COMET/CMS/Fe6 Papers

GreenFIE

Problem:

Motivation:

Solution:

Validation:

Auto-Creation of a Semantically Searchable Family History Data Store

Problem:

Auto-creation of a Web of Searchable Data with imperfect extraction engines.

Motivation:

Hasten the work; convince FS to adopt and support the CMS admin system

Solution:

Use CMS to tune up an ensemble of extraction tools to produce satisfactory PRF for a sampling of pages for a book; then automatically process the full book. Extracted facts are immediately semantically searchable (keyword searchable too, HyKSS-like search).

Validation:

? PRF of a sampling of pages; informal analysis of the value of precision v. recall; expectations of end users and the possibility of achieving a 100% F-score by having them correct and complete the work done by the ensemble of extraction engines.

Accelerating Information Ingest Into an On-line Wiki-Based Family History Repository

Problem:

Ingesting information from family-history books into the tree is tedious, time-consuming, and error-prone.

Motivation:

Reduce the effort required to determine whether persons mentioned in a family-history book are already in the tree and if not, then to add them, link them to related persons, and document the additions.

Solution:

Using COMET to satisfy the requirement of human oversight and automation to do the actual ingest can significantly increase ingest efficiency. Using an ensemble of extraction tools, preprocess pages of family-history books for COMET users. COMET users check results and fix problems. A quality-check loop assists users in finalizing correct information extraction. Post-process information by standardizing person names, dates, and place names and by inferring gender and birth and married names. Generate for each page a gedcomx document of the information and an image of the page with extracted information highlighted, and for each person a person-info report of all BDM-event information and all marriage and parent-child relationships. Using the person-info report, automatically determine whether the person is already in the tree. (Duplicate verification can be human checked.) If not, upload a person-info-generated gedcomx to insert the person into the tree as well as related persons (as needed) and properly link together related individuals. If so, automatically insert any missing and any contradictory information. In all cases, automatically add documentation as image(s) of source page(s) with information highlighted.

Validation:

Compare time, effort, and error-proneness of ingesting information from a family-history book by hand to ingest via COMET. Do several case studies where time is measured for each step in the process and overall. A report of the time to do the computer processing should be included. Effort is related to time, but also related to the number, type, and complexity of human-involvement tasks. Error-proneness is also related to time and task but in a different way. From the same case studies, we can make observations about effort and error-proneness. We cannot do a full-blown user study, but the differences we can observe should be large enough so as to be convincing in favor of COMET.

Besides you and me as authors, we would need to include Joe, and I think we should also include Deryle and Scott. You should be first author with the rest of us following alphabetically.

PDFindexer

Problem: word/space identification

Motivation:

Solution:

Validation:

GreenDDA

Problem:

Motivation:

Solution:

Validation:

Layout-based Information Extraction

Problem:

Motivation:

Solution:

Validation: