Social Science Data Librarianship -
A University Curriculum

Frank Olken & Fredric Gey
Lawrence Berkeley National Laboratory
UC Data Archive & Technical Assistance, University of California
Berkeley, CA

presented to
IASSIST Conference 2006
Ann Arbor, Michigan
May 23, 2006
Disclaimer

These remarks constitute the opinions of the authors and do not reflect the policies of the University of California, Lawrence Berkeley National Laboratory, nor the United States Government.
Who is the curriculum for?

- Students preparing to become:
  - Social Science Data Librarians
    - at academic social science data libraries
  - Users of Social Science Data
    - at HMOs, at healthcare funding organizations
    - at city and regional planning offices
    - at financial and securities organizations
    - public policy think tanks, research/analysis units
  - Social Science Data Generators
    - Census, BLS, BEA, NCHS, EIA, ...
Where should SSDL Programs be taught?

- Computer and Information Science Departments
- Library and Information Science Schools
- Information Systems Depts within Business Schools
- Information Systems Schools
- Social Science Divisions with Letters and Science
- Public Health Schools
What are Social Science Data?

- anthropology
- communications
- criminology
- demography
- economics
- education
- human geography
- linguistics
- political science
- psychology
- sociology
- history
- city and regional planning
- public health
- social welfare
- business administration
- public policy
- public administration
Types of Social Science Data

- Statistical Macrodata: tables, ...
- Statistical Microdata: records, cases, ....
- Social network data
- Event histories
- Linguistic corpora
Social Science Data Librarianship Courses

• Social Science Data Sets (Domestic and Int'l)
• Statistical Data Management
• Metadata and Data Semantics
• Social Science Data Library Operations
• Statistical Disclosure Control
• Network Analysis Methods
Social Science Data Sets

• Survey course in major social science data sets

• Two semesters: domestic & international

• Topics:
  – demographic data
  – employment & labor statistics
  – health data
  – criminological data
  – political data
  – public opinion data
  – housing data

• Topics (cont.)
  – national income accounts
  – trade statistics
  – personal income
  – income distribution
  – price indices
  – financial statistics
  – energy statistics
  – education statistics
  – social welfare statistics
  – immigration statistics
  – migration statistics
Prototype Course at Cornell: Info 747 by Prof. John Abowd

- http://instruct1.cit.cornell.edu/courses/info747/
- Notable for its breadth
- Available as streaming video
- Lecture notes online in PPT and PDF
- Covers ethics of data collection and use (Tuskegee syphilis study, use of Census in WWII Japanese internment)
Statistical Data Management

- Macrodata vs. microdata
- OLAP, ROLAP, ...
- ETL for data warehousing
- Transposed files
- OLAP Operations
- Aggregation & aggregate indices
- Transposition
- Random sampling
- Covariance computations

- Record linkage
- Time series data & operations
- Temporal data, event histories
- Array and matrix data models
- Statistical model management
- Statistical metadata, for survey sample designs
Metadata & Data Semantics

- Ontologies
- Description logic
- Taxonomies
- Faceted Classification
- Mereology
- Temporal Relations
- Spatial Relations
- Dimensional analysis

- Ontology and schema mappings
- metadata registries
- model management
- federated data management (LAV, GAV, GLAV)
- schema/ontology evolution
Social Science Data Library Operations

- Reference services
- Cataloging & metadata creation
- Indexing and Retrieval
- Statistical processing
- Statistical consulting
- Social Hierarchy Management
- Digital Preservation
- Survey research center operations

- Production of official statistics
- Organizational Structures
- Acquisitions
- Outsourcing
Statistical Disclosure Control

• Data Privacy Issues
• Record linkage methods
• Statistical disclosure control
• Privacy preserving data mining
Proposed Degree Programs

• M.S. in SSDL
• Ph.D. in SSDL
• Dual Degrees in SSDL
  - M.S. in SSDL and M.S. in Social Science
    • e.g., MPH, MBA, Statistics,
  - M.S. in SSDL and Ph.D. in Social Science
Existing “SSDL” Graduate Programs

• Quantitative Methods in the Social Sciences,
  - M.A. degree, Columbia Univ., USA

• Joint Program in Survey Methodology
  - M.S. and Ph.D. degrees
  - U. Maryland, U. Michigan, Westat, all USA

• Computational Social Sciences
  - Ph.D. degree, George Mason Univ., USA

• Social / Official Statistics
  - M.S. and Ph.D. at U. Southampton, U.K.
Existing “SSDL” Graduate Programs

• Typically, emphasis is either on:
  – Statistical Methodology
  – Computation (e.g., social network analysis, simulation)

• Very little focus on:
  – data management
  – data semantics
Conclusions

• Proposed curriculum in Social Science Data Librarianship

• Various possible institutional homes

• Core content:
  − Survey of Social Science Data Sets
  − Statistical Data Management
  − Metadata and Data Semantics
  − Statistical Data Confidentiality
  − Network Analysis Methods
  − SSDL Operations
SSDL Web Page


• This page contains links to:
  - SSDL Curriculum IASSIST Paper
    • (and shortly slides)
  - SSDL Manifesto Paper
  - Other SSDL resources (coming soon)
Contact Information:

• Frank Olken
  - Email:  olken@lbl.gov
  - URL:   http://www.lbl.gov/~olken
  - Tel:   510-486-5891
  - Cell:  510-703-2764

• Fredric Gey
  - Email:  gey@berkeley.edu
  - URL:   http://ucdata.berkeley.edu/gey.html
  - Tel:   510-643-1298
Acknowledgements

• Frank Olken is currently supported by EPA and DOD to work on the XMDR project (Extended Metadata Registries). His previous work on random sampling from relational databases was supported by U.S. Dept. of Energy.

• Fred Gey was Assistant Director of UC DATA until July 2005. His research in information retrieval has been supported by IMLS, NSF and DARPA.