Topic Map Presentation Framework: An Approach To Delivering Newspaper Content Over The Web.

Alison Stevenson (NZETC[1]),
Elizabeth Styron (NZETC[1])

Introduction

This paper will discuss some of the requirements for the successful online delivery of newspaper archive content to users and examine an innovative approach taken to fulfil those requirements by using a semantic framework. This approach is based on the digital library delivery system configured by the New Zealand Electronic Text Centre (NZETC), the working name of which is the Topic Map Presentation Framework (TMPF). It has been developed as a production system at the NZETC since 2002 and is currently used to deliver the Centre's own growing collection of digital resources, a nationally important website containing more than 40,000 pages and over half-a-million hyperlinks. Since 2005 the NZETC has been working with APEX to further develop the TMPF as means of providing sophisticated access to digitised newspaper archives wherein the semantic navigation of online resources greatly enhances the user experience in digital libraries. Using this approach, an ontology codifies an analysis of the structure and relationships in the domain of newspaper publishing and archiving, including publishers, issues, articles, pages, clippings, places, and dates. Metadata is automatically harvested from the source materials into this conceptual framework, producing a map of the content. This map is then used to present the content online in a meaningful structure.

This paper will also briefly cover some of the open technologies used, including Topic Maps, the CIDOC Conceptual Reference Model, Apache Cocoon, and Apache Lucene.

Delivering newspaper content over the web

Delivering newspaper content over the web presents many challenges and requires a delivery system that is technically sophisticated on the inside and intuitive to use on the outside. There are challenges to be met in several areas: the forms of content to be delivered, the granularity of access provided, modes of access enabled, extensibility, customization, usability and performance. This section will look briefly at each of these areas in turn to provide a context for the subsequent discussion of the TMPF.

As regards the forms of content to be delivered over the web from a newspaper archive, a key question is whether to provide user with access to the transcribed text of the newspaper, to images of the original printed newspaper, or to both? Delivering both forms of content has several advantages. The images enable access to content which cannot be transcribed such as photographs, illustrations and complex diagrams. They also provide information through visual details such as the size of headline, the style of font, or design of page layout which is not conveyed by the transcribed text. The transcribed text itself can be usefully provided to users because, in addition to providing the raw material for full text searches, in most cases it will be more legible than the original. As text it is also more malleable than an image in that the presentation can be altered to fit users need by altering font sizes, page layout and other variables. In some projects there may be an additional requirement to provide user access to source images as distinct from access images. Consideration is also needed of what tools

[1] New Zealand Electronic Text Centre

and mechanisms will be provided to enable users to engage with and use the content. Tools, for example, to zoom in and out of access images; to provide a version of each newspaper page suitable for printing; to send a URL, page or image to an email address; to downloading or print an entire newspaper issue; to downloading metadata records (e.g. for inclusion in a reference management system like Endnote). It is often the case that a newspaper archive will be incomplete, that there will be missing issues from a publication or missing pages from a particular issue. There may also be printing anomalies in newspaper pages such as incorrect page numbers and incorrect dates. A newspapers delivery system needs to be able to handle such cases, make omissions visible to users and support metadata describing printing anomalies which can also be made visible to users.

The granularity of access provided is an interesting question. Should users be able to navigate or search to the level of a newspaper page, or to the individual articles on those pages? This is an area in which the delivery system is almost wholly dependant on decisions made during the digitisation process. If article-level access images and mark-up have not been created the delivery system can only operate at the page level. In general, it is more challenging to define articles within newspapers than to define structural sub-units of journals and monographs. In most journals, the articles are easily identified as distinct units within the whole. Chapters serve that function for books. Smaller pamphlets are typically treated as a whole, without sub-units. When both commercial publishers of historic newspapers, as well as institutions such as the British Library, poll user groups to find out whether users prefer working with page-level or article-level files, the overwhelming response is that the preference is for article-level access. Users certainly want the ability to view an entire page, but most users seem to find that navigating from a hit list of relevant article citations directly to the article itself makes for a more efficient and satisfying user experience.

The USA's Library of Congress (LOC) is currently in the midst of a project to create a test bed of digitized newspapers for its National Digital Newspaper Program (NDNP), the specifications for which are page-level files. Nevertheless, many of the US State Libraries which are doing the digitization of the issues coming from their own collections are creating two sets of digitized objects: issue-level files for delivery to the LOC and also a set of article-level files for themselves, since the state-level editorial boards selecting the titles for conversion requested them. While it remains to be seen what the LOC will ultimately specify once the test bed is created and users have weighed-in on the results, the vast majority of other institutions involved with digitizing significant quantities of newspaper content are adopting article-based models over page-based models.

When we consider the modes of access to be provided, we are asking what search facilities will be presented to the user, what browsing methods will be enabled? Where an article-based model has been adopted during the digitisation process then the delivery system should make the most of this source material and allow navigation straight to a given article as well as between articles on the same page or in the same issue, and from page to page and issue to issue. Browsing by newspaper or article title, by date, by geographic region, by author, by subject or by any other item of metadata may be appropriate. Ideally simple full text searching of all newspapers should be available as well as a separate "advanced" search interface which includes support for fielded searching of selected metadata fields; standard Boolean search operators, phrase queries, wildcard queries and proximity operators. Allowing searches to be limited by date, publication, or other metadata fields (e.g. newspaper category, article type) would further increase functionality. Finally, as regards modes of access, there is a need to develop support for access by other systems through interoperability tools and protocols such as OpenURL[2], OAI-PMH[3] and SRU / SRW[4]. AT the NZETC we feel strongly that the systems should conform to relevant public guidelines, specifications and standards where possible to achieve a high level of modularity of the system architecture and to facilitate broad interoperability.

[2] OpenURL is ANSI/NISO Standard Z39.88-2004
[3] The Open Archives Initiative Protocol for Metadata Harvesting
[4] Library of Congress standard for web services for search and retrieval based on Z39.50 semantics

Extensibility is the degree to which the system will enable the inclusion, not only of other resources and data types, but also of metadata from other sources so as to enhance existing content. Customization is the extent to which the interface is customizable, both by the institution maintaining the archive (e.g. to meet branding guidelines) and by the user. For a user this might mean a 'MyNewspaperArchive' approach which could allow users to use personalised features, such as personal preferences, favourite newspaper titles, saved pages, saved search result sets. Finally, usability and performance cover a range of issues such as what range of browsers will be supported? What level of context sensitive help will be provided? Will the system meet W3C accessibility guidelines[5]? What performance requirements must be met?

The NZETC Topic Map Presentation Framework

These were the challenges facing the NZETC when we started develop the Topic Map Presentation Framework (TMPF) for newspaper delivery. The work was prompted by the actions of the National Library of New Zealand and the National Library of Australia both of whom issued, in 2005, public requests for proposals to provide online access to their existing newspaper archives. In response to these requests the NZETC starting working with APEX CoVantage to explore what could be achieved by combining experience in developing semantic navigation frameworks at the NZETC with the newspaper digitisation expertise at APEX. This section will describe some of salient features of the TMPF before describing the example newspaper delivery system built using the TMPF.

The TMPF in production at the NZETC is a dynamically-generated semantic framework – a metadata repository implemented using the ISO Topic Map standard 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 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 seven year effort by the CIDOC Conceptual Reference Model group to create a high-level ontology known as the CIDOC CRM[6]. 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 TMPF on the event-based model of the CIDOC CRM as illustrated below.

[5] W3C Recommendation on making web content accessible to people with disabilities. http://www.w3.org/TR/WAI-WEBCONTENT/

[6] CIDOC CRM http://cidoc.ics.forth.gr/

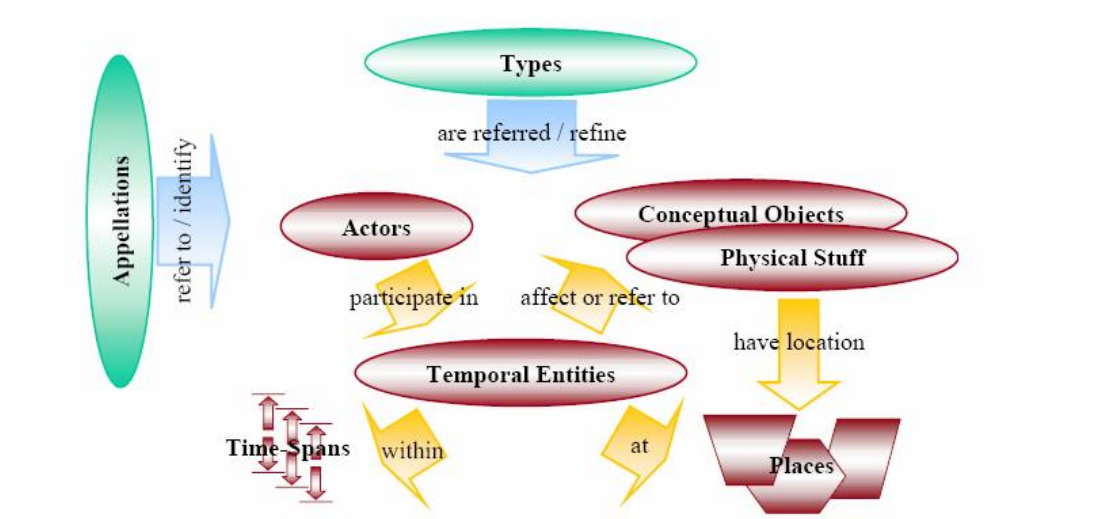

Figure 1 A qualitative metaschema of the CIDOC CRM taken from Martin Doerr "The CIDOC CRM – An Ontological Approach to Semantic Interoperability of metadata AI Magazine", Volume 24, Number 3 (2003)

This allows us to express relationships such as those illustrated in the diagram below.

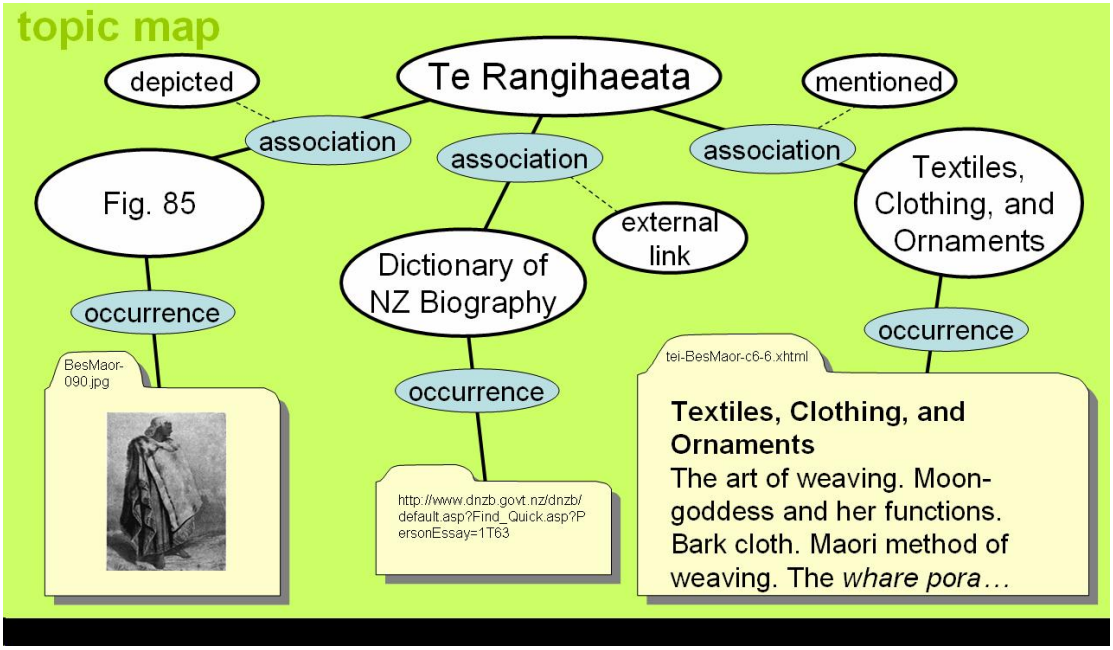

Figure 2 Illustration of relationships encoded in NZETC TMPF

The central topic, Te Rangihaeata, was a chief of 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 the XML file containing the Elsdon Best book. In the centre, 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. But these two topics merged together automatically, leaving us with just one topic with 3 associations.

To construct our topic map, we use XSLT[7] stylesheets to extract metadata from each of our XML text files, and express it in the XTM[8] format. In this way we automatically create hundreds of topic maps, each of which describes one of our texts. We also harvest information about people, places and organisations from a MADS[9] authority file which we construct from what is mentioned in our collection. Finally we merge the harvested topic maps together to create a unified topic map which describes our entire website.

By harvesting not only bibliographic metadata but also references to people, organisations and places, the site provides individual pages for topics of interest, linked automatically to those places they are mentioned or illustrated. Being automatically generated from the source XML files, maintenance is simple and the number and types of topics linked to can be increased simply by adding extra mark-up to the texts.

Each page on the website represents one of these topics, along with any associated topics. The screenshot below is the page for Te Rangihaeata.

Figure 3 Screenshot from NZETC Collection

The topic map manages all the hyperlinks, bibliographic metadata, structural metadata, annotations, classifications, name authorities and glossaries for the entire website. By extracting all the metadata needed from every resource available and merging it all together in the topic map, it is ensured that all relevant information is prepared and readily available to

[7] W3C specification for the syntax and semantics of a language for transforming XML documents into other XML documents http://www.w3.org/TR/xslt
[8] XML Topic Maps http://www.topicmaps.org/xtm/
[9] Metadata Authority Description Schema. A Library of Congress standard for a MARC21-compatible XML format for the type of data carried in records in the MARC Authorities format. http://www.loc.gov/standards/mads/mads.xsd

the presentation system, so that the presentation of every web page can include as much contextual information as is desired.

When TMPF generates a web page about a particular topic (whether a newspaper series, article, page, or any other type of topic), it can query the topic map to find all the information resources related to that topic. This would include the name of the topic, including aliases and names it might have in different languages, background articles, links to external websites, photos, etc. TMPF would then display those resources appropriately, either by generating hyperlinks to those resources or by simply including the content of other resources. For an example of how a page can be enriched by including related content from the topic map, see the hyperlink reference to Governor Grey in the pamphlet "One of England's Little Wars," on the NZETC website:

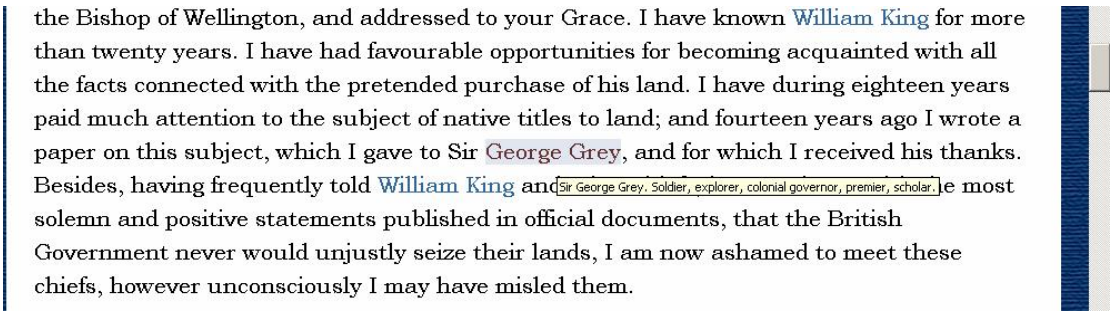

Figure 4 Screenshot from NZETC website. http://www.nzetc.org/tm/scholarly/tei-HadOneO-t1-body.html#name-208095-1

The hyperlink pointing to the page about Grey has a tool tip (an HTML title attribute) saying "Sir George Grey. Soldier, explorer, colonial governor, premier, scholar." This tool tip was drawn not from the encoded text of the pamphlet but from a record about George Grey in an authority file. When the authority file was imported into the topic map, a topic was created to represent George Grey, and this topic was merged with all the references to George Grey in the encoded texts, so that every hyperlink on the website which points to the George Grey page now has the same "authoritative" tool tip.

TMPF makes use of a number of public guidelines, specifications and standards, as listed in the table below.

Purpose	Specification	Reference
Encoding (general)	XML	http://www.w3.org/TR/REC-xml/
Addressing	URI	http://www.w3.org/Addressing/
Text encoding	TEI	http://www.tei-c.org/Guidelines2/
OCR encoding (under development)	ALTO	http://www.loc.gov/ndnp/techspecs.html
Authority control	MADS	http://www.loc.gov/standards/mads/
Transformation of XML source materials into presentation formats	XSLT	http://www.w3.org/TR/xslt
Presentation	HTML	http://www.w3.org/MarkUp/
Page formatting	CSS	http://www.w3.org/Style/CSS/
Image processing	SVG	http://www.w3.org/Graphics/SVG/

Extraction of metadata from XML source materials	XSLT	http://www.w3.org/TR/xslt
Information mapping	ISO Topic Maps	http://www.isotopicmaps.org/
Conceptual reference model	CIDOC CRM	http://cidoc.ics.forth.gr/

Using the TMPF for Newspaper Delivery

When the NZETC started to work with APEX to use the TMPF as a newspaper delivery system we created two small demonstration sites for the National Library of Australia and the National Library of New Zealand using content from their newspaper archives. The Australian site has 7413 web pages, including 204 issues, consisting of 847 newspaper pages and 6362 articles. The New Zealand site is about half that size. The creation of these demonstration sites was work undertaken at the NZETC and APEX to explore how well the Topic Map system worked with newspaper content and does not represent any decision by either National Library to use this technology. The NZETC is just one of many respondents to the RFPs issues by the libraries. The use of the TMPF to deliver example content from the two libraries on these demonstration sites is not indicative of any technology or vendor choices or decisions made by either library. The two demonstration sites are not available to the public.

The APEX Intelligent Zoning and Algorithmic Conversion (IZAAC) was used to convert digital images of the newspapers into XML encoded text[10]. In Apex's process, an Article is defined as a "complete newspaper story." This concept is extended to a number of newspaper elements, including small, continuous units of like content, such as news briefs and classified advertisements, and stand-alone features, such as photographs, illustrations, and crossword puzzles. The IZAAC process relies on human cognition to identify and zone articles, rather than relying on an automated process.

Automated approaches are adequate for extremely simple layouts and standardized formats; but when applied to newspapers, the results are usually unacceptable. Logical analysis alone cannot recognize discontinuous portions of an Article even on the same page, let alone across pages. Perhaps for this reason, many commercial software packages that rely on automated article definition use proxy terms for Articles, such as "chunks". Chunks can miss portions of an Article or may overlap across Articles. Neither is acceptable from a user-experience perspective.

Illustrations that are part of an Article are included with the Article. Illustrations that do not have associated text are treated as separate Articles. Captions are included with the illustrations and are tagged as such, supporting searches by caption. Articles are always associated with the page on which they appear.

The availability of article-level metadata and more accurate OCR text output exponentially increases the points of access into the collection which the TMPF can provide to users. The newspaper demonstration sites provide access to both the access images and to the transcribed text. They also give users the option to hide the transcribed text if they are only interested in viewing the images.

The TMPF for newspaper delivery supports browsing of the full list of newspaper titles by title and by date. Users can browse within a selected newspaper issue and navigate from page to page, from any page to the first page of the issue, from any page to the issue web page, and

[10] Although the TMPF used to deliver the NZETC collection

from any page to the publication web page. Users can also browse from article to article on a given page or in a given article. This is enabled through a topic map which includes topics for a newspaper series (e.g. “The Times”), a newspaper issues (e.g. “The Times 18th June 1942”), pages, articles, article images, publishers and dates. As with the TMPF used in production for the NZETC collection this very simple ontology for the newspaper domain is encoded as an extension of the much more complex and sophisticated CIDOC CRM.

In the near future we expect to include support for browsing the issues of a selected newspaper by date through an interface similar to the calendar pages and browsing by geographic region of newspaper coverage TMPF already harvests publication date information in its topic map, including relating the dates to a calendar of years, decades, and centuries. Exposing these date topics for browsing will require only minor additions to the user interface layer. Although TMPF already includes publication place information in its topic map, this is not quite the same as geographic coverage. However, only minor development work would be required to harvest geographic coverage from newspaper source files and present a browsable user interface, including using a clickable map of NZ. In fact, because TMPF uses a topic map to define browse points, any and all metadata harvested into the topic map (whether from the newspaper source materials or from external data sources) can be used as the basis for browsing the newspaper collection.

To enable users to view the access images of the newspaper pages and articles we used “Zoomify” technology to provide zooming and panning functionality. The free Zoomify EZ encoder creates multiple copies of an image at many resolutions tiers, from the original source resolution down to a thumbnail. Each tier is then cut into many small tiles. All the tiles from all the tiers are combined into a folder of JPEG files with an index of the exact location of every tile. Tile organization is pyramidal in that tiles are stacked from a thumbnail down to the highest resolution, tier upon tier. When the converted image is viewed, the Zoomify Flash Viewer requests tiles from the appropriate tier to fill the display area. Each zoom and pan requests only a small additional number of tiles: those at the level of zoom desired or for the part of the image panned to. These additional tiles are streamed on-demand, to the viewer. No tiles are ever delivered unless required for the current display, or for a display that is anticipated to immediately follow (intelligent pre-fetching).

Figure 5 Pyramidal Tiled Multi-Resolution Image from Zoomification Process. Taken from ***http://www.zoomify.com***

As the screenshot below shows, the user has the ability to zoom in and out of the image, and to pan left, right, up and down around a page or an article. The thumbnail in the top left hand corner is a further navigational aid in that it informs the user of the location of their current view.

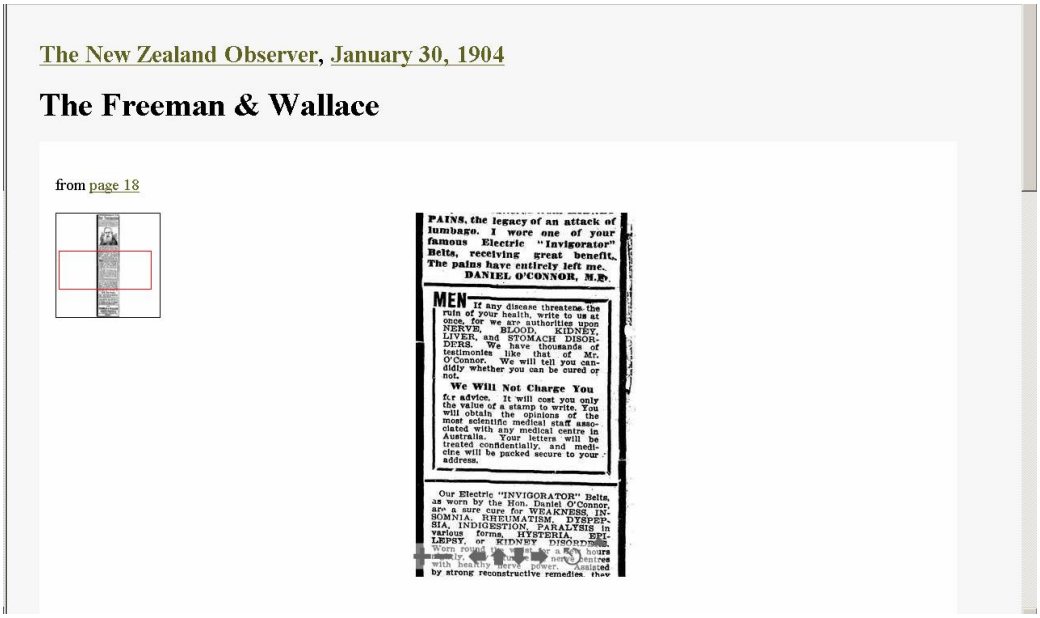

Figure 6 Screenshot of newspaper article in TMPF demo site (Newspaper article from the National Library of New Zealand)

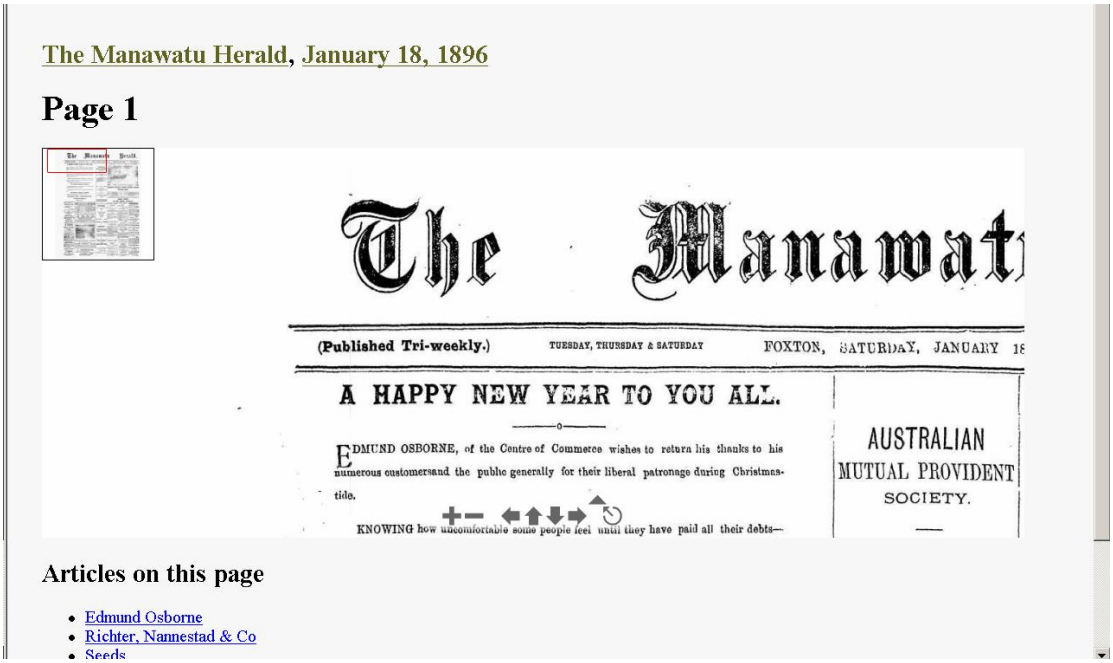

Figure 7 Screenshot of newspaper article in TMPF demo site (Newspaper page from the National Library of New Zealand)

To enable searching over the newspaper archive we used the Apache Lucene[11] search engine, again as we do for the TMPF in production at the NZETC. It provides all the search functionality discussed in the "Delivering Newspaper Content" section above including an optional module for synonym expansion in search queries is using the WordNet thesaurus. Users can limit their search to a particular newspaper title and / or by a date range.

[11] Apache Lucene http://lucene.apache.org/java/docs/index.html

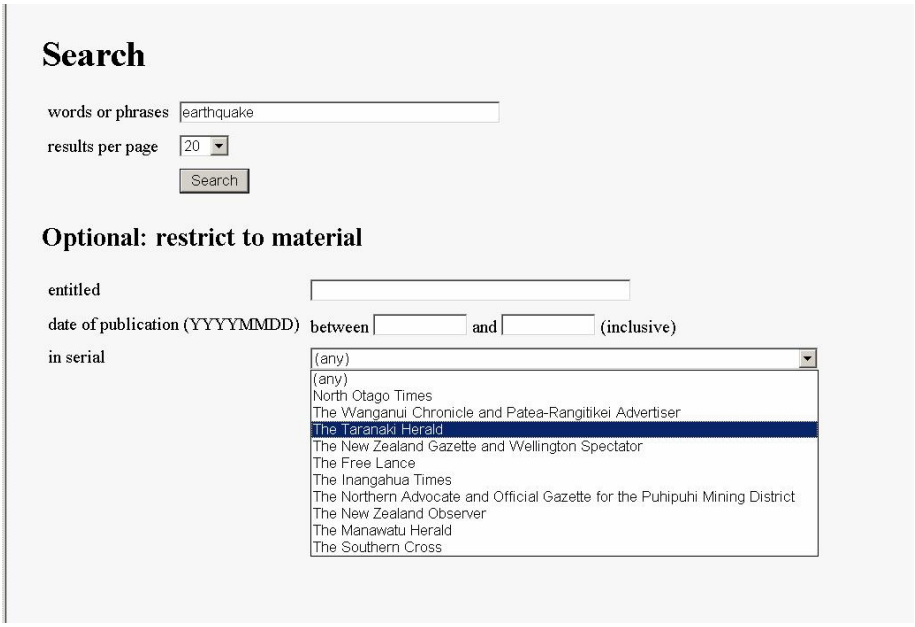

Figure 8 Screenshot of Search Page from TMPF demonstration site (Newspaper titles from the National Library of New Zealand)

Search results are displayed as a list of links to relevant resources with large result sets split across several pages. The user can set the number of results displayed on each query result screen, and navigate easily between result screens. Result pages are also bookmarkable. The Lucene search engine uses a sophisticated relevance ranking algorithm to sort results. Article titles and other important or quality-assured metadata can be boosted in importance when building the Lucene search index, so that Lucene accords hits on these fields a greater weight. Lucene's query syntax also allows users to boost the priority of individual words or phrases in their queries, so that documents containing those terms are regarded as more relevant than unboosted terms.

Background information describing the history of each newspaper can be included on the appropriate publication web page. Such a background article can be written either directly in HTML or in XML. Such an article is linked to the newspaper it describes by using the persistent identifier of the newspaper as the subject of the article, in the article's metadata header. It is important to note that the exact same technique can also be used to associate background information with individual newspaper issues, pages, or articles.

When missing pages or printing anomalies are encountered in the source material the TMPF system handles different types of omissions in different ways. If there is a story behind the omission, this story can be encoded as another text and linked to an appropriate part of the collection: the newspaper series, individual issue, article or page. TMPF would then include and display this explanatory material in the appropriate context. Alternatively, where content is missing from inside a document, the omission may be encoded using appropriate XML mark-up in the source material (such as the <gap> element in the TEI mark-up language) to describe the missing content or explain the omission. Where entire issues are missing, XML files empty of content can be used as placeholders for the missing material. Errors and anomalies in the source materials can be explicitly marked as anomalies using appropriate XML mark-up, such as the <sic> and <corr> elements in the TEI mark-up language.

The TMPF approach means a wide range of textual materials can be included in the delivery system, such as journals and monographs (e.g. pamphlets). TMPF is currently used by the NZETC to present its digital collection which includes books, journals, letters, and pamphlets. TMPF is fully extensible to handle a potentially infinite variety of materials. Different document types can be presented in distinct ways, or, to the extent that they can be interpreted in terms of a common conceptual model, they can then be presented in an identical fashion, to provide

a consistent user experience regardless of the different document formats, encoding practices, and storage technologies. TMPF is designed to be easily extensible to other data types, metadata schemas, and knowledge domains. The use of a topic map for the central metadata repository in TMPF provides an open-ended framework for importing, mapping and meaningfully presenting information from a number of distinct information systems.

TMPF can import and merge pre-existing metadata from other sources. The NZETC has used TMPF to import metadata records from other sources and merge them with metadata describing TMPF's own collection. To import and merge metadata from external sources, the metadata should be exported in some XML format, and an XSLT transformation is used to extract metadata in the common XML Topic Map (XTM) format, which is then imported and merged. Merging of metadata records in general requires only that items of interest can be identified by a URI. For newspapers this might be an ISSN, DOI, URN, or simply an HTTP URL. The topic map metadata repository in TMPF can record mappings between different name authorities and perform cross-walks between sets of metadata using those authorities. All metadata records for resources with a particular identifier are automatically merged. Merging behaviour is a key part of the specification of the Topic Map standard and is a built-in feature of the topic map metadata repository component of TMPF.

In general ease of use is the result of a well thought-out and robust system architecture that lies beneath the interface. This is a characteristic of TMPF which provides a 100% customisable interface providing complete control over the delivery system's look and feel. In designing the interface of TMPF for newspaper delivery we aimed for something which is uncluttered, requires minimal keying and clicking, and minimal opening of new windows. TMPF accommodates context sensitive help by assigning help documentation to any topic in the system. Help can be authored as one or more XML documents and linked to the relevant part of the system by adding subject classification metadata (i.e. the help document is itself tagged as being "about" a class of topic, such as monthly publication histories, newspaper issues, articles, pages, etc.) TMPF allows the use of regular web browser features as users expect. TMPF does not encode session identifiers in URLs, which is a common obstacle to bookmarking. In TMPF, the URLs of web pages are entirely independent of the storage and retrieval mechanism for content. TMPF can be configured to use web page URLs conforming to any desired convention. Other potential obstacles to regular navigation involve client-side JavaScript for linking and the use of frames and pop-up windows. Though TMPF does not prevent the use of these techniques, TMPF has not used them and does not generally recommend them.

The TMPF interface is completely customisable and supports integration of links to other services. Such links may also be classified into different types (e.g. "further reading," "discussion forums," "annotations," etc.), and each type of link presented independently.

These links may be further classified in various ways:

- by access/visibility (e.g. "public," "internal," "QA")
- by perspective (e.g. "historical," "geographical")
- by provenance ("scholarly," "user-contributed," "Te Puna," "Te Papa," "Ministry of Culture and Heritage")
- by natural language (e.g. "English," "Māori")

Multiple user interfaces can present filtered views of these links appropriate for particular audiences.

TMPF is built out of XML-based components; hence it is based entirely on Unicode. This provides the ability to represent all characters in Māori and other Polynesian languages.

As for our TMPF production system at the NZETC, we used Apache Cocoon to transform the XML texts created by APEX into readable documents using XSLT stylesheets. Cocoon can deliver documents in a variety of formats, including HTML, PDF, RTF, SVG, JPEG, PNG, and any other XML-based format. We can also integrate software to produce Microsoft's eBook Reader format.

Cocoon can perform these transformations on demand; i.e. when a request is received from a web browser. Each request is handled by reading the appropriate XML document or documents, and processing the XML data in a succession of stages, first applying logical, then presentational transformations. Each stage is distinct and can be effectively managed by different people. Our web designer can edit the look of the site, the web developer can edit the structure of the site, and the text-editors can edit the content of the site (the e-texts), all independently of each other. To install a new text, the editors can simply upload the XML document and associated image files into the webserver via FTP. The document will then be automatically converted to HTML and divided into separate pages for each chapter, and scaled-down thumbnail versions of the JPEG graphics will be created using the XML graphics format SVG. To change the overall look of the site, the web-designer can upload new design elements such as CSS stylesheets, new versions of the logo, navigation menu, etc, in the same way. When a document is displayed to the reader, the content will be automatically inserted into this new design.

Apache Cocoon is a Java servlet and hence it can be deployed on a wide variety of systems. At the NZETC we run Cocoon inside the Apache Tomcat servlet container (the official reference Implementation for the Java Servlet specification), using JVM version 1.4 from Sun Microsystems.

Conclusion

There are currently numerous large projects going on around the world which aim to create online newspaper archives. So far, much of the public, technical discussion around these projects has focused on the digitisation process – the use of microfilm for scanning, OCR requirements and techniques, file naming conventions, image format choices, storage strategies, and so on. Since the imperative for digitisation is often as much about preservation as it is about access, this is not surprising.

However the delivery of content from the newspaper archive to users is a similarly important process which requires a similar level of commitment to technical research and development if it is to be successful. Of course the successful delivery of newspaper content over the web is predicated on the existence of a high quality, well-structured digital collection which to deliver. No amount of sophisticated search algorithms, novel browsing functionality or intuitive interface design can compensate for inaccurately transcribed text, poor quality images or inaccurate metadata. But it is equally true that the impact and usefulness of even the most interesting, high-quality content will be diminished if that content is not discoverable, navigable and presented in a way that meets user needs.

The TMPF has been developed within academic communities and has not enjoyed the benefits of high-level marketing and promotional efforts. However the semantic framework that it applies to the newspaper domain presents a possible means to fulfil some of the complex requirements which must be met to ensure the successful online delivery of a newspaper archive.