Cornell Node Progress Summary September 2012

Comprehensive Extensible Data Documentation and Access Repository (CED²AR)

The NCRN-Cornell node is building a Comprehensive Extensible Data Documentation and Access Repository (CED²AR) designed to improve the documentation and discoverability of both public and restricted data from the federal statistical system.

Development so far:

- Completed high level functional diagram initial technical diagram of initial CED^{2}AR system (see https://confluence.cornell.edu/display/ncrn/Technical+Documentation)
- Evaluated initial DDI implementation strategy: DDI Lifecycle (3.x) or DDI Codebook (2.5) Settled on DDI Codebook for now; easier to implement and can migrate to future releases as needed. After testing, we decided against using a subset of DDI for the CED^{2}AR: while we may not use all elements, we will store all relevant elements.
- Digitial Object Identifiers (DOI's)
  - Developing formal NCRN Cornell specification for implementing DOI's for datasets.
  - NCRN Cornell will join Datacite, via the California Digital Library. Datacite provides a beta search engine to this metadata repository. The metadata specification for this is an application profile of Dublin Core, and the specification provides a straightforward mapping from DDI 3.12 this metadata format.
- Set up an initial metadata repository, using ICPSR metadata and metadata derived from SIPP Synthetic Beta (SSB) metadata. Subsequent data sets will include QWI, ACS, and IPUMS.

Interaction with Census

- Worked with the SIPP Synthetic Beta team to advise on tools, migrate existing metadata into DDI, and develop tools to maintain the documentation.

Statistical learning and classification

- Modern machine learning techniques for census applications: The goal is to develop boosting and ensemble-based statistical learning techniques to improve the integration, editing and imputation models for various applications in the Census Bureau, for example, assembling the micro-data for longitudinally linked employer-employee database.
- Another useful application is predicting multiple responses (e.g., multiple choices such as the race) commonly seen in survey studies. We have built the statistical models suitable for multiple responses and we have formulated the solutions using logistic regression as well as tree-based boosting algorithms.

Contact

ncrn@cornell.edu