Journal of the Text Encoding Initiative

Issue 5 (June 2013)
TEI Infrastructures

Kirsta Stapelfeldt and Donald Moses

Islandora and TEI: Current and Emerging Applications/Approaches

Electronic reference
Kirsta Stapelfeldt and Donald Moses, « Islandora and TEI: Current and Emerging Applications/Approaches », *Journal of the Text Encoding Initiative* [Online], Issue 5 | June 2013, Online since 24 June 2013, connection on 15 April 2015. URL : http://jtei.revues.org/790 ; DOI : 10.4000/jtei.790

Publisher: Text Encoding Initiative Consortium
http://jtei.revues.org
http://www.revues.org

Document available online on:
http://jtei.revues.org/790
Document automatically generated on 15 April 2015.

Kirsta Stapelfeldt and Donald Moses

Islandora and TEI: Current and Emerging Applications/Approaches

1. About Islandora

1 Developed at the Robertson Library, University of Prince Edward Island, Islandora (Wilcox 2012) is a robust open-source digital asset management system that can be used anywhere that collaboration and digital data stewardship, for the short and long term, are critical. Islandora integrates the Drupal content management system (Drupal 2012) and the Fedora digital repository with a host of functions that:

- provide a public-facing website while at the same time providing a secure private workspace for authenticated members of the group
- enable users to easily customize and brand their website
- provide a set of collaboration and communication tools (e.g., document sharing and editing, data collection, blogs, RSS feeds, tagging, and commenting)
- enable authorized users to manage (create, edit, or delete) any type of digital content in a secure repository
- provide a forms-based interface for users to describe their digital content using the standards of their discipline (for example, describing herbarium samples using the Darwin Core XML schema)
- provide users with discovery tools (search and browse) to display and analyze their content
- provide custom transformation or views of content
- provide best-practice digital preservation services

2 At UPEI, Islandora is used across the entire enterprise —in learning, research, and administration. In 2010, the Atlantic Canadian Opportunities Agency funded a $2.4 million project to help establish the open-source community surrounding Islandora, build the codebase, and release a number of tools ("solution packs") that would provide key functions for different knowledge domains or data management tasks. This grant also helped launch a services company for the software, called discoverygarden inc. (also in 2010). Since 2006, operational funding at UPEI has been dedicated to the management of the codebase. In addition to the work of UPEI staff, all software developed by discoverygarden inc. is released to the public; at the time of writing, the Islandora users' and developers' Google Groups combined have over 500 members, and the project's GitHub account has over 40 repositories.

3 The Islandora website contains additional information about Islandora, its community, software updates and documentation. From the site visitors can experience logging into and administering an Islandora 6 or Islandora 7 installation.

2. System Details

4 Islandora represents a modular approach to system development. At its core are two open-source projects: the Drupal content management system and the Fedora Commons Repository software, which together create a robust digital asset management system that can be used to meet the short- and long-term requirements of digital data stewardship in a collaborative context.

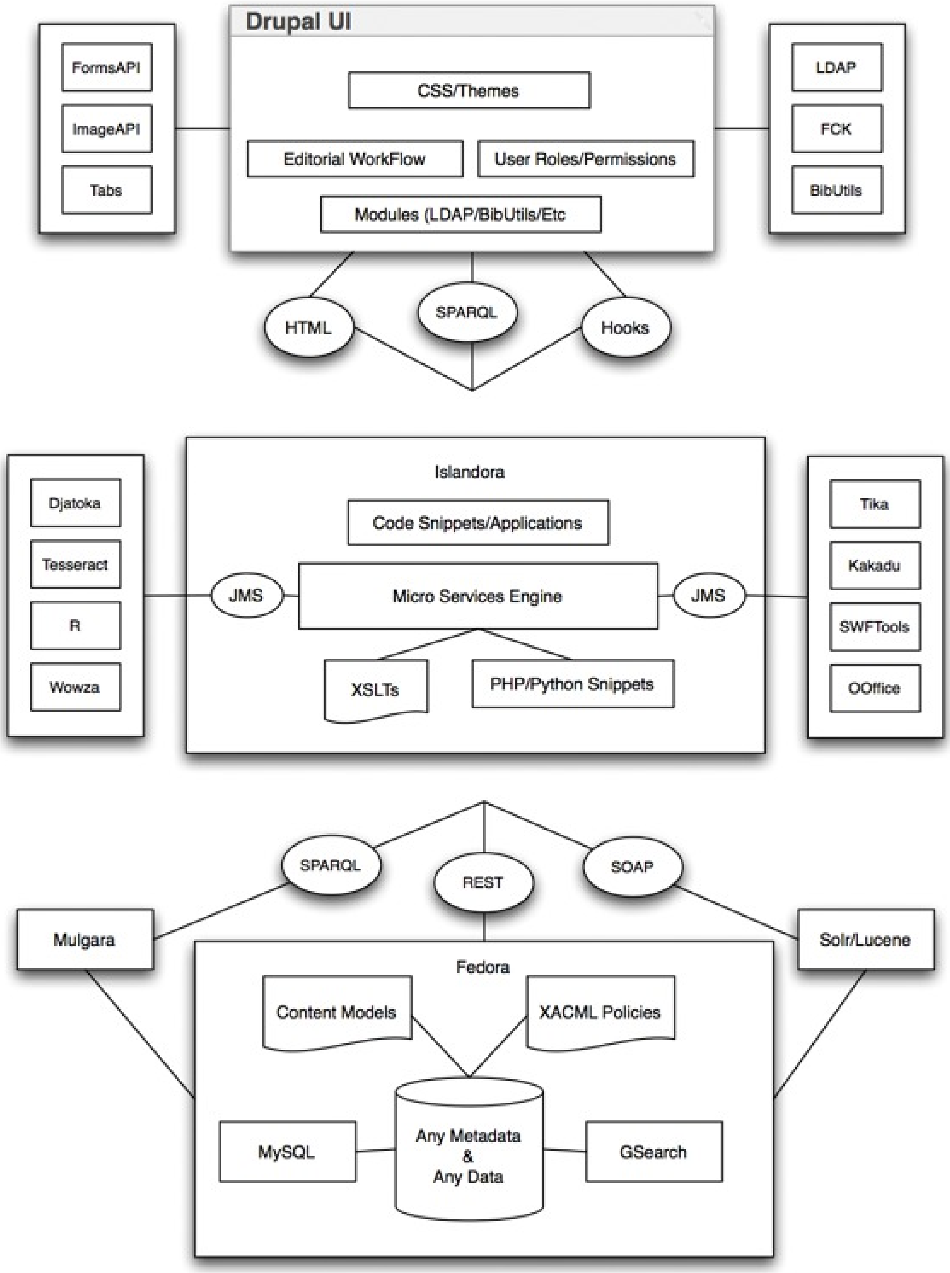


Figure 1. The Islandora architecture

The Drupal content management framework provides a highly customizable interface for collaborative content authoring and publication. It is widely used in the digital humanities community: for example, in TEICHI (Pape, Schöch, and Wegner 2012), *Saint Patrick's Confessio*,[1] and in tutorials such as *Drupal for Humanists*.[2] There are over 20,000 modules available for various versions of Drupal that extend the functionality of a Drupal site; the Islandora project maintains a core Islandora module, and Islandora solution pack modules for both Drupal 6 and Drupal 7. These Islandora modules allow a user to administer an Islandora site in a way that feels much like administering a Drupal site. By utilizing the Islandora modules, implementers can use Drupal to discover and manage assets in a Fedora Commons repository. A Fedora Commons repository provides complex data modeling and datastore better suited to knowledge sharing and long-term data preservation than Drupal. Fedora provides not only web-based services and APIs but also support for metadata sharing and RDF (via a Mulgara triplestore). Apache Solr provides the ability to search and discover Fedora assets in a way that respects Fedora's security layer, as well as providing current best-practice features such as faceted search results. To this base stack a number of additional applications have been added or integrated, including the Tesseract and ABBYY OCR engines and the Djatoka Jpeg 2000 Image Server. Overall, Islandora integrates a number of robust open-source software projects.

6 Additional software can be integrated into an Islandora site via solution packs, which provide features such as content models (which prescribe the nature of a piece of content and its relationship to other pieces of content), ingest workflow and file processing (via modular microservices), metadata management, search profiles, and custom viewers. Other solution packs include utility modules that provide tools for repository managers and other administrative staff: for example, Islandora's batch ingest or workflow solution packs. Islandora also supports alternative discovery strategies to searching and browsing. An example of this is a visual chemical search that allows the user to draw a molecule (or part of it), with the system returning to the user the molecule expressed as a text-based formula, as a 2-D image representation, or a rotatable 3-D object.

3. Modeling Book-Style Content in Islandora

7 To further illustrate the Islandora framework, particularly in the context of digital humanities research, and to introduce the integration of TEI tools into Islandora for IslandLives, the following section discusses the way that content is understood, stored, searched, and displayed in Islandora. Each data asset in Fedora is referred to as an object, and contains two or more datastreams (or files) that are relevant to an object. A "content model object" defines the nature of a data asset object and its relationship to other data asset objects. The resulting atomic approach (which heavily leverages XML and XSLT) allows data structures to be more transparent, reusable, and transferable to other application environments, which is why this model is considered appropriate for the long-term stewardship of digital data.

8 Let us illustrate using the example of a book as it is modeled by the "Book Solution Pack"—not because TEI is only useful when describing a book, but because the book model illustrates a more complex data modeling procedure. A book encountered on a shelf is experienced as a single, self-contained object; however, the same book, when digitized and placed in the Islandora repository, is modeled as numerous objects. The "book object" is defined by the "content model object," and it contains descriptive metadata (the bibliographic record), a thumbnail image considered representative of the book, and RDF statements linking the book to a particular digital collection. The bulk of the book's content—such as OCR'd text, TEI-encoded text, image/text annotations, and high-resolution images of the page—is stored with page objects. These page objects are related to the book object via RDF data, which declare the sequence of the page in the book. This relationship is, again, specified by the content model object for a book and for a page. Pages might belong to multiple books, and books to multiple collections. These RDF statements are written to a datastream known in Fedora as RELS-EXT (represented in fig. 2), which illustrates the relationships written between discrete repository objects to represent the concept of a book. Additional information can be found in the Islandora documentation (Wilcox 2012).

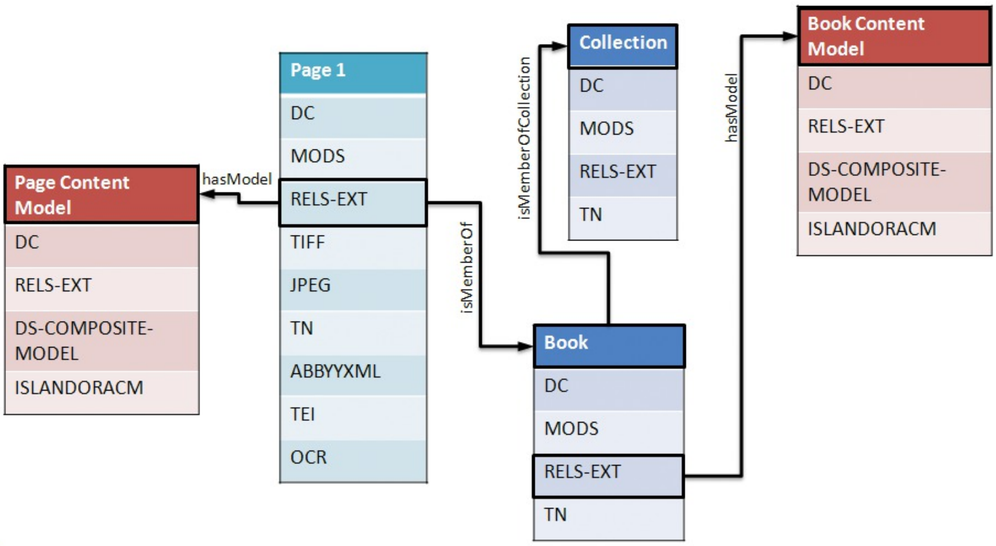


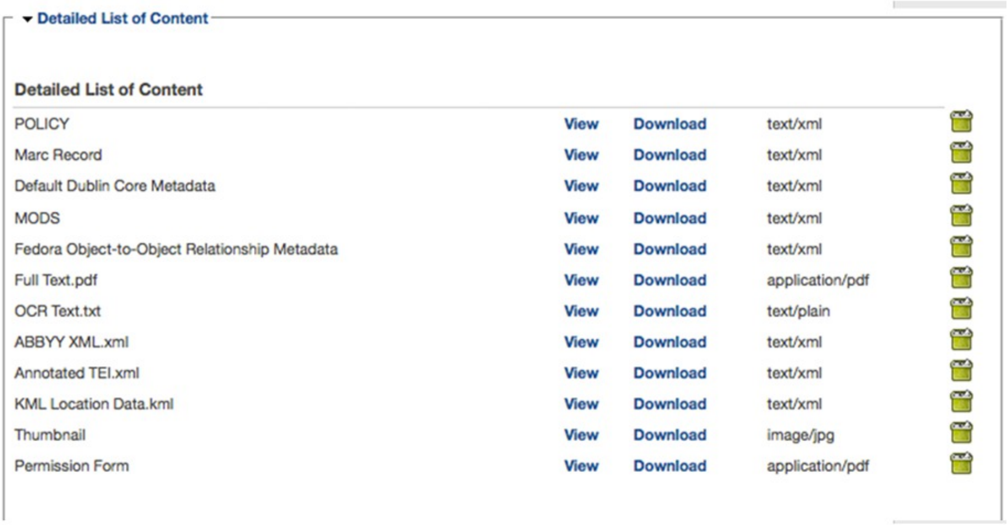
Figure 2. All digital objects in Islandora contain multiple datastreams defined by content models. One of these datastreams (RELS-EXT) stores RDF statements that describe the relationship between a given object and others in the repository. This object-oriented data modeling is flexible, non-hierarchical, and lends itself to atomistic (or highly granular) data models, and, as a result, to complex ontologies.

4. Modeling and Storing TEI in Islandora

9 Drupal's book module can be modified to store TEI data. Defining a TEI datastream within an Islandora content model object (which defines the existence of TEI in the data asset objects that comprise the digital book) is a fairly straightforward procedure: the content model object is modified to define a TEI datastream, which can then be created, read, updated, and deleted by any user with appropriate permissions. From the perspective of data modeling, the challenge is to reconcile the fragmentation of the book into separate pages with the common need to produce and edit single TEI documents which represent an entire book. In other words, if a page's OCR'd text (and TEI-encoded text) are stored in fragments, how can users access a single continuous TEI file, or upload a continuous TEI file that corresponds to multiple pages?

10 The IslandLives project chose to resolve this challenge by duplicating TEI datastreams at the page and book levels. On the one hand, this was a pragmatic decision having to do with system resources—presenting a 500-page TEI document for editing and saving is an expensive computing procedure—as well as with the problem of how to present relevant page-object images alongside the corresponding TEI data for OCR and for encoding correction. On the other hand, this was also a data modeling decision. The Islandora model is designed to keep pages as discrete packages of data (to "atomize" content), and there was the desire to maintain this approach. Apache Solr and the Islandora Solr client allow for selective indexing of datastreams and elements, so there was little concern with contaminating the search index with duplicate data or providing confusing results for users.

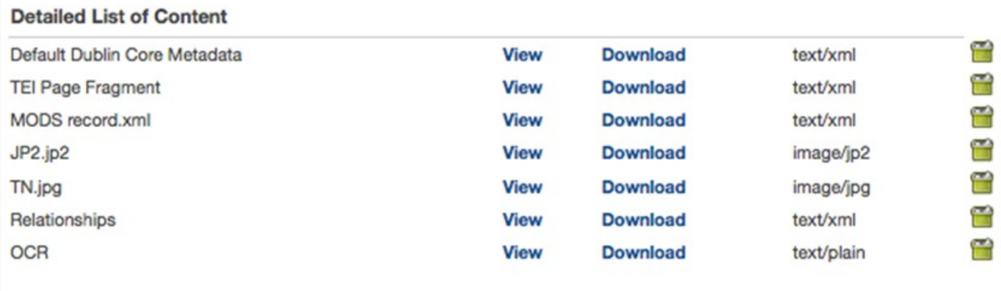
11 In order to provide an overview of the data model in IslandLives, figures 3 and 4 illustrate both the book-level ("Annotated TEI") and page-level ("TEI Page Fragment") datastreams. What these screenshots do not reflect is that the TEI header defined in the book object was duplicated into the page-level TEI fragments.



Detailed List of Content

POLICY	View	Download	text/xml
Marc Record	View	Download	text/xml
Default Dublin Core Metadata	View	Download	text/xml
MODS	View	Download	text/xml
Fedora Object-to-Object Relationship Metadata	View	Download	text/xml
Full Text.pdf	View	Download	application/pdf
OCR Text.txt	View	Download	text/plain
ABBYY XML.xml	View	Download	text/xml
Annotated TEI.xml	View	Download	text/xml
KML Location Data.kml	View	Download	text/xml
Thumbnail	View	Download	image/jpg
Permission Form	View	Download	application/pdf

Figure 3. Datastreams at the book level: A TEI datastream representing all book pages is stored at the book level in the IslandLives project.



Detailed List of Content

Default Dublin Core Metadata	View	Download	text/xml
TEI Page Fragment	View	Download	text/xml
MODS record.xml	View	Download	text/xml
JP2.jp2	View	Download	image/jp2
TN.jpg	View	Download	image/jpg
Relationships	View	Download	text/xml
OCR	View	Download	text/plain

Figure 4. Datastreams at the page level: TEI page fragments are stored at the page level. This stream is used for editing, display, and page-level search resolution.

12 Both the image viewer and the TEI editor are associated with content at the page level, and the book-level TEI document was generated once by a script in order to support those who wish to download or view the complete TEI stream or for use with further transformations like creating an EPUB file from the TEI. This data model poses obvious synchronization problems since any additional work at the page level will not be reflected in the book-level TEI document. In addition, this model does not address the common use case where a project has produced a number of page-level image files and a single continuous TEI document. In the IslandLives project, neither of these issues is particularly pressing. However, both are roadblocks to generalizing for multiple projects.

13 A potential solution to address synchronization between page and book level TEI is to regenerate the book-level TEI document whenever page-level TEI files are edited and saved. A script, to be run by an administrator or run in a nightly cron job, would concatenate all the pages related to a book into a consolidated TEI document and replace the TEI document stored at the top level of the book. Having a corrected book-level TEI document would add a great deal of value, allowing the project to expose its book-level TEI documents. In this case, a page's OCR'd text (and TEI-encoded text) would be stored in fragments, and a user would be able to access a single continuous TEI file, created on demand. However, this is not how IslandLives currently works.

14 The second issue is one of ingesting. In general, uploading a continuous TEI file that corresponds to multiple pages at the time of book object creation remains an issue since a book is comprised of multiple page level objects. However, it is common to encounter projects that have produced multiple page-level images and single, continuous, TEI documents. A script could be created to divide a TEI document for a single repository, but such a generic tool that allows for the uploading of TEI documents and their alignment with digitized assets may pose a problem if different approaches are taken toward encoding for page breaks for different objects.

15 A potential solution to the challenge of addressing page breaks is to use the sequencing information stored with page objects and, on upload, dynamically update the facs attribute of <pb> elements to associate them with the appropriate image URLs in the repository. This approach would assume that a TEI digital object (adhering to a TEI content model in Islandora) is associated with an image digital object (adhering to an Islandora's image content model). The resulting fragments would have to be adequately addressed in the concatenation script so that the complete TEI document would be re-usable outside the repository—another common requirement for TEI systems. The IslandLives approach evolved to address a specific set of requirements, and this evolution is described in the following section.

5. The IslandLives Project: Automating TEI Encoding

16 The IslandLives project began in 2008, with the Robertson Library collecting permissions to digitize and digitally publish a number of the Prince Edward Island local histories in its collection, forming the core of the IslandLives collection. IslandLives is one of many projects illustrating the Robertson Library's commitment to preserving and sharing unique material relating to Prince Edward Island with students, educators, researchers, and others interested in Island culture and heritage.

17 In IslandLives, the TEI encoding was not created by hand but through an automated series of processes developed by the IslandLives team. The books were scanned, and then run through ABBYY FineReader OCR software. ABBYY can produce a number of output formats; ABBYY FineReader XML, the output of the IslandLives digitization process, includes both the OCR text and structural information about the document, including the location of images, lines, and the sizes of pages and paragraphs. For long-term preservation, ABBYY's proprietary format had to be transformed into a recognizable, open standard, and TEI was the clear choice.

18 The IslandLives team used an XSLT stylesheet to transform the ABBYY FineReader XML tags into simple TEI structural tags. Most of the OCR coordinates in the ABBYY XML are lost in the transformation to TEI, so the current IslandLives code, and earlier versions of the book solution pack in Islandora 6, are unable to support highlighting of hits in page images. Current versions of the book solution pack, which uses Tesseract, preserve OCR coordinates as part of the hOCR (Breuel 2010) output. Development of an XSLT stylesheet to transform hOCR to TEI is currently underway. One challenge facing the project is how to edit and update text while preserving page coordinates: currently, the OCR coordinates cannot be corrected in Islandora. A small sample of the ABBYY FineReader XML input, and the resulting TEI output, is provided in example 1.

ABBYY XML (before)

```
<block blockType="Text" blockName="" l="94" t="524" r="1484" b="2092">
  <region><rect l="94" t="524" r="1484" b="2092"/></region>
    <text>
     <par align="Justified" leftIndent="2" rightIndent="3" startIndent="52"
           lineSpacing="71">
          <line l="166" t="536" r="1461" b="589">
              <formatting lang="EnglishUnitedStates">Setting my pen to write
               "Pioneer Days & Shanty Ways" gave</formatting>
          </line>
          <line l="114" t="608" r="1462" b="661">
              <formatting lang="EnglishUnitedStates">me a special reason for
               choosing the title for the book. Having</formatting>
          </line>
          <line l="112" t="679" r="1459" b="733">
          <formatting lang="EnglishUnitedStates">a pioneer fisherman father,
               who told us many stories of the sea,</formatting>
          </line>
          <line l="115" t="751" r="1458" b="805">
               <formatting lang="EnglishUnitedStates">I felt would be an
               interesting place to start. Having a farm girl</formatting>
          </line>
          <line l="112" t="823" r="1463" b="876">
           <formatting lang="EnglishUnitedStates">as a mother, whose stories
    were passed down from her</formatting>
</line>
<line l="111" t="896" r="1462" b="949">
   <formatting lang="EnglishUnitedStates">grandfather's knee, was
    the opportunity I needed to fill the book</formatting>
</line>
<line l="111" t="966" r="1463" b="1020">
    <formatting lang="EnglishUnitedStates">with many treasured memories
    of family life back then. </formatting>
</line>
 <div n="d1e657" rend="94,524,1484,2092">
```

Final TEI (after)

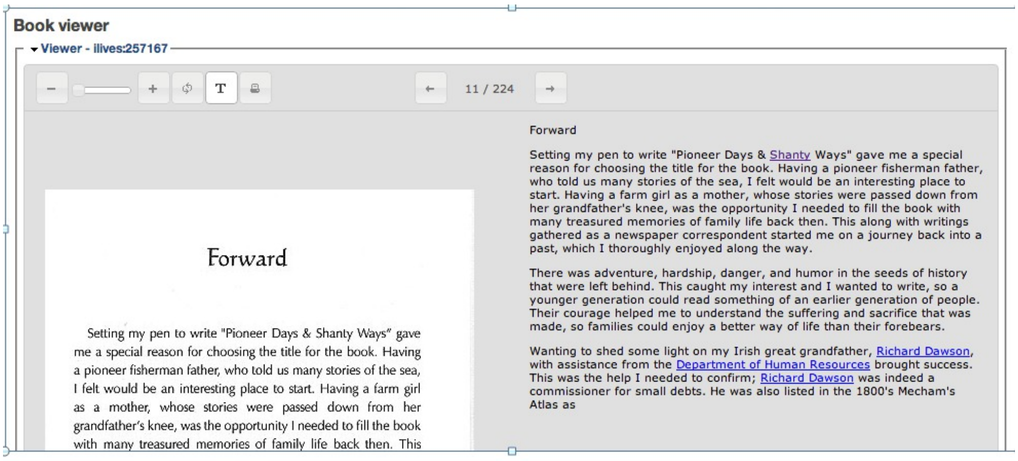
```
<p n="d1e663">
     Setting my pen to write &quot;Pioneer Days &amp; <orgName>Shanty</orgName>
    Ways&quot; gave me a special reason for choosing the title for the book.
     Having a pioneer fisherman father, who told us many stories of the sea,
     I felt would be an interesting place to start. Having a farm girl
     as a mother, whose stories were passed down from her
     grandfather&apos;s knee, was the opportunity I needed to fill the book
     with many treasured memories of family life back then. </p>
 <p n="d1e706">
```

Example 1. During the automated IslandLives process, ABBYY XML (above) becomes simple TEI encoding (below).

19 Beyond just OCR'ing of the page images, the IslandLives team decided to attempt an automated semantic encoding of place names, personal names, dates, and organizations mentioned in the texts to allow for richer connections to be made across the multiple histories. Tools from the General Architecture for Text Engineering (GATE) project (Cunningham et al. 2002) were used with specially created controlled vocabularies of names for this process. The sources of information for these gazetteers varied: for example, given and family names were captured from the Prince Edward Island Baptismal Index,[3] organizational names were derived from corporate subject headings in the library's catalog, and place names came from the Canadian Geographical Names Data Base.[4] Place names are stored in the project's controlled vocabulary alongside their CGNDB database keys, allowing for data associated with the place name (such as the latitude and longitude coordinates) to be retrieved dynamically from the external data source. With sources for semantic and structural information in place, the only remaining piece was the TEI header. All of the books in IslandLives exist in the Library's online catalog, which allows for the export of bibliographic records in MARCXML format. Team members created a crosswalk from MARCXML fields to TEI header elements in order to create headers.

6. The Encoding Tool Chain

20 In April 2009, IslandLives team member wrote scripts that automated the processes described above and the conversion of ABBYY XML to valid TEI XML, generating pages as HTML. During this conversion process, semantic tags are turned into links that, when clicked, will launch a relevant repository search (via a Fedora disseminator) of all works in the collection that contain the same tag (see fig. 5).



Forward

Setting my pen to write "Pioneer Days & Shanty Ways" gave me a special reason for choosing the title for the book. Having a pioneer fisherman father, who told us many stories of the sea, I felt would be an interesting place to start. Having a farm girl as a mother, whose stories were passed down from her grandfather's knee, was the opportunity I needed to fill the book with many treasured memories of family life back then. This

Forward

Setting my pen to write "Pioneer Days & Shanty Ways" gave me a special reason for choosing the title for the book. Having a pioneer fisherman father, who told us many stories of the sea, I felt would be an interesting place to start. Having a farm girl as a mother, whose stories were passed down from her grandfather's knee, was the opportunity I needed to fill the book with many treasured memories of family life back then. This along with writings gathered as a newspaper correspondent started me on a journey back into a past, which I thoroughly enjoyed along the way.

There was adventure, hardship, danger, and humor in the seeds of history that were left behind. This caught my interest and I wanted to write, so a younger generation could read something of an earlier generation of people. Their courage helped me to understand the suffering and sacrifice that was made, so families could enjoy a better way of life than their forebears.

Wanting to shed some light on my Irish great grandfather, Richard Dawson, with assistance from the Department of Human Resources brought success. This was the help I needed to confirm; Richard Dawson was indeed a commissioner for small debts. He was also listed in the 1800's Mecham's Atlas as

Figure 5. Site visitor view of TEI: When a datastream is viewed by a user, the encoding appears as HTML links that launch Solr searches of other works in the collection containing the same TEI tag.

21 The automation that underpins this display takes the following steps:

1. A script retrieves the appropriate MARCXML record from the library catalogue and the output is transformed into a <teiHeader> using an XSLT stylesheet (MARC2teiHeader_UPEIv2.xsl) and stored in a temporary file for use during the process.
2. A source ABBYY XML-encoded document is transformed using an XSLT stylsheet (ABBYY-to-TEIv2.xsl) into a TEI document (structural TEI only) and the <teiHeader> document is integrated into the resulting TEI file.
3. The TEI file is then run through GATE, which encodes people, organizations, dates, and places using its own specific encoding format. Here is an example of a place name with GATE encoding:
   <Location gate:gate gate:matches="38937;39350;39352" county="Queens" locType="community" rule1="Location1" rule2="LocFinal" type="UnincorporatedArea" key="BACDW">Rocky Point</Location>
4. The resulting XML document, with a mix of TEI and GATE tagging, is run through another XSLT stylesheet (GATE-to-TEIv2.xsl) that turns GATE tags into appropriate TEI tags. Here is the same fragment converted into TEI:
   <placeName key="BACDW"><settlement type="UnincorporatedArea">Rocky Point</settlement></placeName>
5. As a final step, an XSLT stylesheet (TEI-Split-Pages.xsl) is used to parse the TEI document and extract individual pages, each of which includes the <teiHeader> of the parent document. Each of these page-level TEI fragments is stored as a datastream with its related page object.

22 Three additional files are created during the process to assist with proofing:

1. A KML file that provides a map of all the place names and a snippet of related text mentioned in a work (created by TEI-to-KML.xsl).
2. An extract of all the terms encoded with the persName element. (created by ExtractPersname.xsl).
3. An html representation of the work (created by TEI-to-HTML.xsl).

23 The XSLT stylesheets referenced in the proceeding steps are available from the Robertson Library's GitHub repository.[5] A high-level view of the transformation process is offered in figure 6.

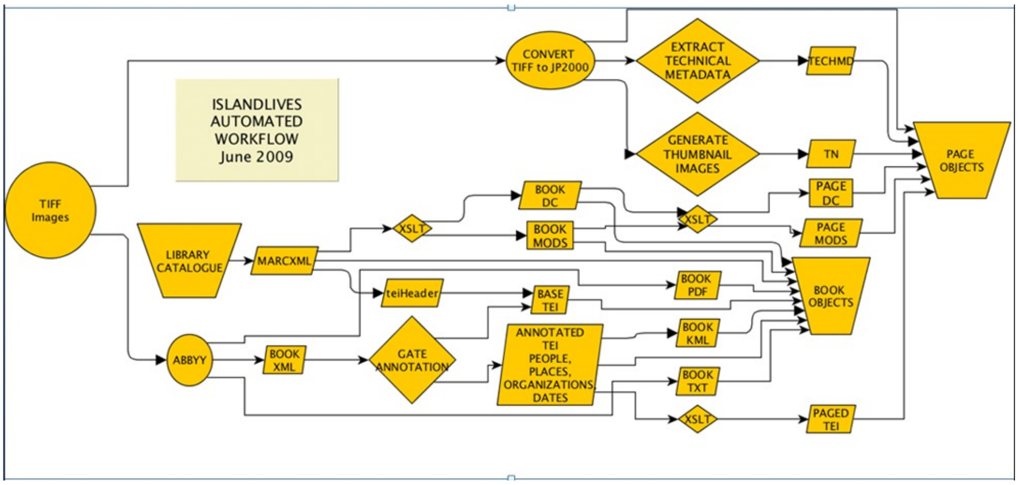


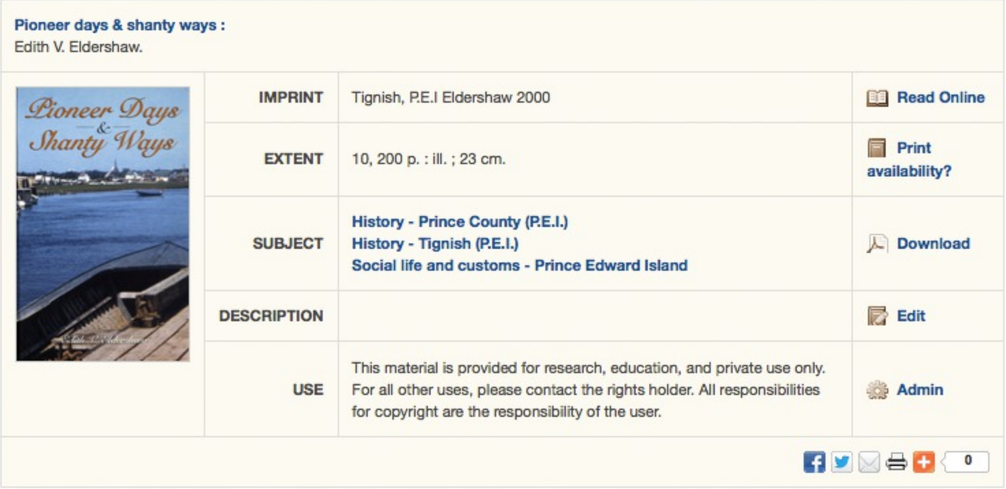
Figure 6. The IslandLives automated workflow takes the output of digitization and produces Islandora objects, including valid TEI data about the contents of book pages.

7. Automation Results and the TEI Editor

24 The IslandLives project team found that the structural TEI markup created by the automation process was much more accurate than the semantic markup created by the automation process. The automated semantic markup was more reliable when tagging place names than personal and organization names. While a certain amount of tuning may produce acceptable results for a semantically similar data set, the efficiency of the automation process for semantic markup is questionable. Where encoding is focused on a very specific goal (such as tagging place names) and on semantically similar data, there is some success. However, the more complex the scholarly work and the desired semantic encoding, the less likely that an automated approach of the kind employed by IslandLives would be satisfying.

25 In the IslandLives project, the errors in the automatically encoded XML needed to be corrected. To address this requirement, a group of fourth-year computer science students at the University of Prince Edward Island were approached to create an XML editor for the inserted semantic tags, which would provide administrative users with an interface to view the TEI encoding, add and modify tags, validate the document against a custom TEI schema, and update the relevant datastreams in the underlying repository. There was also the requirement for the editor to work alongside Islandora's original book viewer. After the initial work was complete, a developer was contracted to complete the work required to finish developing the editor. The editor is still in use for the IslandLives project.

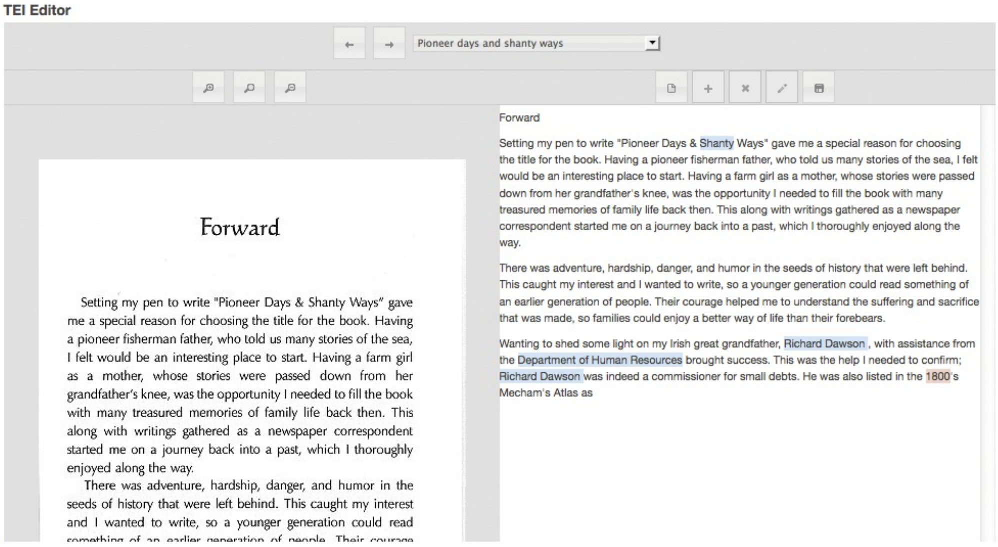
26 To access the TEI editor, a user logs into the site as an administrator and navigates to the relevant book record (see fig. 7). The "Edit" link provides access to the TEI editor.


Pioneer days & shanty ways :
Edith V. Eldershaw.

	IMPRINT	Tignish, P.E.I Eldershaw 2000	Read Online
	EXTENT	10, 200 p. : ill. ; 23 cm.	Print availability?
	SUBJECT	History - Prince County (P.E.I.) History - Tignish (P.E.I.) Social life and customs - Prince Edward Island	Download
	DESCRIPTION		Edit
	USE	This material is provided for research, education, and private use only. For all other uses, please contact the rights holder. All responsibilities for copyright are the responsibility of the user.	Admin

Figure 7. Administrator's view of a book record.

27 When administrative users select the "Edit" link, they are presented with a viewer for book pages, with the OCR text on the right-hand side. Semantic encoding is presented using a light-colored background (see fig. 8), the colors of which are configurable.



TEI Editor

Forward

Setting my pen to write "Pioneer Days & Shanty Ways" gave me a special reason for choosing the title for the book. Having a pioneer fisherman father, who told us many stories of the sea, I felt would be an interesting place to start. Having a farm girl as a mother, whose stories were passed down from her grandfather's knee, was the opportunity I needed to fill the book with many treasured memories of family life back then. This along with writings gathered as a newspaper correspondent started me on a journey back into a past, which I thoroughly enjoyed along the way.

There was adventure, hardship, danger, and humor in the seeds of history that were left behind. This caught my interest and I wanted to write, so a younger generation could read something of an earlier generation of people. Their courage

Forward

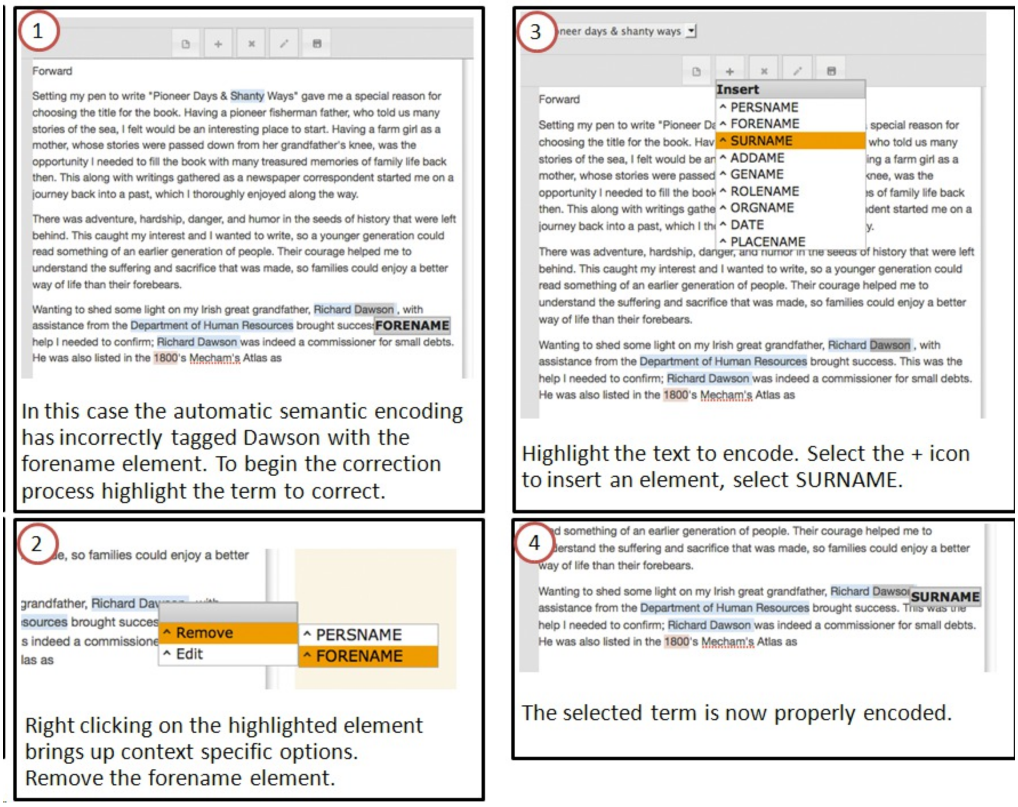
Setting my pen to write "Pioneer Days & Shanty Ways" gave me a special reason for choosing the title for the book. Having a pioneer fisherman father, who told us many stories of the sea, I felt would be an interesting place to start. Having a farm girl as a mother, whose stories were passed down from her grandfather's knee, was the opportunity I needed to fill the book with many treasured memories of family life back then. This along with writings gathered as a newspaper correspondent started me on a journey back into a past, which I thoroughly enjoyed along the way.

There was adventure, hardship, danger, and humor in the seeds of history that were left behind. This caught my interest and I wanted to write, so a younger generation could read something of an earlier generation of people. Their courage helped me to understand the suffering and sacrifice that was made, so families could enjoy a better way of life than their forebears.

Wanting to shed some light on my Irish great grandfather, Richard Dawson , with assistance from the Department of Human Resources brought success. This was the help I needed to confirm; Richard Dawson was indeed a commissioner for small debts. He was also listed in the 1800's Mecham's Atlas as

Figure 8. Beside each page, the editor presents the text of the page, with the TEI tags produced through automation highlighted.

28 The discovery and editing of an erroneous tag is depicted in figure 9.



Forward

Setting my pen to write "Pioneer Days & Shanty Ways" gave me a special reason for choosing the title for the book. Having a pioneer fisherman father, who told us many stories of the sea, I felt would be an interesting place to start. Having a farm girl as a mother, whose stories were passed down from her grandfather's knee, was the opportunity I needed to fill the book with many treasured memories of family life back then. This along with writings gathered as a newspaper correspondent started me on a journey back into a past, which I thoroughly enjoyed along the way.

There was adventure, hardship, danger, and humor in the seeds of history that were left behind. This caught my interest and I wanted to write, so a younger generation could read something of an earlier generation of people. Their courage helped me to understand the suffering and sacrifice that was made, so families could enjoy a better way of life than their forebears.

Wanting to shed some light on my Irish great grandfather, Richard Dawson , with assistance from the Department of Human Resources brought succes FORENAME help I needed to confirm; Richard Dawson was indeed a commissioner for small debts. He was also listed in the 1800's Mecham's Atlas as

In this case the automatic semantic encoding has incorrectly tagged Dawson with the forename element. To begin the correction process highlight the term to correct.

e, so families could enjoy a better
grandfather, Richard Da
sources brought succes
s indeed a commissione
las as

Right clicking on the highlighted element brings up context specific options. Remove the forename element.

Wanting to shed some light on my Irish great grandfather, Richard Dawson , with assistance from the Department of Human Resources brought success. This was the help I needed to confirm; Richard Dawson was indeed a commissioner for small debts. He was also listed in the 1800's Mecham's Atlas as

Highlight the text to encode. Select the + icon to insert an element, select SURNAME.

d something of an earlier generation of people. Their courage helped me to
derstand the suffering and sacrifice that was made, so families could enjoy a better
way of life than their forebears.

Wanting to shed some light on my Irish great grandfather, Richard Dawso SURNAME assistance from the Department of Human Resources brought success. This was the help I needed to confirm; Richard Dawson was indeed a commissioner for small debts. He was also listed in the 1800's Mecham's Atlas as

The selected term is now properly encoded.

Figure 9. A user discovers and edits a tag that has been improperly encoded in the tool chain.

29 For users needing to modify attributes and their values, an attribute editor is provided (depicted in fig. 10). For more advanced users, the TEI encoding can be edited manually (depicted in fig. 11).

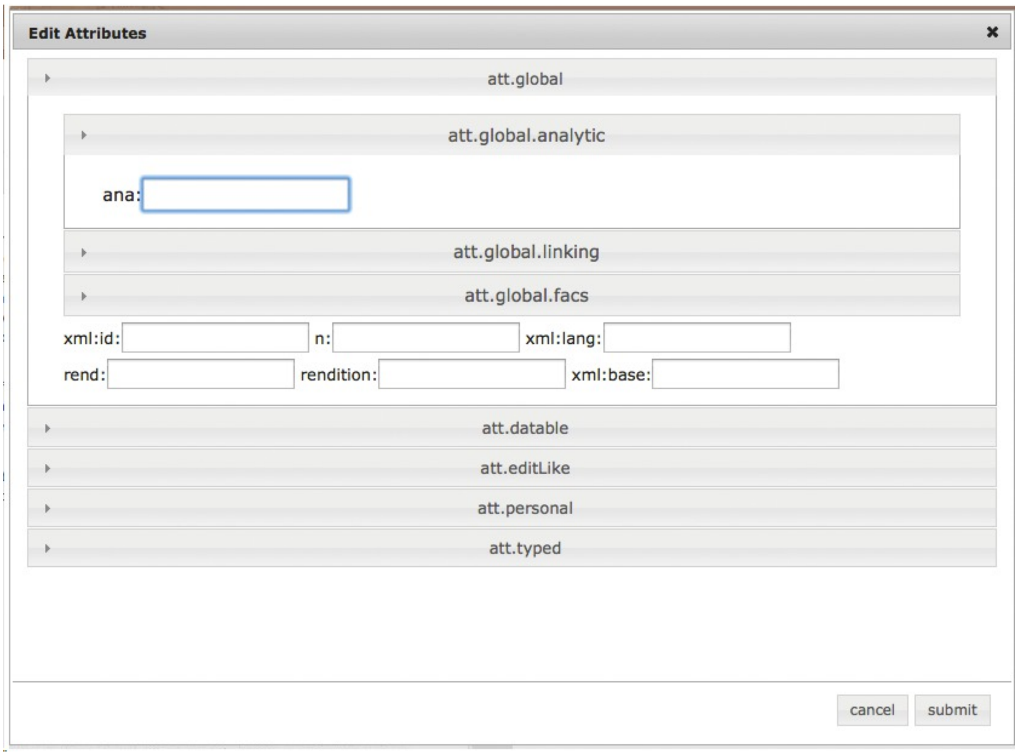
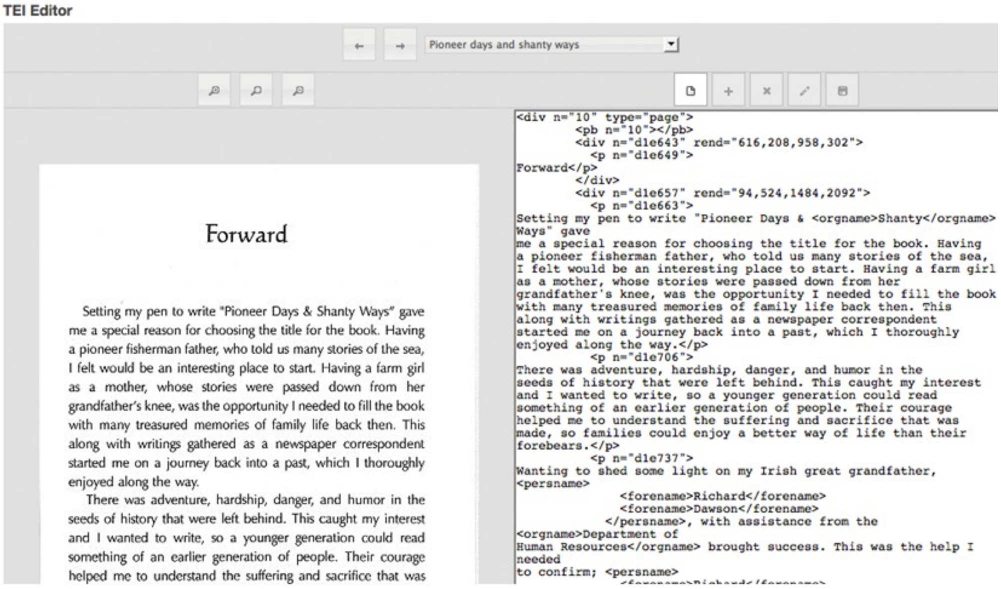


Figure 10. Additional flexibility is also provided via an attribute editor, which allows each tag to be enriched by an administrative user.



Forward

Setting my pen to write "Pioneer Days & Shanty Ways" gave me a special reason for choosing the title for the book. Having a pioneer fisherman father, who told us many stories of the sea, I felt would be an interesting place to start. Having a farm girl as a mother, whose stories were passed down from her grandfather's knee, was the opportunity I needed to fill the book with many treasured memories of family life back then. This along with writings gathered as a newspaper correspondent started me on a journey back into a past, which I thoroughly enjoyed along the way.

There was adventure, hardship, danger, and humor in the seeds of history that were left behind. This caught my interest and I wanted to write, so a younger generation could read something of an earlier generation of people. Their courage helped me to understand the suffering and sacrifice that was

```
<div n="10" type="page">
        <pb n="10"></pb>
        <div n="d1e643" rend="616,208,958,302">
          <p n="d1e649">
Forward</p>
        </div>
        <div n="d1e657" rend="94,524,1484,2092">
          <p n="d1e663">
Setting my pen to write "Pioneer Days & <orgname>Shanty</orgname>
Ways" gave
me a special reason for choosing the title for the book. Having
a pioneer fisherman father, who told us many stories of the sea,
I felt would be an interesting place to start. Having a farm girl
as a mother, whose stories were passed down from her
grandfather's knee, was the opportunity I needed to fill the book
with many treasured memories of family life back then. This
along with writings gathered as a newspaper correspondent
started me on a journey back into a past, which I thoroughly
enjoyed along the way.</p>
          <p n="d1e706">
There was adventure, hardship, danger, and humor in the
seeds of history that were left behind. This caught my interest
and I wanted to write, so a younger generation could read
something of an earlier generation of people. Their courage
helped me to understand the suffering and sacrifice that was
made, so families could enjoy a better way of life than their
forebears.</p>
          <p n="d1e737">
Wanting to shed some light on my Irish great grandfather,
<persname>
              <forename>Richard</forename>
              <forename>Dawson</forename>
            </persname>, with assistance from the
<orgname>Department of
Human Resources</orgname> brought success. This was the help I
needed
to confirm; <persname>
```

Figure 11. Users familiar with TEI can also view and edit the encoding manually. TEI is validated before it is saved.

30 The TEI editor is based on JavaScript and was originally inspired by Drupal's integration of CKEditor. When the Islandora TEI Editor code (which resides in a Drupal module) is installed, it defines access rules for the editor and provides information about how a page is rendered in the editor. The code retrieves a TEI document from the repository (specified by the page's persistent identifier), permits editing, and validates the edited TEI XML against the custom TEI schema before saving the document back to Fedora. The validation occurs within the TEI Editor module against a RELAX NG file that was generated using Roma (a tool for working with TEI customizations) and stored locally as part of the module's files. A document that does not pass validation produces a dialog box to inform the user that their edits will not be saved to Fedora.

31 In this way, the editor is designed to combine WYSIWYG elements with a few editing options for more advanced TEI users who may prefer to tag more directly. The WYSIWYG code was

designed to be extended: the TEI schema is defined (in JSON) separately from the code for the editor, allowing new TEI elements and attributes to be added easily (see example 2).

```
// Format of TEI elements that are defined in the TEI Editor
  {
    this.persname = {
     name: 'persName',
     bgColor: '#d5e5f1',
     attributes: [att.global, att.datable, att.editLike, att.personal, att.typed],
        };
```

Example 2. Format of TEI elements defined in the TEI editor.

32 If tags and elements were to be added to this JSON file and the associated CSS file, they would appear immediately in the XML editing tools and be functional. The user would also need to ensure that the element was defined and able to be validated within their custom TEI schema.

33 The code has not been developed beyond what was required for the project, although messages posted to the Islandora list in April 2012 suggested that the community may be interested in reviving the code base and bringing it into line with current versions of Islandora. The code is available at Islandora's GitHub repository.[6] At the same time as the community is showing a renewed interest in this project, other projects using Islandora are emerging to address more complex scholarly requirements for TEI.

8. Current TEI-related Initiatives in Islandora

34 Editing Modernism in Canada (EMiC) is a $3.8 million project funded by a SSHRC Strategic Knowledge Cluster / Strategic Network Grant, and is designed to run through 2015 (Pelham 2010). The project aims to create, disseminate, and foster public discourse about the literature of Canadian modernism (early to mid-twentieth century), develop a suite of tools for the production of digital texts, and provide training to, and develop relationships among, digital humanists. As part of the mandate to forge partnerships, EMiC has also partnered with the Canadian Writing and Research Collaboratory (CWRC), the Modernist Versions Project, and Mukurtu. As such, the project represents a number of key forces affecting tool and project development in the digital humanities. Based largely on an exploration of the IslandLives project, EMiC contracted with discoverygarden inc. to develop key tools for use in Islandora, including a web-based TEI editing interface that would enable users to create, edit, and validate TEI documents against a RELAX NG schema, and provide some WYSIWYG or even WYSIWYM ("what you see is what you mean") functions in addition to straight editing of raw XML. The resulting TEI will be stored, versioned, and managed from the underlying Fedora Commons repository.

35 The project is currently integrating successive versions of CWRC-Writer tool to meet EMiC's TEI requirements. The CWRC-Writer tool is being actively developed by the Canadian Writing and Research Collaboratory project (CWRC). The tool, currently in version 0.3, extends the JavaScript-based TinyMCE editor using jQuery. Broadly speaking, the goal for the development of CWRC-Writer is to facilitate the production of enriched documents by scholars with limited knowledge of RDF and XML markup. At the same time, the application is designed to be useful to those who prefer to work in code. Committed to open standards and information sharing, the project's goals include:

- Close-to-WYSIWYG editing and enrichment of scholarly texts with meaningful visual representations of markup
- Ability to add named entity annotations to texts
- Ability to combine TEI markup for the text and stand-off RDF for named entities
- Ability to export using "weavers" that recombine the plain text, the TEI, and the RDF into different forms (including an embedded TEI-compliant XML)
- Documented code to allow editorial projects to incorporate CWRC-Writer into their environments. (Rockwell et al. 2012)

36 Additional information about CWRC-Writer, including information on how to test-drive a current demo, is available at the project website.[7] Experimental code for integrating CWRC-

Writer into Islandora is available in GitHub but is rapidly changing as CWRC-Writer itself is incomplete.

37 Islandora's integration with the SharedCanvas (Sanderson et al. 2011) image annotation tool is more complete, and will be released as a separate "image annotation" solution pack in 2013, for Islandora 7. Beta versions of the module are available in the Islandora repository. SharedCanvas is being developed in a partnership between Stanford University, Los Alamos National Laboratory, the Open Annotation Collaboration, and the Mellon Foundation. The SharedCanvas application facilitates complex, layered image annotation and complies with the Open Annotation Collaboration's standard for the production of open annotations (Sanderson and Van de Sompel 2011).

38 Since the IslandLives project was begun, a number of web-based XML editing tools have emerged that may facilitate the development of a completely integrated web-based XML editor. A December/January thread on TEI-L under the subject "web-based XML editors" included over a dozen suggestions for possible open-source tools that would allow users to validate against a RELAX NG schema, and provide some WYSIWYG or WYSIWYM functions in addition to simply editing raw XML. As the IslandLives project continues to evolve, and new tools are built by EMiC, CWRC, and others, the Islandora project and UPEI hope that community input will improve the quality of generic tools available in Islandora for TEI scholars. That said, the Islandora framework is designed so that most web-based applications can be integrated and users can develop custom workflows and tools required to support specific TEI projects.

Bibliography

Breuel, Thomas, ed. 2010. "The hOCR Embedded OCR Workflow and Output Format." Last modified March 2010. https://docs.google.com/document/preview?id=1QQnIQtvdAC_8n92-LhwPcjtAUFwBlzE8EWnKAxlgVf0.

Cunningham, Hamish, Diane Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications." *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002. http://eprints.aktors.org/90/.

Pape, Sebastian, Christof Schöch, and Lutz Wegner. 2012. "TEICHI and the Tools Paradox: Developing a Publishing Framework for Digital Editions." *Journal of the Text Encoding Initiative* 2 (Feb.). http://jtei.revues.org/432.

Pelham, Amanda. 2010. "Speaking Volumes." *Dal News* [Dalhousie University], October 29. http://www.dal.ca/news/2010/10/29/volumes.html.

Rockwell, Geoffrey, Susan Brown, James Chartrand, and Susan Hesemeier. 2012. "CWRC-Writer: An In-Browser XML Editor." Poster presented at Digital Humanities 2012 (Hamburg, Germany, July 16–20). http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/cwrc-writer-an-in-browser-xml-editor/

Sanderson, Robert, Benjamin Albritton, Rafael Schwemmer, and Herbert Van de Sompel. 2011. "SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination." Paper presented at 11th ACM/IEEE Joint Conference on Digital Libraries (Ottawa, Ontario, June 13–17, 2011). http://www.arxiv.org/abs/1104.2925.

Sanderson, Robert, and Herbert Van de Sompel, eds. 2011. "Open Annotation: Beta Data Model Guide." Last modified August 10, 2012. http://www.openannotation.org/spec/beta/.

Wilcox, David. 2012. "Creating Custom Content Models." Last modified June 25, 2012. https://wiki.duraspace.org/display/ISLANDORA6122/Creating+Custom+Content+Models

Resources (Sites Mentioned)

ABBYY Products & Services: http://finereader.ABBYY.com/

Apache Solr project: http://lucene.apache.org/solr/

Canadian Writing and Research Collaboratory: http://www.cwrc.ca/en/

Drupal: http://drupal.org/

Editing Modernism in Canada Project: http://editingmodernism.ca/

Fedora Commons: http://fedora-commons.org/

Islandora: http://www.islandora.ca/

IslandLives: http://www.islandlives.ca/

Modernist Commons (The EMiC project): http://modernistcommons.ca/

TEI-L archive: http://listserv.brown.edu/tei-l.html

Notes

[1] "Saint Patrick's Confessio Hypertext Stack Project," accessed December 12, 2012, http://www.confessio.ie/.

[2] Quinn Dombrowski and Elijah Meeks, *Drupal for Humanists*, http://drupal.forhumanists.org/.

[3] Prince Edward Island Baptismal Index, Public Archives and Records Office, accessed December 12, 2012, http://www.gov.pe.ca/archives/baptismal/.

[4] "Geographical Names of Canada," Natural Resources Canada, last modified March 1, 2011, http://www.nrcan.gc.ca/earth-sciences/products-services/mapping-product/topographic-maps/5776.

[5] https://github.com/roblib/TEI

[6] https://github.com/Islandora/islandora_tei_editor

[7] https://sites.google.com/site/cwrcwriterhelp/

Cite this article

Electronic reference

Kirsta Stapelfeldt and Donald Moses, « Islandora and TEI: Current and Emerging Applications/Approaches », *Journal of the Text Encoding Initiative* [Online], Issue 5 | June 2013, Online since 24 June 2013, connection on 15 April 2015. URL : http://jtei.revues.org/790 ; DOI : 10.4000/jtei.790

Authors

Kirsta Stapelfeldt
Kirsta Stapelfeldt, MA, MLIS, is the manager of the Islandora project at the University of Prince Edward Island's Robertson Library, and has served as a subject matter expert for discoverygarden inc. on Editing Modernism in Canada and other projects.

Donald Moses
Donald Moses, MLIS, is the digital initiatives and systems librarian at UPEI's Robertson Library, and manages the IslandLives project as part of his role as manager of the Island Archives project, an umbrella project containing numerous digital history initiatives undertaken at UPEI.

Copyright

Abstract

Islandora is an open-source software framework developed since 2006 by the University of Prince Edward Island's Robertson Library. The Islandora framework is designed to ease the management of security and workflow for digital assets, and to help implementers create custom interfaces for display, search, and discovery. Turnkey options are provided via tools and modules ("solution packs") designed to support the work of a particular knowledge domain (such as chemistry), a particular content type (such as a digitized newspaper), or a particular task (such as TEI encoding). While it does not yet have native support for TEI, Islandora provides a promising basis on which digital humanities scholars could manage the creation, editing, validation, display, and comparison of TEI-encoded text. UPEI's IslandLives project, with its forthcoming solution pack, provides insight into how an Islandora version 6 installation can support OCR text extraction, automatic structural/semantic encoding of text, and web-based TEI editing and display functions for site administrators. This article introduces the Islandora framework and its suitability for TEI, describes the IslandLives approach in detail, and briefly discusses recent work and future directions for TEI work in Islandora. The authors hope that interested readers may help contribute to the expansion of TEI-related services and features available to be used with Islandora.

Index terms

Keywords : Islandora, annotation, RDF, CWRCWriter, SharedCanvas, Drupal, EMiC