INTRODUCTION TO THE TEXT ENCODING INITIATIVE (TEI) FOR HISTORICAL DOCUMENTS

Caitlin Pollock, @stilettobrarian
 HILT 2017, Austin, Texas

Background

➔ Machine readable texts
example: MARC (library) records

➔ Text Encoding for the Humanities
- ◆ 1949
- ◆ IBM and Father Roberto Busa*
- ◆ Busa began using computing to perform literary analysis on the works of Thomas Aquinas
- ◆ Searching for the concordance of words in Aquinas text corpus
- ◆ Index Thomisticus was printed, but later released on CD-ROM and the web.

➔ TEI traces back to Vassar College and the ACH.
➔ Now a large research community or consortium

*See: Gold, Matthew K. (2012)"The Digital Humanities Movement." In Matthew K. Gold (Ed.), Debates in Digital Humanities. http://dhdebates.gc.cuny.edu/debates

Sooo... What is TEI?

What is TEI-C and TEI?

From the website: The Text Encoding Initiative Consortium is an internal organization whose mission is to develop and maintain guidelines for the digital encoding of literary and linguistic texts. The Consortium publishes the Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange: an international and interdisciplinary standard that is widely used by libraries, museums, publishers, and individual scholars to represent all kinds of textual material for online research and teaching.

Shorter: TEI is a standard for encoding text using an XML format.

The TEI Guidelines

The TEI Guidelines in English can be located here:

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ (and there is a copy in the GitHub repo.)

Two most important parts of TEI:
 Text Body
 Element References

The TEI Header

File Description `<fileDesc>`

◆ *Title Statement `<titleStmt>`
◆ Edition Statement `<editionStmt>`
◆ Extent `<extent>`
◆ *Publication Statement `<publicationStmt>`
◆ Series Statement `<seriesStmt>`
◆ Notes Statement `<notesStmt>`
◆ *Source Description `<sourceDesc>`

Encoding Description `<encodingDesc>`

Profile Description `<profileDesc>`

Revision Description `<revisionDesc>`