



UNC  
SCHOOL OF INFORMATION  
AND LIBRARY SCIENCE

# Data Science at SILS

Presented by

Barbara M. Wildemuth

Associate Dean for Academic Affairs

School of Information & Library Science, UNC-CH

[wildemuth@unc.edu](mailto:wildemuth@unc.edu)

# Current research initiatives

- Data Intensive Cyber Environments (DICE) Center, <http://dice.unc.edu/>
  - DataNet Federation Consortium – building national data infrastructure through federation of existing systems, <http://datafed.org/>
  - DataBridge – analyzing formation of collaboration communities through use of similar data sets, <http://databridge.web.unc.edu/>
  - Syndicate – collaboration on content-delivery system for academic data
- BitCurator, <http://www.bitcurator.net/>
  - BitCurator project – analyze disks for presence of protected data
  - BitCurator Access

# Infrastructure provisioning at UNC-CH

- Integrated Rule Oriented Data System (iRODS, <http://irods.org/>) – policy based data management
  - SILS LifeTime Library – digital library for student research
  - Carolina Digital Repository – research collections
  - In collaboration with RENCi, data management for:
    - ADCIRC storm surge modeling
    - HydroShare – hydrology watershed analysis automation
    - VISR – North Carolina virtual institute for social research
    - National Consortium for Data Science
    - iRODS Consortium
    - UNC-CH Genomics data grid
    - Lineberger Cancer Institute data grid
  - Institute for the Environment – hydrology watershed analysis automation
  - Odum DataVerse archive

# Current curriculum initiatives

- BSIS & MSIS – established, with strong offerings related to information retrieval, database management, and data science
- Post-Master's Certificate in Data Curation, <http://sils.unc.edu/programs/graduate/post-masters-certificates/data-curation>
- Graduate Certificate in Digital Curation, [http://sils.unc.edu/programs/certificates/digital\\_curation](http://sils.unc.edu/programs/certificates/digital_curation)
- Professional Science Master's in Digital Curation (proposal in development)
- Digital Curation Curriculum (DigCCurr) projects, <http://www.ils.unc.edu/digccurr/>
- Carolina Health Informatics Program (CHIP), <http://chip.unc.edu/>
  - PSM in Biomedical and Health Informatics
  - Certificates in clinical information science, nursing, and public health
  - PhD program in health informatics (proposal in development)

# Core components of data science

- Application/domain area(s)
  - Defining research questions, data needs and collection
- Data/metadata quality and validity
  - Description/metadata, provenance, intellectual control, handling heterogeneous data types, data management
- Storage, access, and preservation
  - Storage models and mechanisms, access to data sets, security, preservation
- Analysis and visualization
  - Statistical analysis, data/text mining, machine learning algorithms, algorithmic pre-processing of data, visualization
- Evaluation and interpretation
  - Evaluating the validity of the results, interpreting the results, understanding the implications of the results
- Legal and social issues
  - At beginning and end of process; privacy, intellectual property, security, socially impactful uses of results (negative or positive)

# Mapping strengths in data science at UNC

	SILS	CS	STOR	Math	Appl Phys Sci	Pharm	RENCI	Odum	...
Applications/domain area(s)									
Data/metadata quality and validity									
Storage, access, and preservation									
Analysis and visualization									
Evaluation and interpretation									
Legal and social issues									

Legend:

Less		More		
1	2	3	4	5