Creating Digital Texts

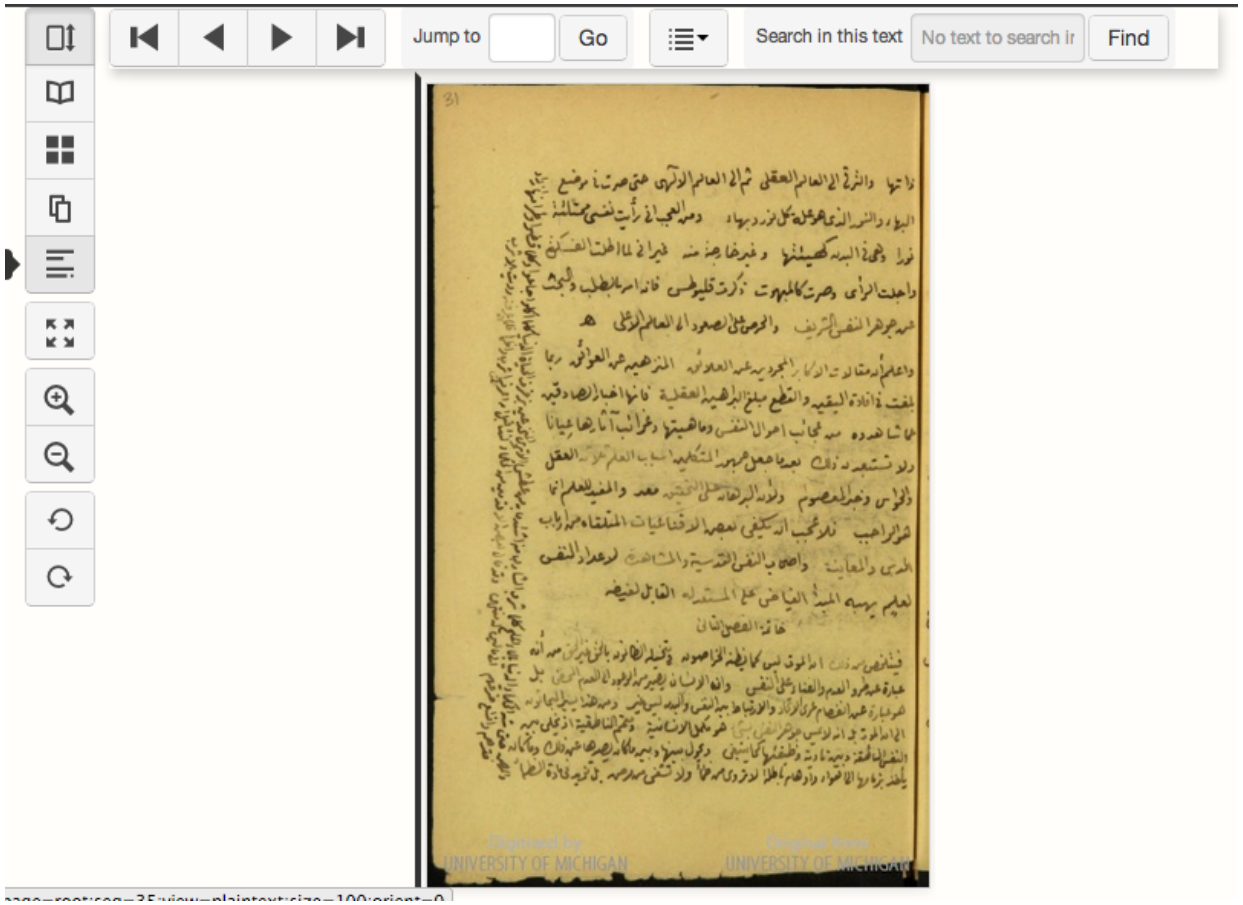
Rebecca Welzenbach
Text Creation Partnership | Michigan Publishing
University of Michigan Library

Overview

- What is text encoding, and how does it fit into a digitization project?
- How do scholars and researchers use XML, specifically, TEI to enhance their research?
- How can librarians and archivists use TEI to improve a digital collection, and to support the needs of researchers?
- Questions/discussion

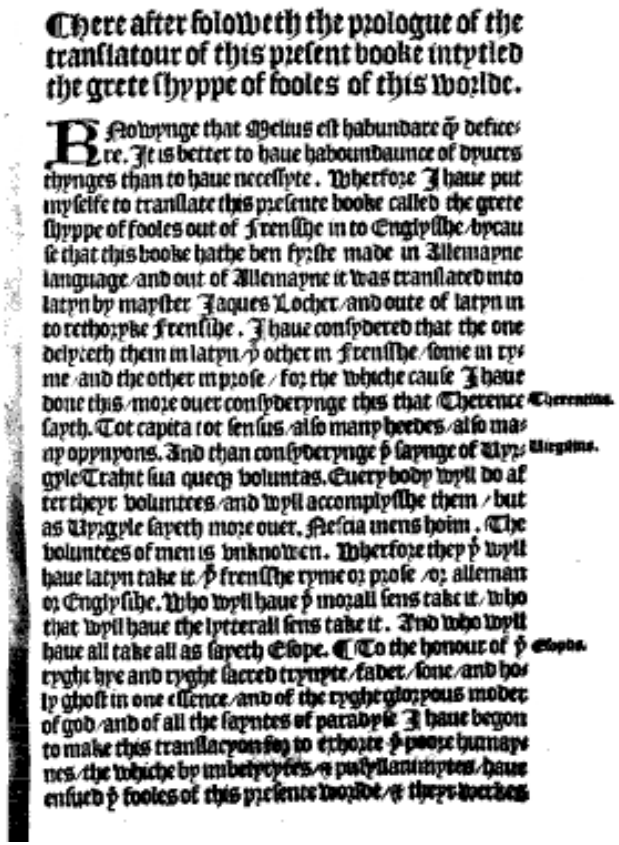
What is text encoding?
How does it fit into the digitization process?

Pt. 1: Scanning



What is text encoding?
How does it fit into the digitization process?

Pt. 2: Text capture (OCR or manual transcription)

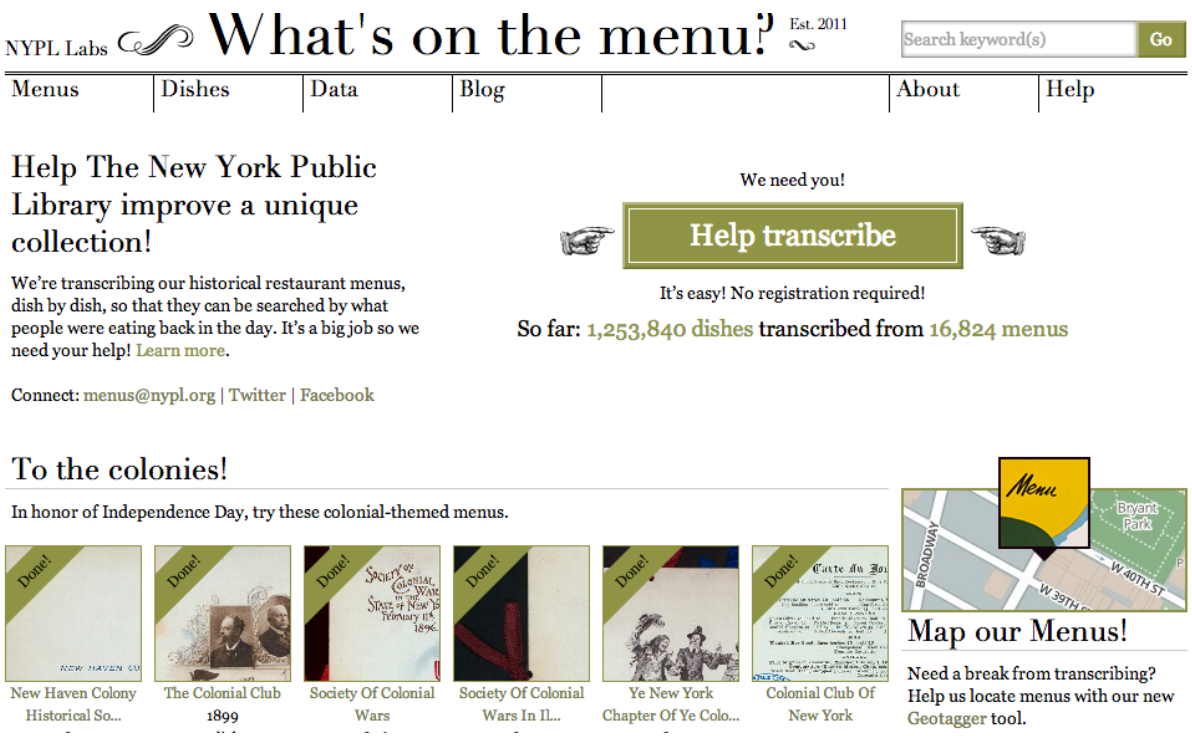


Here after foloweth the prologue of the translatour of this present booke intytled the grete shyppe of fooles of this worlde.

KNowynge that Melius est habundare q̃ deficere. It is better to haue haboundaunce of dyuers thynges than to haue necessyte. Wherfore I haue put myselfe to translate this presente booke called the grete shyppe of fooles out of Frensshe in to Englysshe / bycause that this booke hathe ben fyrste made in Allemayne language / and out of Allemayne it was translated into latyn by mayster Jaques Locher / and oute of latyn in to rethoryke Frensshe. I haue consydered that the one delyteth them in latyn / ye other in Frensshe / some in ryme / and the other in prose / for the whiche cause I haue done this / more ouer consyderynge this that Therence * sayth. Tot capita tot sensus / also many heedes / also many opynyons. And than consyderynge ye saynge of Uyrgyle * Trahit sua que{que} voluntas. Euery body wyll do after theyr voluntees / and wyll accomplysshe them / but as Uyrgyle sayeth more ouer. Nescia mens hoim. The voluntees of men is vnknowen. Wherfore they y^t wyll haue latyn take it / ye frensshe ryme or prose / or alleman or Englysshe. Who wyll haue ye morall sens take it / who that wyll haue the lytterall sens take it. And who wyll haue all take all as sayeth Esope. ¶ To the honour of y^e* ryght hye and ryght sacred trynyte / fader / sone / and holy ghost in one essence / and of the ryght gloryous moder of god / and of all the sayntes of paradyse I haue begon to make this translacyon for to exhorte ye poore humaynes / the whiche by imbelycytes / & pusyllanimytes / haue ensued y^e fooles of this presente worlde / & theyr werkes

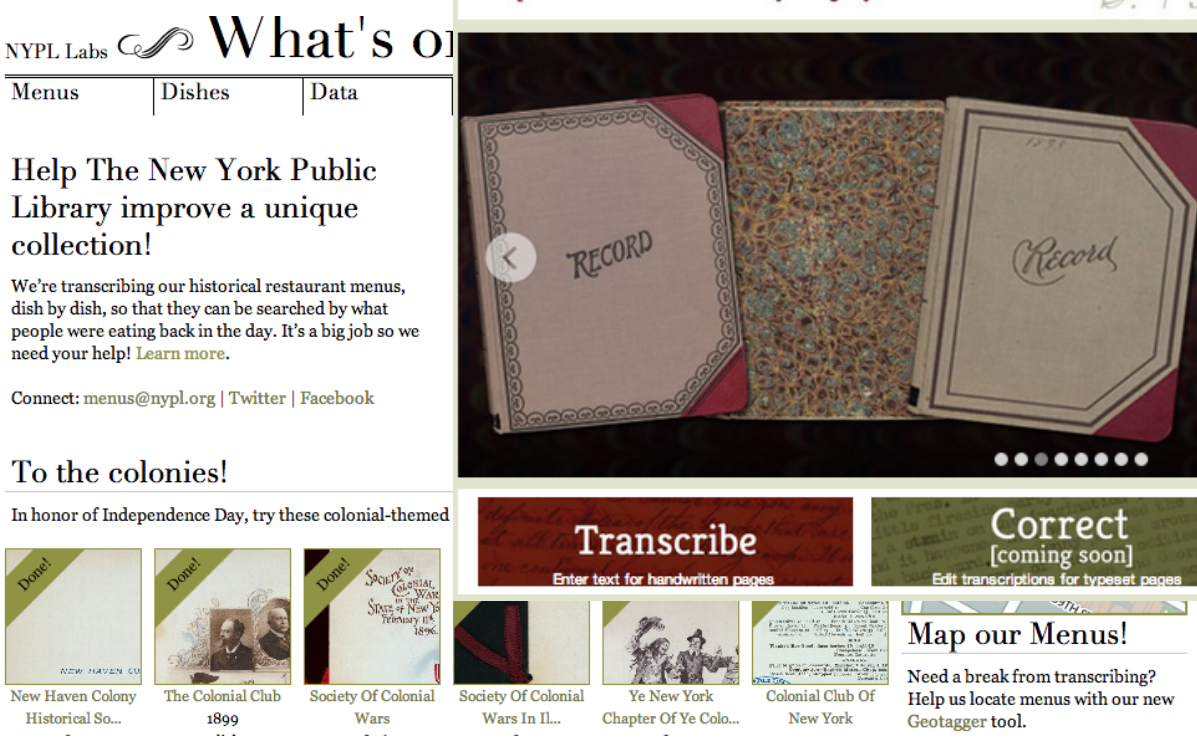
What is text encoding?
How does it fit into the digitization process?

Pt. 2: Text capture (OCR or manual transcription)



What is text encoding?
How does it fit into the digitization process?

Pt. 2: Text captu



What is text encoding?
How does it fit into the digitization process?

Pt. 2: Text captu

NYPL Labs What's o

Menus | Dishes | Data

Help The New York Public Library improve a unique collection!

We're transcribing our historical restaurant men
dish by dish, so that they can be searched by wha
people were eating back in the day. It's a big job s
need your help! Learn more.

Connect: menus@nypl.org | Twitter | Facebook

To the colonies!

In honor of Independence Day, try these colonial

New Haven Colony Historical So...

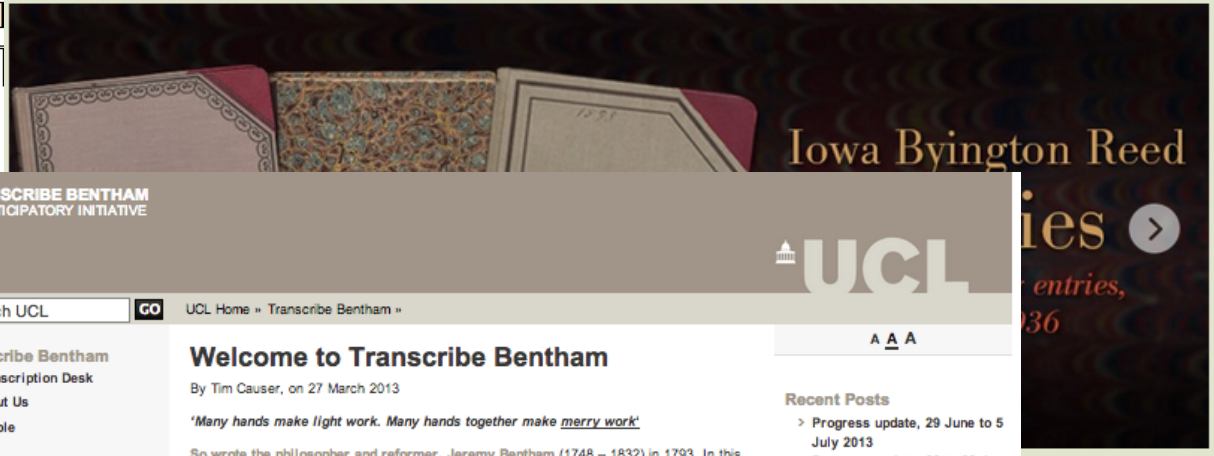
The Colonial Club 1899

Society Of War

DIY|History

Help build the historical record by doing it yourself!

The University of Iowa LIBRARIES



TRANSCRIBE BENTHAM
A PARTICIPATORY INITIATIVE

UCL

Search UCL GO

UCL Home » Transcribe Bentham »

Transcribe Bentham
- Transcription Desk
- About Us
- People
- FAQ
- Talks and publications
- Jeremy Bentham
- Awards
- Publicity
- Contact Us

Education
- Information for Schools
- A-levels and Scottish Highers
- Inside and Outside the Classroom
- Palaeography

Related Sites
- Bentham Project
- Bentham Papers Digital

Welcome to Transcribe Bentham

By Tim Causer, on 27 March 2013

'Many hands make light work. Many hands together make merry work'

So wrote the philosopher and reformer, Jeremy Bentham (1748 – 1832) in 1793. In this spirit, we cordially welcome you to *Transcribe Bentham*, a double award-winning collaborative transcription initiative, which is digitising and making available digital images of Bentham's unpublished manuscripts through a platform known as the 'Transcription Desk'. There, you can access the material and—just as importantly—transcribe the material, to help the work of UCL's Bentham Project, and further improve access to, and searchability of, this enormously important collection of historical and philosophical material.

This is an exciting opportunity to make a genuine difference to research and scholarship by contributing to the production of the new edition of *The Collected Works of Jeremy Bentham*, and to help create for posterity a vast digital repository of Bentham's writings. We warmly invite you to take part in this endeavour: no special skills are required, you do not require approval to participate, and every contribution—no matter how small—is of great value to *Transcribe Bentham*.

Please consult the *Transcribe Bentham* FAQ for more details on taking part.

You can also read more about Jeremy Bentham, his thought and his importance, and consult resources on deciphering historical handwriting.

Find out more about the consortium behind *Transcribe Bentham*, and talks and

A A A

Recent Posts
- Progress update, 29 June to 5 July 2013
- Progress update, 22 to 28 June 2013
- Testing times: Transcribe Bentham's new interface
- Progress update, 15 to 21 June 2013
- Progress update, 8 to 14 June 2013

Archives
- July 2013
- June 2013
- May 2013
- April 2013
- March 2013
- February 2013
- January 2013
- December 2012
- November 2012
- October 2012

What is text encoding?
How does it fit into the digitization process?

Pt. 2: Text capture (OCR or manual transcription)

Argument in the shyppe of Fooles of this worlde.

O: the felycyte and salute of all the humayne gendre is compyled and dyscrete the shyppe of fooles of this transyeory worlde in the whyche abcenderh all they that wagerh from ꝑ waye of trouthe and from the playne ecbortacyon of ꝑ intellectyue vnderstondynge in transmutable and obscure thoughtes of ꝑ frayle body. Wherfore this pre-sent boke may be called sarysc notwithstondynge that the fyrste auctoure dyde delyte hym in the newe intytu-lacyon of this present booke for ryght so as by the pre-sent boke and fyrcyons the auncyent poetes dyde correcte ꝑ vyces and fragylyteis of mortall men. ¶ Semblably this present pagyne specifyeth before theyr syght the es-tate and condycyon of men to the ende that as a mye-roure they beholde the meurs and rectytude of lyfe. Ne ueertheles thynke not ye lectours that I haue wrote by worde dyscrete and reduced this presente booke out of Frenshe in to our maternall tongue of Englyshe for I haue onely as redyteth Flaccus taken entyrely the substaunce of the scrypture in esperaunce that myn au-dace presumpteuous sholde be pardonned of the lectour-es hauynge aspecte vnto ꝑ capacyte of my tendre ye-res and the imbecylyte of my lytel vnderstondynge in leuynge the egrellyons poetryques and fabulous obscur-tees in adiuerynge the werke in facyle sentence and fa-mylyar style in supplyenge al the reders to haue me for

The shyppe of.

Ment in che lhpppe okKooles
ozwe. :»c^H-^H^^^H'
M chcfclpcytc and sälutc of all che
hulnapl« gcnd« ls compz'led and
dp^cne che lhyppc of tooles ok chl»
nansM)pwo:ldc mtlic vHlchc ab
crndcch all chey chat vagcch fron» ^
Wgpc of trouchc and from chc plap-
N ne crho'.tacj'on of ß lntcllccy'llc vns
dcrlwndzmye mtransmutablcand
cchlcurcchouFtncsof ^fraylebodz'.Vinclfo:ctnlS p.cs
smtboke map de cgllcd say'ic motWltlittondmFc chat
chc fyistc auttoure dydc delytc tiM «n tyc ncwc mtM^
lacvon of chls p:c<cnt bookc fo: ryyht lo asbz' chcpocs
lt'es and fvccyons M auncycm pocres dydc co:rcctch
v«rs änd kragylptecs of nwuaU mm. CAcnwlalilp
rini, mtscnr päP'nc lpcnfycch betont tl^yr sMt ll)c cb
r<nc and nmdycj'on of mm to chc cndc chat ab a mxr>
wurc chcp vcl^ldc chc mcurs and rccr^mdc of lz'fc.Oe
ucrchclcs rly'lUicllotxclccrourschgt^ hauclro:dcbp
wo:dcdxzcc«and reduccd thlbp«lcntc bookcoutc of
Lrrnllhc m to om- nlätcrnall tonyuc ok GnylylNic to:
7 l,auc owij' gsrclytcch Llaccus^taKcn cnti'cicIz' chc
llbttaulucof tnc scrxpturc m esperaunccll>atmyngu
oacc p^sumpnwuslholdcbcpardonncdot chclcctou^
«cs vauyngc aspccre vnw ^ capan'tc of mx tcnlncyc
res andllicImvyl^tcokmylylclvnocrttondxngc m
lcuznigc che cgrclh'ons voctpqucs and fzdlflloue odlcur
5ccs u,acln'curnycll^v»Kcu»fac?lcscntnvcandfg
»nplf'cr ttvlc in llmplvcngc gl chc rcdcwto liauc n,c so?
«Niclhz'ppcof. Hl.

Image and OCR from Hathi Trust
http://hdl.handle.net/2027/wu.89007024094

What is text encoding?
How does it fit into the digitization process?

Pt. 2: Text capture (OCR or manual transcription)

EMOP is trying to use specialist knowledge of early modern typefaces to train OCR engines to do a better job on previously indecipherable texts. The goal is not perfection, but to produce texts good enough that automated OCR is worthwhile, and then can be reviewed and corrected by the crowd.

http://emop.tamu.edu/

What is text encoding?
How does it fit into the digitization process?

Pt. 3: Putting the text into some kind of structure

PERFORMANCE STUDIES, MASS CULTURE, & THE JEWISH PROBLEM 3

The provisional and contested quality of acting Jewish distinguishes it on the one hand from performing Judaism, a concept based on the assumption of an "authentic" Jewish culture, and on the other hand from representations of the Jew, a concept that locates the generation of meaning exclusively in the performance text. My argument is based on the concept of double coding, the specific means and mechanisms by which a performance can communicate one message to Jewish audiences while simultaneously communicating another, often contradictory message to gentile audiences.

As the late performance theorist Dwight Conquergood explained, "Subordinate people do not have the privilege of explicitness, the luxury of transparency, the presumptive norm of clear and direct communication, free and open debate on a level playing field that the privileged classes take for granted."5 Jewish American artists in theater, film, and television, while certainly enjoying access to the "means of production" at a level far from subordinate, nonetheless have tended to approach their creative work from just such an outsider's point of view. Because of real or perceived anti-Semitism, Jewish characters and themes are frequently "reformed" for performance. David Zurawik, for example, writes in The Jews of Prime Time, ""Too Jewish' is an expression that echoes all too loudly across the history of Jewish characters on network television in its use-most often by Jewish programmers and network executives-as a tool to distort, disguise, or altogether eliminate depictions of Jewish identity from American prime-time television."6 Yet this does not stop audiences from attempting to "decode hidden Jewish identity,"7 a pastime that I will argue is itself a source of community formation for many American Jews.

http://hdl.handle.net/2027/mdp.39015062623320

Text Encoding Initiative (TEI)

Pt. 3: Putting the text into some kind of structure

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. (http://www.tei-c.org/index.xml)

Text Encoding Initiative (TEI)

Pt. 3: Putting the text into some kind of structure

XML is:

- Open, non-proprietary standard
- Stored in ‘plain text’ but usually thought of as contrasting with it
- Marks beginning and ends of spans of text using tags:

<sentence>This is a sentence.</sentence>

Adapted from "Creating Digital Editions: an Introduction to the Text Encoding Initiative (TEI)"
http://www.lib.umich.edu/publishing-production/creating-digital-editions-introduction-text-encoding-initiative-tei

Text Encoding Initiative (TEI)

Pt. 3: Putting the text into some kind of structure

In an XML document, spans of text must nest properly:

Wrong:

```
<sentence>Overlap is <emphasis>not allowed!
</sentence></emphasis>
```

Right:

```
<sentence>Overlap is <emphasis>not allowed!
</emphasis></sentence>
```

Adapted from "Creating Digital Editions: an Introduction to the Text Encoding Initiative (TEI)"
http://www.lib.umich.edu/publishing-production/creating-digital-editions-introduction-text-encoding-initiative-tei

Text Encoding Initiative (TEI)

Pt. 3: Putting the text into some kind of structure

In an XML document, you have:

Elements (tags), attributes, values, content

<sentence type="declarative">This is a sentence.</sentence>

<sentence type="interrogative">Is this is a sentence?</sentence>

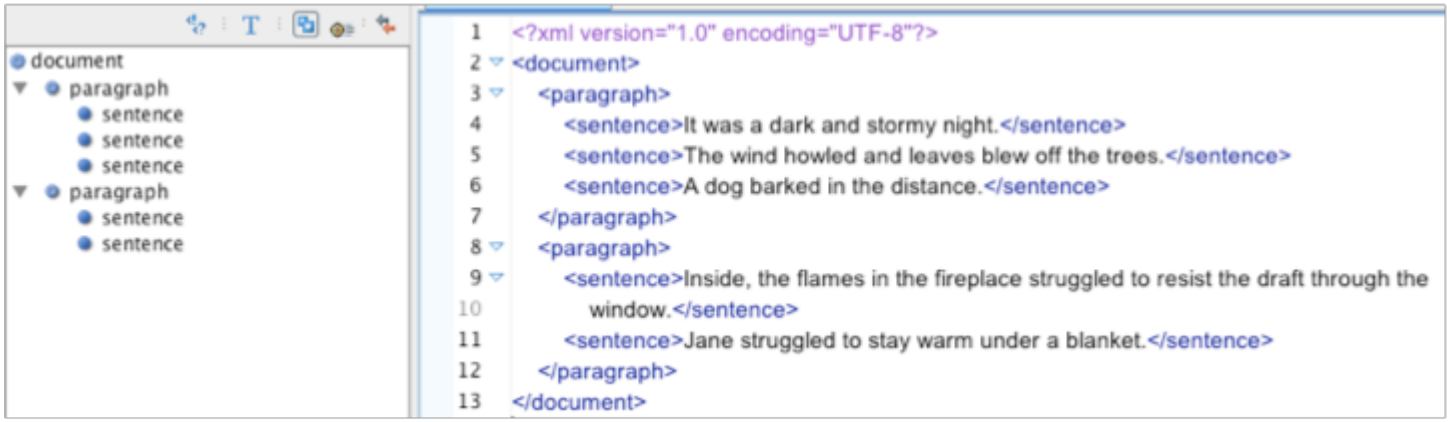
Elements may have zero, one, or more than one attribute

Adapted from "Creating Digital Editions: an Introduction to the Text Encoding Initiative (TEI)"
http://www.lib.umich.edu/publishing-production/creating-digital-editions-introduction-text-encoding-initiative-tei

Text Encoding Initiative (TEI)

Pt. 3: Putting the text into some kind of structure

In an XML document, you have a tree structure:



Adapted from "Creating Digital Editions: an Introduction to the Text Encoding Initiative (TEI)"
http://www.lib.umich.edu/publishing-production/creating-digital-editions-introduction-text-encoding-initiative-tei

Text Encoding Initiative (TEI)

Pt. 3: Putting the text into some kind of structure

A “header,” attached to the beginning of each XML document, expresses metadata about the source material and the creation of the XML file itself.

```
<HEADER>
     <FILEDESC>
          <TITLESTMT><TITLE TYPE="245" I2="2">A new petition to
          the Kings most Excellent Majestie The humble petition
          of the knights, gentry, free-holders, of the county of
          Yorke, who have subscribed hereto: presented to His
          Majestie at Yorke, April 30, 1642. Desiring a happy
          union betwixt the King and the Parliament.</TITLE>
          </TITLESTMT>
     <EXTENT>Approx. 4 KB of XML-encoded text transcribed from 1
     1-bit group-IV TIFF page image.</EXTENT>
     <PUBLICATIONSTMT><PUBPLACE>Ann Arbor, MI ; Oxford (UK) :
     </PUBPLACE>
     <PUBLISHER>Text Creation Partnership,</PUBLISHER><DATE>2008-
     09 (EEBO-TCP Phase 1).</DATE>
          <IDNO TYPE="DLPS">A90004</IDNO>
          <IDNO TYPE="stc">Wing N701</IDNO>
          <IDNO TYPE="stc">Thomason 669.f.6[9]</IDNO>
          <IDNO TYPE="stc">ESTC R211829</IDNO>
          <IDNO TYPE="eebo citation">99870521</IDNO>
          <IDNO TYPE="proquest">99870521</IDNO>
          <IDNO TYPE="vid">160870</IDNO>
     <AVAILABILITY><P>This keyboarded and encoded edition of the
     work described above is co-owned by the institutions
     providing financial support to the Early English Books
     Online Text Creation Partnership. Searching, reading,
     printing, or downloading EEBO-TCP texts is reserved for the
```

Adapted from "Creating Digital Editions: an Introduction to the Text Encoding Initiative (TEI)"
http://www.lib.umich.edu/publishing-production/creating-digital-editions-introduction-text-encoding-initiative-tei

Text Encoding Initiative (TEI)

Pt. 3: Putting the text into some kind of structure

Outside your XML document, you will need a schema, or a syntax for your XML documents, specifying:

- Which elements are allowed
- Which elements may nest inside of others
- In what order these elements must occur
- How many times they may repeat
- What attributes they may have
- What values those attributes may have

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ST.html#STIN

Adapted from "Creating Digital Editions: an Introduction to the Text Encoding Initiative (TEI)"
http://www.lib.umich.edu/publishing-production/creating-digital-editions-introduction-text-encoding-initiative-tei

Text Encoding Initiative (TEI)

Pt. 3: Putting the text into some kind of structure

Outside your XML document, you will need a platform, application, or software in order to process or render your TEI documents

What is text encoding?
How does it fit into the digitization process?

Pt. 3: Putting the text into some kind of structure

```
62 <BODY>
63 <DIV1 N="1" TYPE="chapter">
64 <PB REF="3">
65 <HEAD>THE
66 AMERICAN CRISIS<NOTE N="*" PLACE="foot">An&s;wer to the Proclamation of General Ga
67 <HEAD>No. I.</HEAD>
68 <EPIGRAPH>
69 <Q><L>Projicit <HI>ampullas.</HI> <BIBL>HOR.</BIBL></L>
70 <LG>
71 <L>On <HI>&s;ouls</HI> of &s;lav'ry more than death afraid,</L>
72 <L><HI>Gage</HI> wa&s;tes his <HI>pardons</HI> and his <HI>ga&s;conade.</HI></L>
73 </LG></Q>
74 </EPIGRAPH>
75 <P>GENERAL GAGE's Proclamation lies before
76 me, though it is not a &s;ubject for critici&s;m, yet
77 it de&s;erves notice; we may overlook the &s;tyle, but
78 mu&s;t dete&s;t the doctrine; with what judgment the
79 General may command, with what &s;pirit he
80 may fight, with what prudence he may retreat, is
81 yet unknown; but if his conduct in the field is
82 equal to his compo&s;ition in the clo&s;et, the &s;ucce
```

Page [unnumbered]

THE AMERICAN CRISIS, AND A LETTER TO *SIR GUY CARLETON*, ON THE MURDER OF CAPTAIN HUDDY, AND THE INTENDED RETALIATION ON CAPTAIN ASGILL, OF THE GUARDS.

BY THOMAS PAINE, AUTHOR OF COMMON SENSE—RIGHTS OF MAN—AGE OF REASON—AND THE DECLINE AND FALL OF THE ENGLISH SYSTEM OF FINANCE.

LONDON: PRINTED AND SOLD BY DANIEL ISAAC EATON, No. 74, NEWGATE-STREET.

Page [unnumbered]

IN CONGRESS, AUGUST 26, 1785.

RESOLVED, That the early, unsolicited and continued labours of Mr. THOMAS PAINE. in explaining and enforcing the principles of the late revolution, by ingenious and timely publications* upon the nature of liberty and civil government, have been well received by the citizens of these States, and merit the approbation of Congress▪ and that in consideration of these services, and the benefits produced thereby, MR. PAINE is entitled to a liberal gratification from the United States.

OCTOBER 3, 1785.

RESOLVED, That the board of treasury take order for paying to Mr. THOMAS PAINE, the sum of three thousand dollars, for the conlsiderations mentioned in the resolution of the 26th of August last.

Page [unnumbered]

How do scholars use TEI to enhance their research?

What kinds of features do scholars (users?) mark up?

- Tag instance of people, places, dates, locations to make these searchable or to analyze this data
- Mark up linguistic features such as parts of speech
- Document observations or claims about a source document (e.g. scribal corrections; changes in hand
- Attach scholarly interpretation (e.g., the occurrece of a particular theme, attaching a person's name to a pronoun used in the text) directly to a text
- Mark rhyme and meter
- Leave original spelling in place while attaching normalizations to aid in searching
- ....etc.!

How do scholars use TEI to enhance their research?

What is the end product of their research?

- A specialized digital edition of a single work or set of works
- A corpus of words ready for analysis
- An ebook
- A website
- A graphic visualization
- ...etc.!

How can librarians and archivists use TEI in digital collections?

There is no one right way to mark up a document. Decisions about text encoding will be driven by several factors:

- The goal of your project (what are you trying to achieve? Who is your audience?)
- How much money and time you have
- What do you know about your users?
- Where do you think cultural heritage institutions fit into the picture?

How can librarians and archivists use TEI in digital collections?

Best Practices for TEI in Libraries:

"A guide for mass digitization, automated workflows, and promotion of interoperability with XML using the TEI"

Outlines five recommended "levels" of encoding to guide libraries, archives, and other collections in creating digital texts consistently and sanely

http://www.tei-c.org/SIG/Libraries/teiinlibraries/

How can librarians and archivists use TEI in digital collections?

Level 1	The text is generated through OCR, is subordinate to the page image, and is not intended to stand alone as an electronic text (without page images). Encoding is done to assist in full text searching.
Level 2	The text is generated through OCR and is mainly subordinate to the page image, though navigational markers (textual divisions, headings) are captured.
Level 3	The text is created by conversion, either by way of OCR or keyboarding. Some structural elements of the text are encoded. The text may be used with or without page images.
Level 4	The text is generated either through corrected OCR or keyboarding and is able to stand alone without page images in order for them to be read by students, scholars, and general readers.
Level 5	The text is generated either through corrected OCR or keyboarding and is able to stand alone without page images, as in Level 4. In addition, the tagging requires substantial human intervention by encoders with subject knowledge.

http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html#index.xml-body.1_div.1

How can librarians and archivists use TEI to help researchers?

- Document and publish your encoding decisions so that users may understand what has been captured and not captured in your texts
- If your texts are in the public domain, or if you otherwise have rights to them, apply a Creative Commons license and make the raw XML file available for download. Users can save time and resources by building on your transcription rather than starting from scratch

Resources for TEI novices

Reading:

- Spectrum of "Gentle Introductions"
  - Hawkins, "An Introduction to XML for Text"
  - Mueller, "A Very Gentle Introduction to the TEI Markup Language"
  - Burnard & Sperberg-McQueen, "A Gentle Introduction to SGML"
- The Guidelines themselves: tei-c.org
- TEI listserv
- TEI Bibliography/Reading list

Resources for TEI novices

Teaching/Training:

- Workshops such as:
  - University of Victoria's Digital Humanities Summer Institute or Oxford's Digital Humanities Summer School (week-long)
  - Workshops at Brown University (weekend)
  - Occasional workshops here at the University of Michigan (day long)
- TEI by Example and other online training materials

Resources for TEI novices

Tools for generating and publishing TEI:

- OxGarage: Produce basic TEI documents out of Word and other formats
- TAPAS (in beta): Service to host and publish basic TEI texts on the web
- TEI Boilerplate: software to publish basic TEI texts on the web
- <oXygen/> text editor

Resources for Document Transcription

Tools for transcribing and displaying historical documents:

- T-Pen: “is a web-based tool for working with images of manuscripts. Users attach transcription data (new or uploaded) to the actual lines of the original manuscript in a simple, flexible interface.”

- DIY History software environment (as in use at Iowa): http://diyhistory.lib.uiowa.edu/code.html (makes use of Omeka and MediaWiki)

- Vendors, such as Apex CoVantage and SPi Global

Thank you! Questions?