110528.html
Go
Firefox Help
Firefox Support
Plug-in FAQ
Go
Site search
Browse Collection By...
Authors
Works
Projects...
Victoria University of Wellington History
The Print History Project
Golder Poetry
Kōtare
The Official History of New Zealand in the Second World War
NZETC
NEW ZEALAND ELECTRONIC TEXT CENTRE
HOME
ADVANCED SEARCH
ABOUT
Done
Start
JCB- Notes.doc - Microsof...
Te Rangihaeata | NZE...
NoteTab Pro - C:\Progra...
C:\Program Files\Apache ...
Desktop
3:45 p.m.

TE RANGIHAEATA (PERSON)

Image Gallery
Works Mentioning this person

Image Gallery

Mentioned in

- The Maori As He Was : A Brief Account of Maori Life as it was in Pre-European Days
  - TEXTILES, CLOTHING, AND ORNAMENTS

External Links

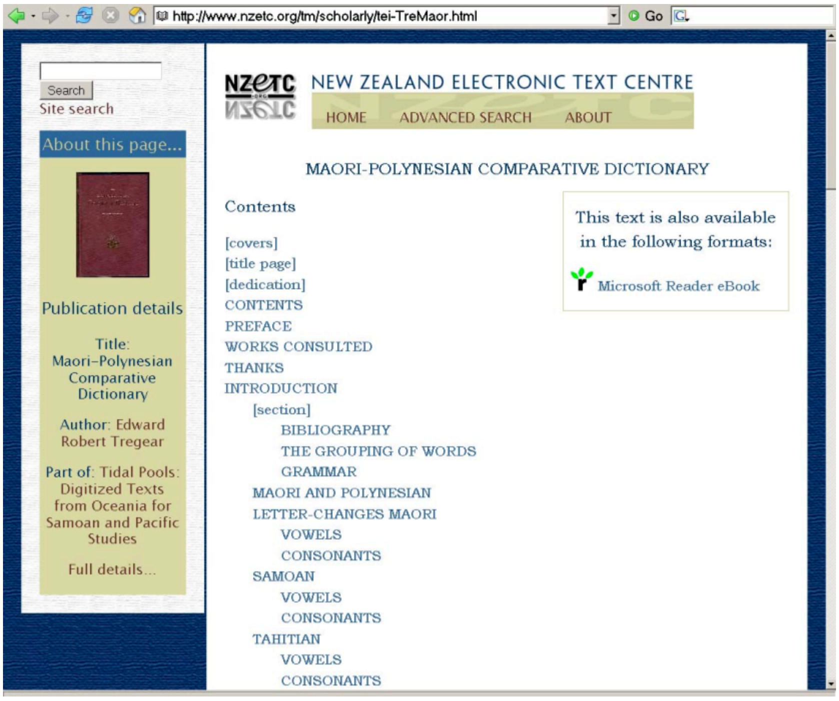
Dictionary of New Zealand Biography

Home | Advanced Search | About
© 2005 Victoria University of Wellington | Contact us | Conditions of use

Te Rangihaeata web page

Topic Map supports exploration

The topic map underlying the website now allows for an exploratory style of navigation.
The web page below shows a dictionary written by a scholar called “Edward Tregear”. Notice that his name appears (as a hyperlink) in a sidebar on the left. This sidebar is visible not only here on the contents page, but throughout the book.
This hyperlink is an expression of an association in the topic map between the topic representing Edward Tregear, and the topic representing the writing of the dictionary. This association was harvested from a TEI *author* element in the *teiHeader* element of the text


http://www.nzetc.org/tm/scholarly/tei-TreMaor.html
Go
Search
Site search
NZETC
NEW ZEALAND ELECTRONIC TEXT CENTRE
HOME
ADVANCED SEARCH
ABOUT

About this page...

Publication details

Title: Maori–Polynesian Comparative Dictionary

Author: Edward Robert Tregear

Part of: Tidal Pools: Digitized Texts from Oceania for Samoan and Pacific Studies

Full details...

MAORI-POLYNESIAN COMPARATIVE DICTIONARY

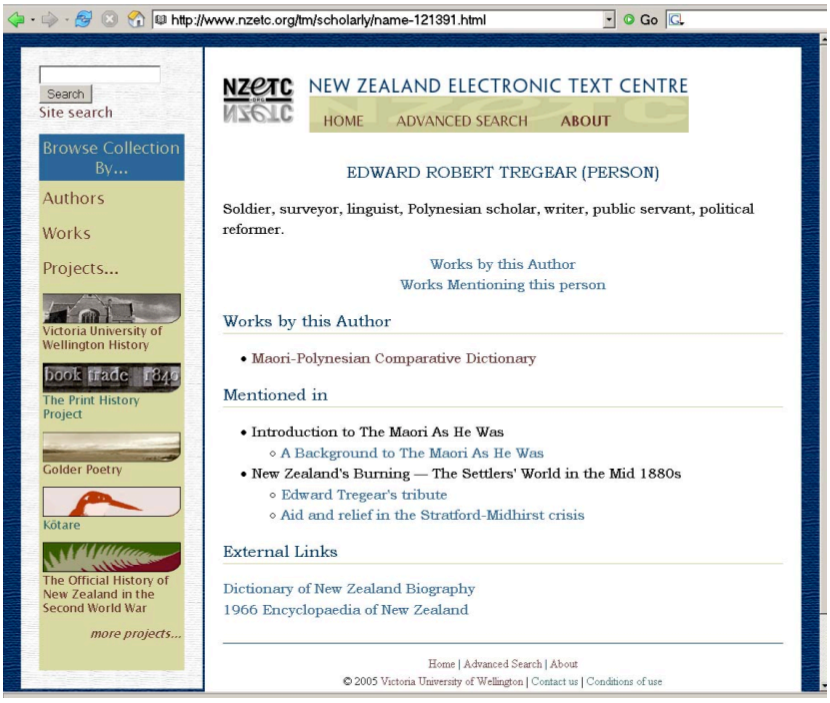
This text is also available in the following formats:

Microsoft Reader eBook

Contents

Edward Tregear's *Maori-Polynesian Comparative Dictionary*

Clicking on his Edward Tregear's name takes you to a page about him:


http://www.nzetc.org/tm/scholarly/name-121391.html
Go
Search
Site search
NZETC
NEW ZEALAND ELECTRONIC TEXT CENTRE
HOME
ADVANCED SEARCH
ABOUT
Browse Collection By...
Authors
Works
Projects...
Victoria University of Wellington History
The Print History Project
Golder Poetry
Kōtare
The Official History of New Zealand in the Second World War
more projects...

EDWARD ROBERT TREGEAR (PERSON)

Soldier, surveyor, linguist, Polynesian scholar, writer, public servant, political reformer.

Works by this Author
Works Mentioning this person

Works by this Author

- Maori-Polynesian Comparative Dictionary

Mentioned in

- Introduction to The Maori As He Was
  - A Background to The Maori As He Was
- New Zealand's Burning — The Settlers' World in the Mid 1880s
  - Edward Tregear's tribute
  - Aid and relief in the Stratford-Midhirst crisis

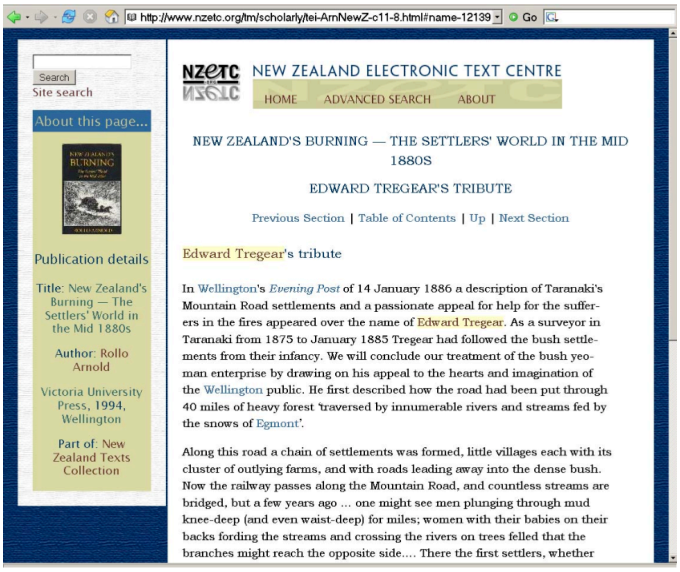
External Links

Dictionary of New Zealand Biography
1966 Encyclopaedia of New Zealand

Home | Advanced Search | About
© 2005 Victoria University of Wellington | Contact us | Conditions of use

Edward Tregear web page

This page contains links to everything on the site related to Edward Tregear. Note that the *Maori-Polynesian Comparative Dictionary* is listed as one of his works.
Note also the list of works where Edward Tregear is mentioned. Clicking on one of these links will display the page of the topic which represents the text where he is mentioned.


http://www.nzetc.org/tm/scholarly/tei-ArnNewZ-c11-8.html#name-12139
Go
Search
Site search
About this page...
NZETC
NEW ZEALAND ELECTRONIC TEXT CENTRE
HOME
ADVANCED SEARCH
ABOUT

NEW ZEALAND'S BURNING — THE SETTLERS' WORLD IN THE MID 1880S

EDWARD TREGEAR'S TRIBUTE

Previous Section | Table of Contents | Up | Next Section

Publication details

Title: New Zealand's Burning — The Settlers' World in the Mid 1880s

Author: Rollo Arnold

Victoria University Press, 1994, Wellington

Part of: New Zealand Texts Collection

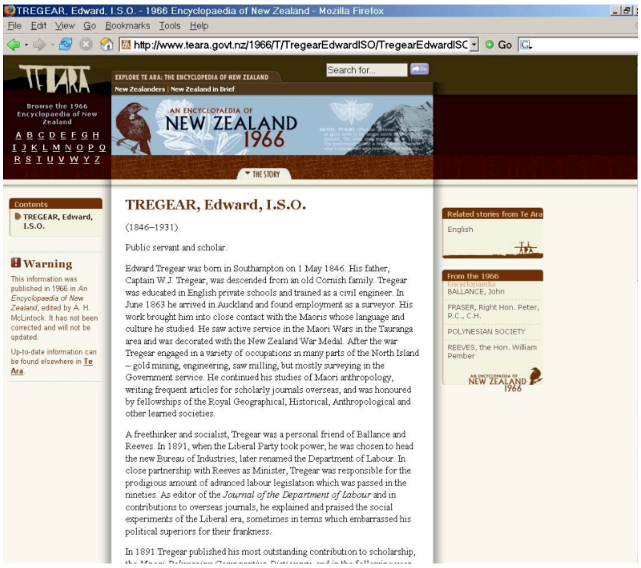
Edward Tregear's tribute

In Wellington's *Evening Post* of 14 January 1886 a description of Taranaki's Mountain Road settlements and a passionate appeal for help for the sufferers in the fires appeared over the name of Edward Tregear. As a surveyor in Taranaki from 1875 to January 1885 Tregear had followed the bush settlements from their infancy. We will conclude our treatment of the bush yeoman enterprise by drawing on his appeal to the hearts and imagination of the Wellington public. He first described how the road had been put through 40 miles of heavy forest 'traversed by innumerable rivers and streams fed by the snows of Egmont'.

Along this road a chain of settlements was formed, little villages each with its cluster of outlying farms, and with roads leading away into the dense bush. Now the railway passes along the Mountain Road, and countless streams are bridged, but a few years ago ... one might see men plunging through mud knee-deep (and even waist-deep) for miles; women with their babies on their backs fording the streams and crossing the rivers on trees felled that the branches might reach the opposite side.... There the first settlers, whether

Rollo Arnold's book *New Zealand's Burning* makes a mention of Edward Tregear

The topic map also defines the content of the list of external links on Tregear's page. Clicking on one of these links loads a page from an external site. This external link was harvested along with the Edward Tregear topic from our name authority file.


TREGEAR, Edward, I.S.O. - 1966 Encyclopaedia of New Zealand - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.teara.govt.nz/1966/T/TregearEdwardISO/TregearEdwardISC
Go
TE ARA
Search for...
EXPLORE TE ARA: THE ENCYCLOPEDIA OF NEW ZEALAND
New Zealanders | New Zealand in Brief
Browse the 1966 Encyclopaedia of New Zealand
A B C D E F G H I J K L M N O P Q R S T U V W Y Z
AN ENCYCLOPAEDIA OF NEW ZEALAND 1966
THE STORY
Contents
TREGEAR, Edward, I.S.O.
Related stories from Te Ara
English
From the 1966 Encyclopaedia
BALLANCE, John
FRASER, Right Hon. Peter, P.C., C.H.
POLYNESIAN SOCIETY
REEVES, the Hon. William Pember

TREGEAR, Edward, I.S.O.

(1846–1931).

Public servant and scholar.

Edward Tregear was born in Southampton on 1 May 1846. His father, Captain W.J. Tregear, was descended from an old Cornish family. Tregear was educated in English private schools and trained as a civil engineer. In June 1863 he arrived in Auckland and found employment as a surveyor. His work brought him into close contact with the Maoris whose language and culture he studied. He saw active service in the Maori Wars in the Tauranga area and was decorated with the New Zealand War Medal. After the war Tregear engaged in a variety of occupations in many parts of the North Island – gold mining, engineering, saw milling, but mostly surveying in the Government service. He continued his studies of Maori anthropology, writing frequent articles for scholarly journals overseas, and was honoured by fellowships of the Royal Geographical, Historical, Anthropological and other learned societies.

A freethinker and socialist, Tregear was a personal friend of Ballance and Reeves. In 1891, when the Liberal Party took power, he was chosen to head the new Bureau of Industries, later renamed the Department of Labour. In close partnership with Reeves as Minister, Tregear was responsible for the prodigious amount of advanced labour legislation which was passed in the nineties. As editor of the *Journal of the Department of Labour* and in contributions to overseas journals, he explained and praised the social experiments of the Liberal era, sometimes in terms which embarrassed his political superiors for their frankness.

In 1891 Tregear published his most outstanding contribution to scholarship,

Warning

This information was published in 1966 in *An Encyclopaedia of New Zealand*, edited by A. H. McLintock. It has not been corrected and will not be updated.

Up-to-date information can be found elsewhere in Te Ara.

Edward Tregear external link