Computational Linguistics for Metadata Building: Enhancing Subject Access Through Linguistic Techniques

Judith Klavans, Jimmy Lin, Rebecca Passonneau, Eileen Abels, Carolyn Sheffield, Dagobert Soergel, Tandeep Sidhu

Typical Record: Image with minimal metadata

Problems in image indexing and access
- Many images available with limited access points
- Subject-oriented cataloging is expensive, time-consuming, and requires domain expertise

Solution
- Apply computational linguistic techniques over text about images
- Identify potential subject terms
- Populate existing catalog records

Computational Linguistics Toolkit
- Available for download
- Being tested with catalogers

Benefits for Catalogers
- Provides candidate terms for catalogers to judge
- Links to thesauri (e.g. AAT, ULAN, TGN)
- Exports automatically to multiple metadata formats

Methodology
- Mine scholarly texts for subject-oriented metadata
- Identify relevant text segments and conceptual function
- Disambiguate terms and phrases by mapping to integrated thesauri (AAT, ULAN, TGN)
- Gather user feedback and perform iterative improvements

Export terms to catalog record

Mapping extracted terms to thesauri

Texts contain rich subject terms

University of Maryland

www.umiacs.umd.edu/~climb