Plain Text Corpora

Introduction to Digital Scholarly Editing

Gustavo Fernández Riva

January 2019

Fellow CONICET (Argentina)
Visiting Scholar IEM – FCSH – Universidade Nova Lisboa (Portugal)

Why Plain Text?

Due to its simplicity, plain text can be easily created, stored and analyzed:

- Creation: OCR (ex: AbbyFineReader) or manual transcription.
- Storage: as .txt. Occupies very little memory.
- Easy to import and process in any programming language (Python, R); excelent for some machine assisted analysis tasks (for example, stylometry).
- The Unicode encoding is very simple and explicit: it has no risk of becoming unreadable in the future.
- Foundation for later enrichment (for example, as XML)

Corpora

A corpus:

- A corpus is a sample, designed to represent a textual domain in a language.
- A digital corpus is a collection of “machine-readable” texts that enables different forms of computer-based analysis.
- A corpus will contain some canonical works, but it is not designed as a digital container for a canon.

Corpora Examples

- Project Gutenberg
- Open Medieval French
- Textos Portugueses Medievales
- CLIGS: TextBox
- Hispanic Seminary
- Wikipedia Corpus
- Corpus do Portugues