IST 400/600 – John D’Ignazio

Lecture: Data and Users

- Defining communities
- Needs assessment
- Evaluating use
- Dataset characteristic analysis

Dr. Joel Trexler’s Aquatic Ecology Labgroup FIU

Defining the “who”

- Understanding the community clarifies purpose of a managed data archive
- Helps guide decisions on
  - Documentation (level, amount, terminology)
  - Formats
  - Presentation
- Digital and web-based means data may cross disciplinary or publication boundary
A Broadening Circle

• Starts with the researcher associated lab group or science team
  – Restricted, proprietary access to allow for publications
  – Shared terminology, sense of purpose
  – Shared understanding
    • context for the questions asked
    • context for the decisions made

• Roles and position
  – expertise, access, methods

A Broadening Circle

• Funders
  – Want to leverage investment
  – increasingly requiring data publication:
  • provision (deposit) vs. Mandate (creation) of archive
  • degree of specification/standards

• Disciplinary subfield of researchers
  – Similar expertise with labgroup
  • education/terminology/knowledgebase
  • investigation of similar phenomena
  – Lack the context of the local lab group
  • assumptions, imperatives, compromises
A Broadening Circle

• Other/inter disciplines
  – Adopt for their research agenda
  – Integrate with different tools, techniques, assumptions
    • Increased burden to clarify data constraints

• Public
  – Define terminology, provide summary/metaphor
    • News media
    • Technology development
    • Education

Needs Assessment: Guiding what could be

• Matching what people need with what can be built for them
  – Standards
    - Systems
    - Specifications

• Systems-oriented tools to reveal
  – Unspoken and undocumented practices
  – Theoretical and methodological orientation
  – Tacit knowledge, paradigms
Needs Assessment: tools

• Cognitive work analysis
  – Studies interplay of work, actors, and activities
  - Allows detailed mapping of relations and knowledge

• UML
  – Sequential process
  – Team needs generate requirements
  – Requirements generate system design

Evaluating use: Capturing what is

• Counting uses
• Eliciting user practices and tacit knowledge
  - Surveys
  - Interviews
  - Focus Groups
• Observation
Counting uses -- which kinds

• Concentrate on place
  Amount / Categories of data
  – Research programs
  – Collections

• Concentrate on system
  – log analysis
  – Cost assessments

• Usage patterns over time
  – highlight use of particular data
  – judge by relative cost to support

Counting uses -- how to’s

• Determine type of data appropriate to question
  – accumulation versus sampling
  – automatic versus manual
  – Examples: Page hits / IP addresses, lab visits, query logs, instrument time…

• Determine analysis
  – what can the data say?
  – appropriate tools and techniques (SW, stats pkgs)
Eliciting user practices-- which kinds

• Surveys
  – Reports of patterns & frequency of use, ease & success
  – Relevance and ranking
  – Satisfaction/perceptions of quality
  – Can sort data by discipline / institution

• Interviews & focus groups
  – Reveals greater depth/complexity -- needs to be focused
  – Direct access: “in their own words”
  – individual/group level

Eliciting user practices-- how to’s

• Surveys
  – Determine format (telephone, web-based, in-person)
  – Develop questions and presentation order related to scope
  – Validity requires sampling method and number

• Interviews & focus groups
  – Train moderator according to guide/script
  – Set conducive environment
  – Solicit appropriate participants
Observation -- which kinds

• Aspect of use-oriented task performance
  – Think alouds for system use
  – Capture samples in time and space
• Identify problems in existing implementation
  – Where are problems? What type occur?
  – Verify improvements worked

Observation -- how to’s

• Protocol-based: Structured and Defined
• Capture samples in periodic time and consistent space
• Record with minimal interpretation
• Data analysis
The goal: An informed data management solution

- Matching people with dataset characteristics results in a stronger system
- Barrier to use and reuse of data is lowered
  - Knowledge is created
  - Citations and publications grow
- Scientists/Funders are happy!

Findings from the Snow & Ice Data Center (Parsons & Duerr)

- Challenge your assumptions about the knowledge of your data users and creators
- Write in language and detail appropriate to your user community
- Understand and describe data uncertainty
- Keep data formats, metaformats, and profiles simple and flexible
- Address issues of transparency, interoperability, extensibility, and storage volume