The Bodleian TEI Catalogue Consolidation Project

James Cummings and Andrew Hankinson

About the Project

The Bodleian's TEI Catalogue Consolidation project was a joint Bodleian / IT Services project at the University of Oxford, designed to implement a single consolidated and sustainable infrastructure for delivering TEI manuscript description catalogues. The project incorporated user research, scoping, technical design, implementation, user testing, and subsequent iteration. The aims of the project were to determine a suitable technical architecture for the storage and indexing of TEI manuscript descriptions across multiple collections; engage in user testing of the existing TEI-based catalogues at the Bodleian and decide on the functional improvements required for the front end interface(s); build or implement the 'back-end' technical architecture as scoped; build a new, user-friendly interface for searching and browsing TEI; and design and implement a sustainability plan around training, communications and standards.

The Bodleian Libraries holds well over 20,000 manuscripts and serves readers from all academic divisions of the University of Oxford as well as thousands of external readers from around the world. More than half of the manuscript collections in the Bodleian are effectively hidden from scholars because they are not described online, or because the existing descriptions are not well indexed or even available to users. To try to solve this problem, the Bodleian spent five years creating TEI-based catalogues for descriptions of its manuscripts. However, the result was eight separate TEI catalogues, the largest of which (http://www.fihrist.org.uk/) is the union catalogue fo the description of Islamic manuscripts in the UK. Six other catalogues (Armenian, Genizah, Georgian, Hebrew, Senmai, Tibetan) represent portions of the Bodleian Libraries Oriental collections. The most recent addition is the TEI catalogue for western medieval manuscripts, funded by *The Tolkien Trust*. Prior to the consolidation project, small improvements were made to each subsequent catalogue, but funding was not available to rollout those improvements retrospectively, or to consolidate the code base. TEI is rapidly becoming the de-facto language for encoding descriptions of manuscripts collections (in the Bodleian and internationally) and this project was an opportunity to build on its use in a concrete way, by providing a clear set of software, guidelines, and training capacities for the large number of staff engaged with describing manuscripts. Consolidating this infrastructure has significantly lowered the operating costs of support and feed directly in the University's strategy for enhancing access to collections.

As part of the project, the existing TEI manuscript description schemas were consolidated into a single TEI customization. By using TEI, the catalogue descriptions were made more easily expandable to include full transcriptions of manuscripts, leading the way towards the mark-up and extraction of important biographical, geographical, and scientific data. In addition the existing catalogue descriptions were migrated to this single TEI customization. As well as showcasing the project as a whole, the poster will focus on the consolidation process including the key aims and objectives of the project, which were:

- To decide on a technical architecture for the storage and indexing of TEI-XML manuscript descriptions across multiple collections.
- Engage in user testing of the existing TEI-based catalogues at the Bodleian to decide the functional improvements required for the front end interface(s).
- Develop a technical requirements document with the necessary details to implement the recommendations during the subsequent build project.
- Build the 'back-end' technical architecture as scoped.
- Build a new user-friendly public interface for searching and browsing the TEI.
- Migrate the existing TEI applications into the new solution.
- Design and implement a sustainability plan around training, communications and standards.

Schema Consolidation Report

A schema consolidation report was produced based on an examination of all of the elements and attributes (and their values) in each of the catalogues. The full report is available at: http://research.it.ox.ac.uk/MSS/BodSchemaCreation.pdf

Reports on Individual Bodleian collections:

- Armenian catalogue: http://research.it.ox.ac.uk/MSS/armenian.html
- Fihrist catalogue: http://research.it.ox.ac.uk/MSS/fihrist.html
- Genizah catalogue: http://research.it.ox.ac.uk/MSS/genizah.html
- Georgian catalgue: http://research.it.ox.ac.uk/MSS/geogian.html
- Hebrew catalogue: http://research.it.ox.ac.uk/MSS/hebrew.html
- Medieval catalogue: http://research.it.ox.ac.uk/MSS/medieval.html
- Senmai catalogue: http://research.it.ox.ac.uk/MSS/senmai.html
- Tibetan catalogue: http://research.it.ox.ac.uk/MSS/tibetan.html

These were also compared to the Cambridge University Library's TEI manuscript catalogue:

- Cambridge catalogue: http://research.it.ox.ac.uk/MSS/cambridge.html

Catalogue	XML Files	Distinct XML Elements	XML Elements	Attributes	Attribute Values
Armenian	124	80	20602	9812	9851
Fihrist	11677	143	1573493	809102	821456
Genizah	225	69	151101	92938	93260
Georgian	94	81	13332	5638	5676
Hebrew	473	101	70372	29027	30589
Medieval	9173	123	810749	462124	463942
Senmai	1051	60	118166	40520	42007
Tibetan	132	94	13748	4860	4975
Total:	22949	165	2771563	1454021	1471756

Customizations of the TEI are made using its own literate programming customization language 'ODD' (One Document Does-it-all). From these customisation files it is usual to generate both human readable documentation and schemas which document instances will be validated against. The major divisions in the TEI are that of modules. The TEI modules used by each of the collections are detailed in the table below:

Proposed	Armenian	Fihrist	Genizah	Georgian	Hebrew	Medieval	Senmai	Tibetan	Cambridge
core	core	core	core	core	core	core	core	core	core
figures	-	figures	-	-	figures	figures	-	-	figures
header	header	header	header	header	header	header	header	header	header
linking	-	linking	-	-	-	linking	-	-	linking
msdescription	msdescription	msdescription	msdescription	msdescription	msdescription	msdescription	msdescription	msdescription	msdescription
namesdates	namesdates	namesdates	namesdates	namesdates	namesdates	namesdates	namesdates	namesdates	namesdates
tei	tei	tei	tei	tei	tei	tei	tei	tei	tei
textstructure	textstructure	textstructure	textstructure	textstructure	textstructure	textstructure	textstructure	textstructure	textstructure
transcr	transcr	transcr	transcr	transcr	transcr	transcr	-	transcr	transcr

These are precisely the modules that one would expect manuscript description to use. The core module contains the basic elements that occur in many TEI documents. The figures model provides elements for dealing with figures and tables. The header module contains elements for the general TEI metadata stored in the teiHeader. The linking module has elements for linking, segmentation, and alignment of texts. The manuscript description, by far the most detailed module of the TEI used for manuscript description catalogues, is for this specialised form of metadata. The names and dates module increases the basic name and date elements with additional elements and attributes.

Additional TEI elements added to Bodley-MSS TEI ODD Customization:

am	distributor	msFrag	relatedItem
authority	district	musicNotation	schemaRef
cit	ex	objectType	scriptDesc
correspAction	expan	orig	scriptNote
correspContext	geogFeat	pb	sponsor
correspDesc	l	projectDesc	surfaceGrp
dim	lg	reg	zone

Legacy Data Migration

The migration of existing catalogue files was done using an XSLT stylesheet. This copied through existing XML structures but replaced and updated these where possible.

Major types of changes

- Standardization of elements to most specific form e.g <persName> instead of <name type="person">
- Normalization of @xml:lang and similar attributes e.g. fre to fr, geo to ka
- Removal of entirely empty or template elements from legacy templates used
- Renaming of non-TEI elements e.g. folia, format, marginalia, Rubric, inscribed to TEI equivalents
- Standardization of @type attributes on <bibl>, <decoNote>, <dimensions>, <name>, <title>, etc.
- Updating to current TEI practice where this has changed e.g. need for <publisher>, <distributor>, <authority> as first element in <publicationStmt> or <altIdentifier> in <msPart> needing to be inside <msIdentifier>
- Attempting to ensure pointing elements have real URIs or URI-fragments such as <date calendar="#gregorian">
- Provision of standard @xml:id attributes for all <msDesc>, <msPart>, <msFrag>, and <msItem> elements (up to 6 levels of <msItem> nesting)
- Ensuring elements required in a particular order are given in that order

Check for truly empty elements

```
<xsl:function name="jc:checkEmpty" as="text()">
 <xsl:param name="node" as="node()"/>
 <xsl:variable name="output">
 <xsl:choose>
  <!-- if either of these is true, it isn't empty -->
  <xsl:when test="($node//text()[string-length(normalize-space(.)) gt 1])
   or ($node//@*[string-length(normalize-space(.)) gt 1])">false</xsl:when>
  <xsl:otherwise>true</xsl:otherwise>
 </xsl:choose>
 </xsl:variable>
 <xsl:value-of select="normalize-space($output)"/>
</xsl:function>
```

Split certain medieval catalogue manuscripts altIdentifiers

```
<xsl:function name="jc:splitAltIdentifier" as="item()*">
 <xsl:param name="altIdentifier" as="item()"/>
 <xsl:choose>
  <xsl:when
   test="$altIdentifier/idno[@type='SCN' and not(contains(., 'Not in SC')) and
         contains(., ',')] | $altIdentifier/idno[@type='TM']">
   <xsl:for-each select="tokenize($altIdentifier/idno, ',')">
    <altIdentifier><xsl:copy-of select="$altIdentifier/@*"/>
     <idno><xsl:copy-of select="$altIdentifier/idno/@*"/>
      <xsl:value-of select="normalize-space(.)"/>
     </idno>
    </altIdentifier>
   </xsl:for-each>
  </xsl:when>
  <xsl:otherwise>
   <altIdentifier><xsl:apply-templates select="$altIdentifier/@*|$altIdentifier/node()"/></altIdentifier>
  </xsl:otherwise>
 </xsl:choose>
</xsl:function>
```

Standardise MS Identification Numbers

```
<xsl:function name="jc:normalizeID">
 <xsl:param name="ID" as="item()"/>
 <xsl:variable name="pass0">
  <xsl:choose>
   <xsl:when test="matches($ID, '[0-9]\.[0-9]')">
    <xsl:variable name="part">
     <xsl:analyze-string select="$ID" regex="([a-zA-Z]+)\.">
      <xsl:matching-substring><xsl:value-of select="regex-group(1)"/></xsl:matching-substring>
      <xsl:non-matching-substring><xsl:value-of select="."/></xsl:non-matching-substring>
     </xsl:analyze-string>
    </xsl:variable>
    <xsl:value-of select="translate(normalize-space($part), '`!£$%^[_]°()}{,', '')"/>
   </xsl:when>
  <xsl:otherwise>
   <xsl:value-of select="translate(normalize-space($ID), '`!£$%^[_]°()}{,.', '')"/>
  </xsl:otherwise>
  </xsl:choose>
 </xsl:variable>
 <xsl:variable name="pass1">
  <xsl:value-of select="replace(normalize-space($pass0), ' - ', '-')"/>
 </xsl:variable>
 <xsl:variable name="pass2">
  <xsl:value-of select="replace(normalize-space($pass1), '\*', '-star')"/>
 </xsl:variable>
 <xsl:variable name="apos">&apos;</xsl:variable>
 <xsl:variable name="pass3">
  <xsl:value-of select="replace(normalize-space($pass2), $apos, '')"/>
 </xsl:variable>
 <xsl:value-of select="translate(normalize-space($pass3), ' \/',',_..')"/>
</xsl:function>
```

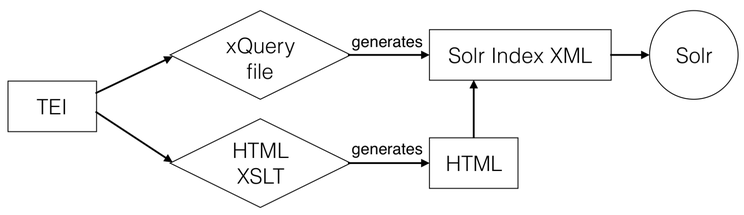
Catalogue Websites

Each catalogue site is delivered using the Blacklight application, which provides a front-end to a Solr index. The Solr index for four different entities—manuscripts, people, places, works—are generated using xQuery. We use the saxon xQuery processor to create a Solr XML file, which is then uploaded to the Solr server.

As a separate process for the manuscript records we produce an HTML rendering of the TEI record to use as the primary record display. This record is created using XSL, generating a single HTML file for each TEI record. This HTML is then embedded in a Solr field.

Each catalogue will be self-contained; that is, the data for the catalogue and the xQuery and XSLT files needed to produce the catalogue are present in the same GitHub repository.

For an example catalogue see https://medieval.bodleian.ox.ac.uk/



Bodleian Libraries
UNIVERSITY OF OXFORD

UNIVERSITY OF OXFORD

The Tolkien Trust

GitHub Repositories:

https://github.com/bodleian/armenian-mss
https://github.com/bodleian/fihrist-mss
https://github.com/bodleian/genizah-mss
https://github.com/bodleian/georgian-mss
https://github.com/bodleian/hebrew-mss
https://github.com/bodleian/medieval-mss
https://github.com/bodleian/senmai-mss
https://github.com/bodleian/tibetan-mss
https://github.com/bodleian/consolidated-tei-schema