Citation Key

for more information see: http://open.umich.edu/wiki/CitationPolicy

Use + Share + Adapt

{ Content the copyright holder, author, or law permits you to use, share and adapt. }

PD-GOV **Public Domain – Government**: Works that are produced by the U.S. Government. (17 USC § 105)

PD-EXP **Public Domain – Expired**: Works that are no longer protected due to an expired copyright term.

PD-SELF **Public Domain – Self Dedicated**: Works that a copyright holder has dedicated to the public domain.

CC ZERO **Creative Commons – Zero Waiver**

CC BY **Creative Commons – Attribution License**

CC BY-SA **Creative Commons – Attribution Share Alike License**

CC BY-NC **Creative Commons – Attribution Noncommercial License**

CC BY-NC-SA **Creative Commons – Attribution Noncommercial Share Alike License**

GNU-FDL **GNU – Free Documentation License**

Make Your Own Assessment

{ Content Open.Michigan believes can be used, shared, and adapted because it is ineligible for copyright. }

PD-INEL **Public Domain – Ineligible**: Works that are ineligible for copyright protection in the U.S. (17 USC § 102(b)) *laws in your jurisdiction may differ

{ Content Open.Michigan has used under a Fair Use determination. }

© FAIR USE **Fair Use**: Use of works that is determined to be Fair consistent with the U.S. Copyright Act. (17 USC § 107) *laws in your jurisdiction may differ

Our determination **DOES NOT** mean that all uses of this 3rd-party content are Fair Uses and we **DO NOT** guarantee that your use of the content is Fair.

To use this content you should **do your own independent analysis** to determine whether or not your use will be Fair.

SI 675 Digitization for Preservation

Week 5 – Text and Image in Digitization

Outline

Outline

- 1. Markup as an intellectual pursuit
- 2. Forms of representing text
- 3. Text Encoding Initiative
- 4. Examples
- 5. "Library" roles

• McGann, *Radiant Textuality (2001)*

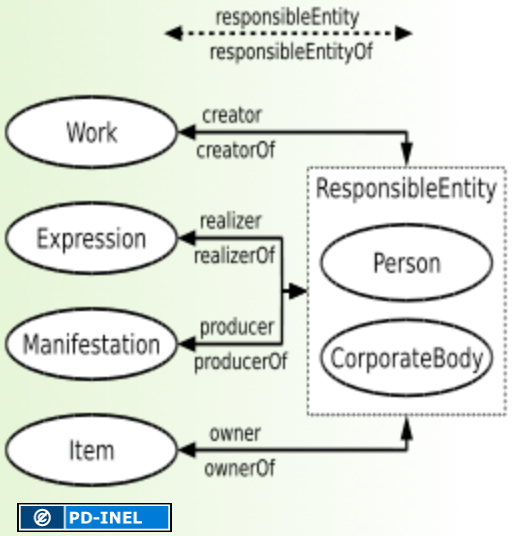
Theory of Markup

- Advancing the art and science of documentary editing – from compilation to critique to self-critique.
- McGann: "An edition is conceivable that might undertake as an essential part of its work a regular and disciplined analysis and critique of itself." (p. 81)
- Pragmatics of theory (speculation versus construction) (p. 83)
- Buzzetti: Markup is: "the use of embedded codes, known as tags, to describe a document's structure" (p. 67)

• **Buzetti, *Digital Representation and Text (2002)***

Text Markup

FRBR [IFLA/LC]



- What is an adequate digital representation?
  - Exhaustivity [beyond a simple text string]
  - "all of the information contained in the "literary material as originally written by an author" (p. 61)
  - Support for automated processing
  - "extract information from it and to represent it" (p. 62)
- Multipurpose representation: "The data model upon which the digital representation of the text is founded must be capable of transposing, by way of algorithms, the procedures for textual criticism and interpretative textual analysis. The model must be able to satisfy the needs of the philologist and the editor, as well as those of the historian and the literary critic.

• **Buzetti, *Digital Representation and Text (2002)***

Distinction: Content v. Expression

Markup

Representing Text

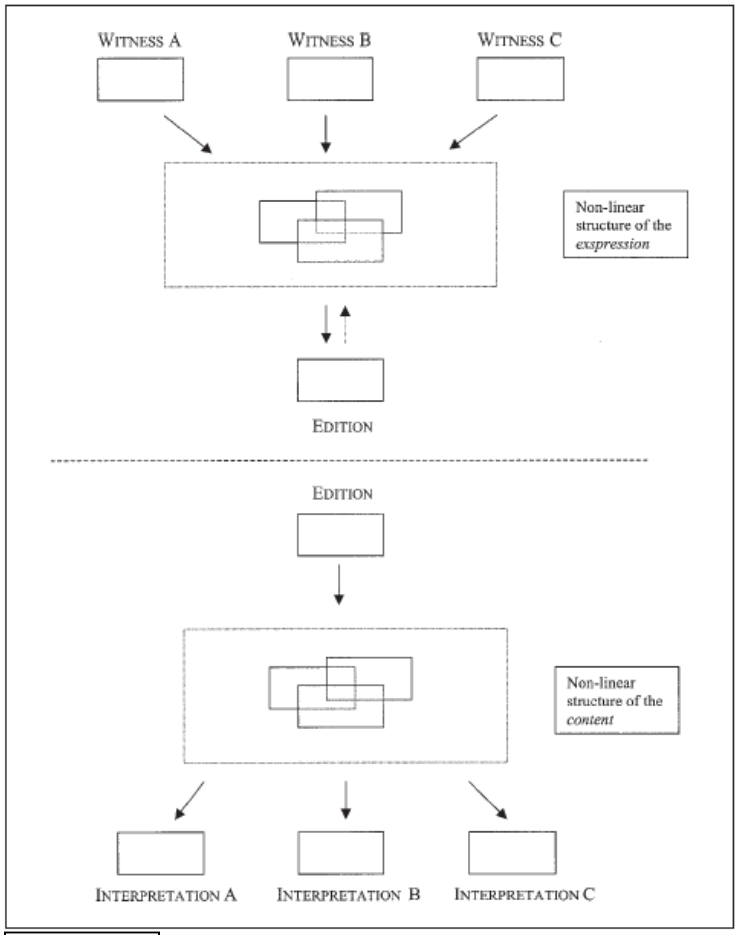
TEI

Examples

Support

"The indefiniteness of the relationship between the expression and the content is what assures the dynamism and mobility of the text: to any single expression many different contents may correspond, and to any single content many different expressions may correspond."

Buzetti, pp. 79-80.



• Renear (2004) and van der Weel (ND)

Text Encoding Initiative

- Roots in digital text processing
- International collaboration [1987]
- Data exchange as a primary motivating factor (NINES)
- Language used to identify particular text objects, in context
- Whitman guidelines as an example

Gentle Introduction to TEI. http://www.tei-c.org/Guidelines/P4/html/SG.html

•Renear, *Text Encoding (2004)*

OHCO: Ordered Hierarchy of Content Objects

Increasing analytical complexity challenges usefulness of a single structural mark-up model.

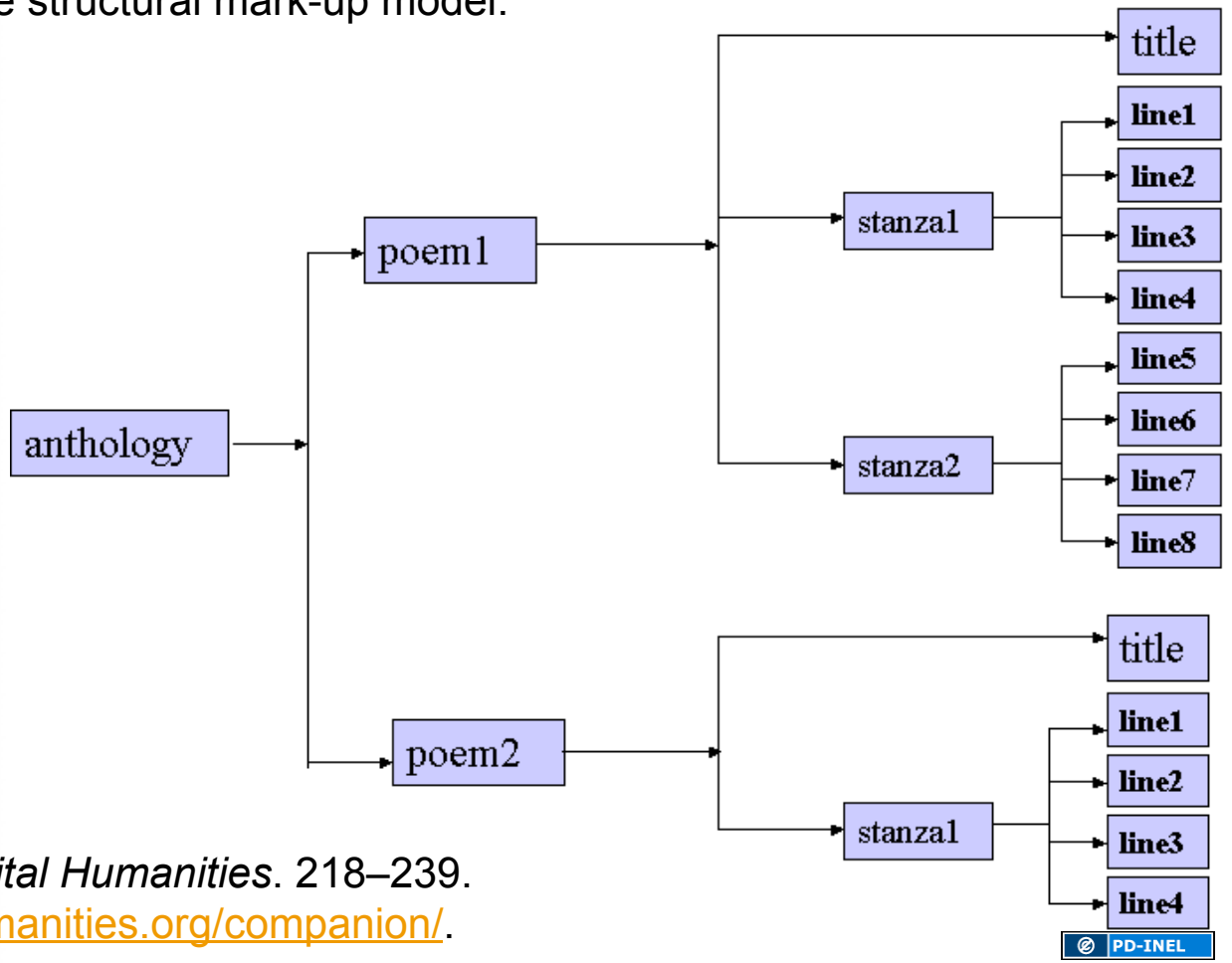
Markup

Representing Text

TEI

Examples

Support



A Companion to Digital Humanities. 218–239. http://www.digitalhumanities.org/companion/.

• Chapman, *Managing Text Digitization (2003)*

Text and Image

- Text is "Ordered Hierarchy of Content Objects" (OHCO)
- What is not text? [Biggs]

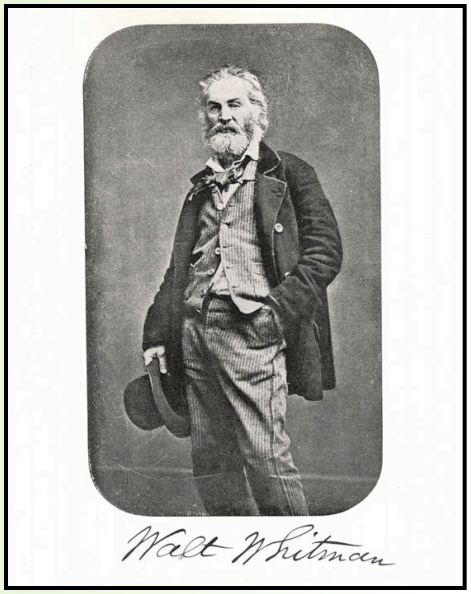
		linguistic content	non-linguistic content
1	sentential structure	<text>	
2	distributed-sentential structure	<notation> (maths, logic, etc)	<notation> (music)
3	diagrammatic structure		<figure> (graphics)

- Text, image, hybrid
- Presentation preferences
- Imaging and OCR
- Quality issues (image, text, product)

Biggs (2004) "What Characterizes Pictures and Text *Literary and Linguistic Computing* 19: 271.

• Barney, "Ordering Chaos" (2005)

Walt Whitman Archive



PD-EXP

"Visitors to Walt Whitman's home in Camden, New Jersey, described the poet, in his final years, as living in a tossed sea of scattered manuscripts that littered his floor." (p. 205)

- Item level information on all manuscripts
- Comprehensive index
- Elaborately marked up text [TEI]
- Images associated with text
- Integrated EAD finding aid as organizational tool

Whitman Archive: http://www.whitmanarchive.org/

Other Examples

- Rossetti Archive: http://www.rossettiarchive.org/
- Documenting the American South: http://docsouth.unc.edu/
- Making of America (UM): http://quod.lib.umich.edu/m/moagrp/
- Making of America (Cornell): http://cdl.library.cornell.edu/moa/
- NINES: http://www.nines.org/
- MONK: http://www.monkproject.org/
- Gutenberg: http://www.archive.org/details/gutenberg

• Sukovic, "Beyond the Scriptorium" (2002)

The Role of the Library

- Collectors or creators?
- Libraries: organization, access, dissemination
- Skills and support
- New alliances needed
- Sponsor digital humanities activities
  - IATH: http://www.iath.virginia.edu/
  - Digital Humanities Centers: http://digitalhumanities.pbwiki.com/
  - centerNet: http://digitalhumanities.org/centernet/

References

- Barney et al. [2005] "Ordering Chaos" *Library and Linguistic Computing*. [Whitman]
- Biggs, M. A. R. [2004] "What Characterizes Pictures and Text?" *Literary and Linguistic Computing* 19 (3).
- Buzzetti . [2002] "Digital Representation and the Text Model." *New Literary History* 33(1): 61–88 http://muse.jhu.edu/journals/new_literary_history/v033/33.1buzzetti.html
- Chapman. [2003] "Managing Text Digitization." *Online Information Review* 27 (1): 17-28.
- Functional Requirements for Bibliographic Records [FRBR]. http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records
- McGann. [2001] *Radiant Textuality: Literature After the World Wide Web*. New York, NY: Palgrave Macmillian, 2001. Chapter 3, "Editing as a Theoretical Pursuit," pp. 75-87.
- Renear. [2004] "Text Encoding." Susan Schreibman, Ray Siemans, and John Unsworth (eds.) *A Companion to Digital Humanities*. 218–239. http://www.digitalhumanities.org/companion/.
- Sukovic. [2002] "Beyond the Scriptorium: The Role of the Library in Text Encoding." *D-Lib* 8(1) http://www.dlib.org/dlib/january02/sukovic/01sukovic.html
- van der Weel. [ND] "The Concept of Markup." *Digital Text and the Gutenberg Heritage*. http://www.let.leidenuniv.nl/wgbw/~adriaan/Gut/Ch03_Concept_of_markup.fn.pdf.

Thank you!

Paul Conway
Associate Professor
School of Information
University of Michigan
www.si.umich.edu