Teaching Scientific Data Management in Data Science Education and Workforce Development Programs for Science Communities

Robert R. Downs
rdowns@ciesin.columbia.edu

NASA Socioeconomic Data and Applications Center (SEDAC)
Center for International Earth Science Information Network (CIESIN)
The Earth Institute, Columbia University

2012 Microsoft eScience Workshop
Chicago, IL
Session: Educating Data Scientists for Scientific Data
October 9, 2012
Recent popularity of data science has led to increased recognition of the need for education and workforce development in data science. However, definitions of the term, data science, vary and often focus on techniques for data analytics and visualization, omitting scientific data management and related topics associated with data policy, stewardship, and preservation. Scientific data management encompasses a variety of concepts and methods to foster continuing access and long-term stewardship of data for current and future users. Considering the needs for scientific data management knowledge and capabilities to facilitate improved and persistent accessibility and use of scientific data throughout the data lifecycle, instruction on topics in scientific data management is recommended for data science education and workforce development programs for science communities.
“Big Data and Data Scientists -- It's An Issue Of Degree(s)”

Data Science Skills

• “Data science requires skills ranging from traditional computer science to mathematics to art” (Loukides, M. 2011 What is Data Science? O’Reilly Media) *

• “computer science, mathematics and analytical skills in the context of practical application” … “analyze and extract meaning from extremely large data sets” … “statistics, machine learning, text retrieval and natural language processing to analyze data and interpret results” (University of Washington Certificate in Data Science. Retrieved 10/8/2012 from http://www.pce.uw.edu/certificates/data-science.html) *

* Italics added
Examples of Data Science Programs

• Undergraduate
  – B.S. Degree in Computational and Data Sciences, George Mason University,

• Graduate
  – M. S. Degree in Data Science, University of Dundee (being planned),
    http://www.computing.dundee.ac.uk/study/postgrad/degreedetails.asp?17

• Certificate
  – Syracuse University Certificate of Advanced Study in Data Science,
    http://ischool.syr.edu/future/cas/datascience.aspx
• CODATA Data Science Journal http://www.codata.org/dsj/
  • Scope: Data, Database, Processing, Complexity, Scalability, Distribution, Interaction, Applications Interfaces, Models

• EPJ Data Science http://www.epjdatascience.com/
  • Scope: Data extraction, analysis, generation, new empirical laws or fundamental theories on systems

• The Journal of Data Science http://www.jds-online.com/
  • Scope: Collecting, analyzing, modeling, applications, all aspects of applied statistics, general applications of quantitative methods
Examples of Data Management Learning Units

• Data management planning
• Data policy
• Data accessibility and use
• Data stewardship and preservation
Examples of Data Management Planning Topics

- Data management plans
- Archive selection
- Organizational commitment, governance, and sustainability
- Data management system organization and expertise
- Data project planning
- Data authorship
- Data curation
- Distribution
Examples of Data Policy Topics

- Intellectual property rights
- Privacy
- Confidentiality
- Restrictions
- Security concerns
Examples of Data Accessibility and Use Topics

- Data description and documentation
- Standards for content and metadata
- Interoperability and integration
- Data dissemination
- Provenance
- Persistent identification
- Open access
- Data citation
- Assessing data use and impact
Examples of Data Stewardship Topics

- Preparing data for archiving
- Data submission and acquisition
- Data appraisal
- Software dependencies
- Integrity validation and verification
- Data migration
- Data service levels
- Long-term preservation
Challenges for Teaching Data Management

• Time allocation
  – Additional topics require more time
  – Could reduce time for other topics and experiences

• Lack of interest
  – Immediate benefits of knowledge not obvious
  – Practical applications less apparent

• Need expertise
  – Additional knowledge required
  – Draws on various disciplines
Teaching Intellectual Property Rights in Practice

- Archival acquisition, storage, and management
  - Rights to be assigned to the archive during submission
- Use
  - Rights to be assigned to users
  - Qualifiers or restrictions on uses
- Integrate
  - Rights assigned to derivative products
- Distribute
  - Rights assigned to distributor
- Others distribute
  - Rights for 3rd party distribution
- Long-term use
  - Expiration of rights
Demonstrating Choices for Assigning Rights

• Public Domain – not copyrighted, permission is not required
  – Works created by some government agencies may be in the public domain
  – Works in the public domain can be marked to identify them
  – Rights can be waived by applying a Creative Commons 0 (CC0) license

• CreativeCommons.org licenses for copyright holder to allow use, dissemination, derivation, integration
  – CC By License: requires attribution (cite the source)
  – CC By-ND: requires attribution, no derivatives allowed
  – CC By-NC License: requires attribution, no commercial use allowed
  – CC By-SA License: requires attribution, share alike using same license

• Apache.org license for copyright holder of scripts and computer code
  – Applied to code to allow “worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute” (Apache License 2.0, 2004)

Improving Understanding on Open Science

- Open Access
  - Publications
  - Data

- Open Source
  - Software (code, scripts, algorithms, models, documentation, guides, …)

- Open Standards
  - Terms, definitions, procedures
  - Schemas
Ways to Teach Data Mgt in Data Science Programs

• Introduce data science with data management topics
  – Data archiving enables use by future users
  – Data dissemination enables use by others

• Include data management in ethics modules
  – Data science also needs instruction on ethical issues
  – Data management illustrates and addresses many ethical issues
  – Providing rights enables archiving, dissemination, and use

• Include analysis of data management in statistical modules
  – Archival processing status
  – Data citations in publications by dataset, producer, or repository
Example: Teaching Data Management in Data Science

“confidential data management, repository requirements and assessment, effective documentation practices, and how to create, comply with, and evaluate required data management plans” (ICPSR Summer Program, Applied Data Science: Managing Data for Re-use. Retrieved 10/8/2012 from http://www.icpsr.umich.edu/icpsrweb/sumprog/courses/0149)*
Benefits of Data Management in Data Science

- **Discovery**
  - Data that have been described in catalogs can be found

- **Accessibility**
  - Data that are available from archives can be accessed

- **Reuse**
  - Documented data available without restrictions can be reused

- **Interoperability**
  - Data adhering to standards can be integrated and used with other data

- **Stewardship**
  - Curated open data can be preserved for future use