Cyberinfrastructure for data curation

Greg Janée
UC Santa Barbara; CDL

- Data curation:
  - management of data throughout its lifecycle,
  - such that it can be kept usable in the future,
  - affordably

- Three eras of science data

  1. analog
  2. digital
  3. online

- New norms for science data
  - discoverable by searching
  - available: at any time, immediately, in the future
  - usable
  - persistently identified and citable
  - supportive of replication (versioned)
  - meta-information: reviews, uses, provenance
  - linked to and within scholarly literature
- Who's going to do all this?

- Data curation research at UCSB
  - survey, faculty interviews, case studies
  - great interest (1/3 response to survey)
  - broad applicability (90% of departments)
  - curation mandated (50%)
  - researchers personally responsible for data (90%)
  - http://tinyurl.com/ucsb-data-curation

- But...
  - requested help with every aspect of curation
  - researcher motivations not aligned with curation

- Prototypical researcher:
  - focused on research area
  - resourceful in utilizing new tools, techniques
  - not knowledgeable of curatorial aspects of tools
  - not expert in data management
  - time- and resource-constrained
  - views data management as important but secondary

- What about libraries?
  - cultural heritage institutions
  - curation expertise
  - metadata, cataloging, search expertise
  - missing: experience working with data earlier in the lifecycle

- Cyberinfrastructure for data curation
  - Services
    - generic (figshare), discipline-specific (GenBank)
    - systemwide (CDL)
    - campus (institutional repositories)
  - Libraries
    - awareness, identification of curation issues
    - navigation of service space
    - education
    - assistance with projects
    - relationships with researchers
  - Researchers
    - motivations unchanged