Data Curation @ UCSB:
Preliminary findings & recommendations

Greg Janée & James Frew

November 5, 2013

Outline

- *Survey recap*
- *In-depth investigations*
  - 6 observations
- *3 recommendations*
- *Next steps*

Survey findings

- *Curation of digital data is a concern for a significant proportion of UCSB faculty and researchers…*
  - …and for almost every department and unit on campus.

- *Researchers almost universally view themselves as personally responsible for the curation of their data…*
  - …and view curation as a collaborative activity and collective responsibility.

- *Departments have different curation requirements, and therefore may require different amounts and types of campus support.*

Survey findings

- *Researchers desire help with all data management activities related to curation, predominantly storage…*
  - …but may be underestimating the need for help using archival storage systems and dealing with attendant metadata issues.
- *There are many sources of curation mandates, and researchers are increasingly under mandate to curate their data.*
- *Researchers under curation mandate are more likely to collaborate with other parties in curating their data…*
  - …and request more help with all curation-related activities; put another way, curation mandates are an effective means of raising curation awareness.
- *The survey reflects the concerns of a broad cross-section of campus.*

In-depth investigations

- *Informal process*
  - Emailed questions, background research, sit-down interviews
- *Departments represented*
  - ERI
  - ISBER
  - Geography
  - Geology
  - LTERs
  - MCDB
  - MSI
  - Theater and Dance
  - (others in progress)

Observation: transitional time

- *From our introductory talk, new norms for science data:*
  - online
  - instantly and forever available
  - (re)usable
  - discoverable
  - identified and citable
  - hyperlinked into fabric of scholarly communication

- *But tools haven't caught up yet*

Ten Simple Rules for Reproducible Computational Research

PLOS computational biology, doi:10.1371/journal.pcbi.1003285

- *For every result, keep track of how it was produced*
- *Avoid manual data manipulation steps*
- *Archive the exact versions of all external programs used*
- *Version control all custom scripts*
- *Record all intermediate results, when possible in standardized formats*
- *For analyses that include randomness, note underlying random seeds*
- *Always store raw data behind plots*
- *Generate hierarchical analysis output, allowing layers of increasing detail to be inspected*
- *Connect textual statements to underlying results*
- *Provide public access to scripts, runs, and results*

Chris Lynnes' (NASA GSFC) response on ESIP-preserve-l

"Those rules may be simple, in that they are written pithily, but there is a world of complexity beneath nearly every single one of them. […]

If only there were a cross-platform, easy-to-install, easy-to-manage 'science management system' designed for individual scientists, with full support for managing:

- *archiving*
- *provenance*
- *workflows*
- *analysis results*
- *annotations*
- *versions (of every artifact above)*
- *secure public access*

How hard could it be? :-)"

Observation: transitional time

- *Researchers comfortable with traditional tools, techniques*
  - Local storage
  - Local tools
  - Copy data to FTP directories
- *Discomfort with/unawareness of newer tools*
  - Version control systems
  - Persistent identifiers, automating tracking
- *Lack of tools*
  - Provenance

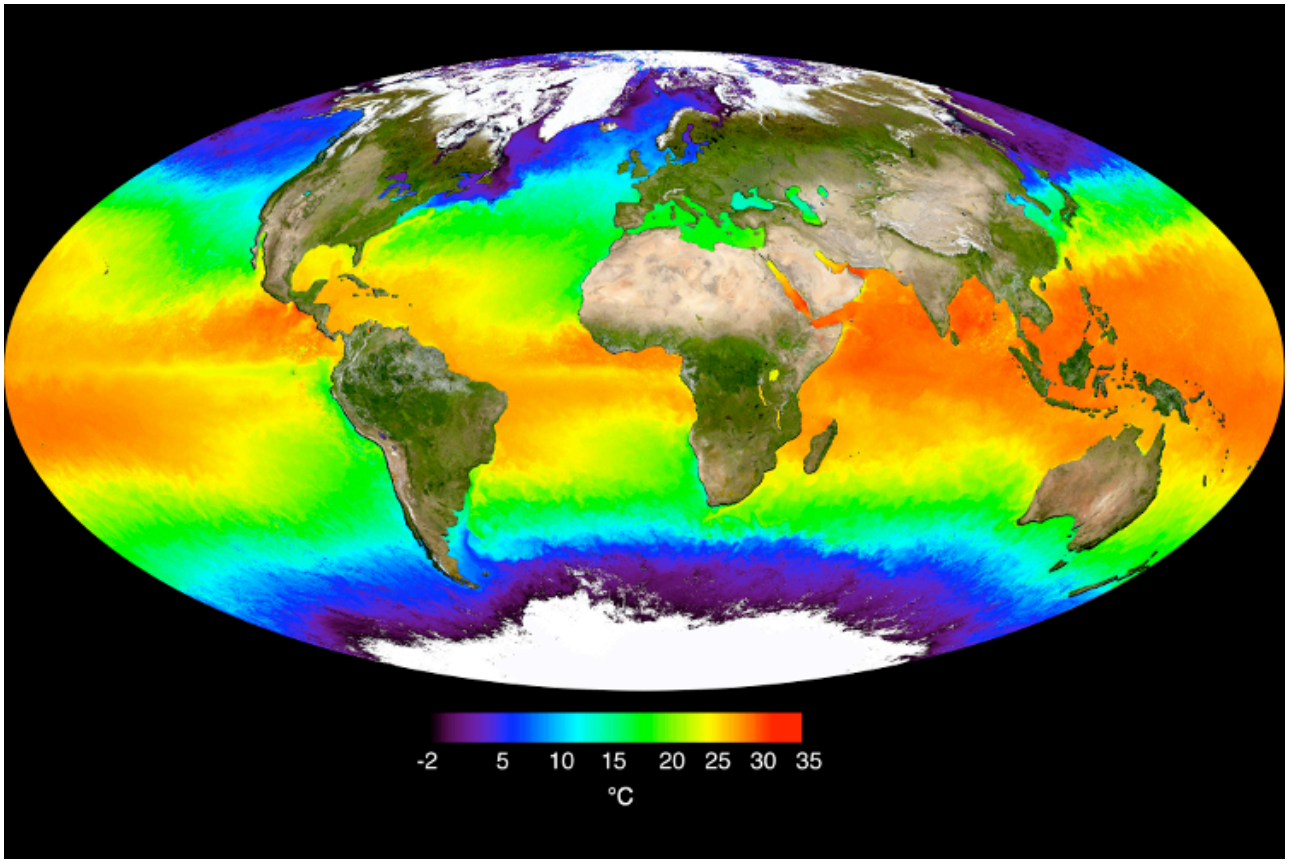
Ten Simple Rules for Reproducible Computational Research

PLOS computational biology, doi:10.1371/journal.pcbi.1003285

- ✘ *For every result, keep track of how it was produced*
- ✘ *Avoid manual data manipulation steps*
- ✘ *Archive the exact versions of all external programs used*
- ✘ *Version control all custom scripts*
- ✓ *Record all intermediate results, when possible in standardized formats*
- *For analyses that include randomness, note underlying random seeds*
- ✓ *Always store raw data behind plots*
- *Generate hierarchical analysis output, allowing layers of increasing detail to be inspected*
- ✘ *Connect textual statements to underlying results*
- ✘ *Provide public access to scripts, runs, and results*

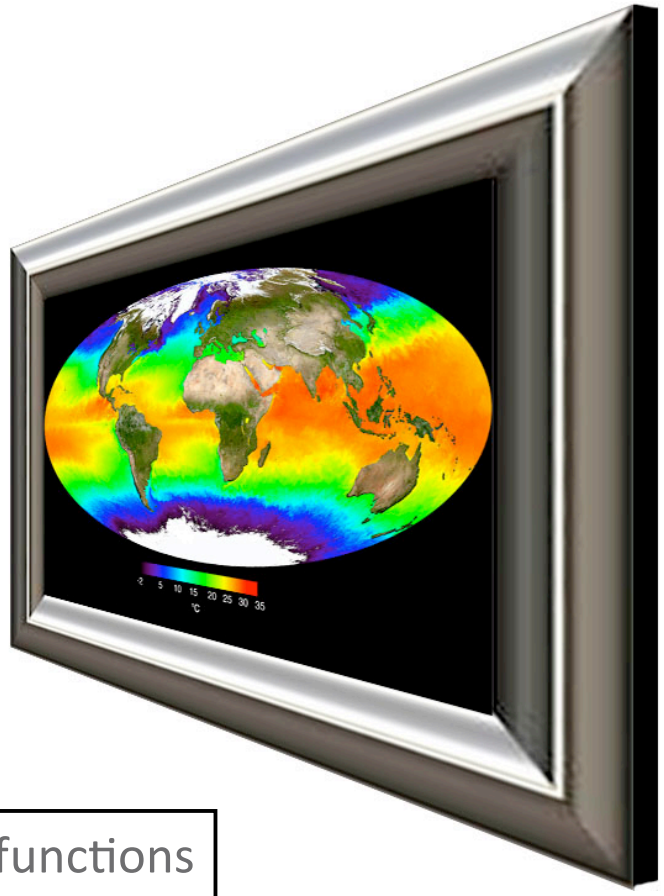
Observation: research focus

- *Researchers are focused on their research, their data*



Observation: research focus

researcher →



← curator

Implication: curatorial functions that must be performed by researchers need to be easy and/or automated.

Observation: limited resources

- *As with all of us:*
  - What must get done, gets done
  - Some of what is desired gets done
  - Much is left undone

curation

Observation: complexity

- *Surface simplicity often belies deeper complexities*
- *Example: fossil image collection*

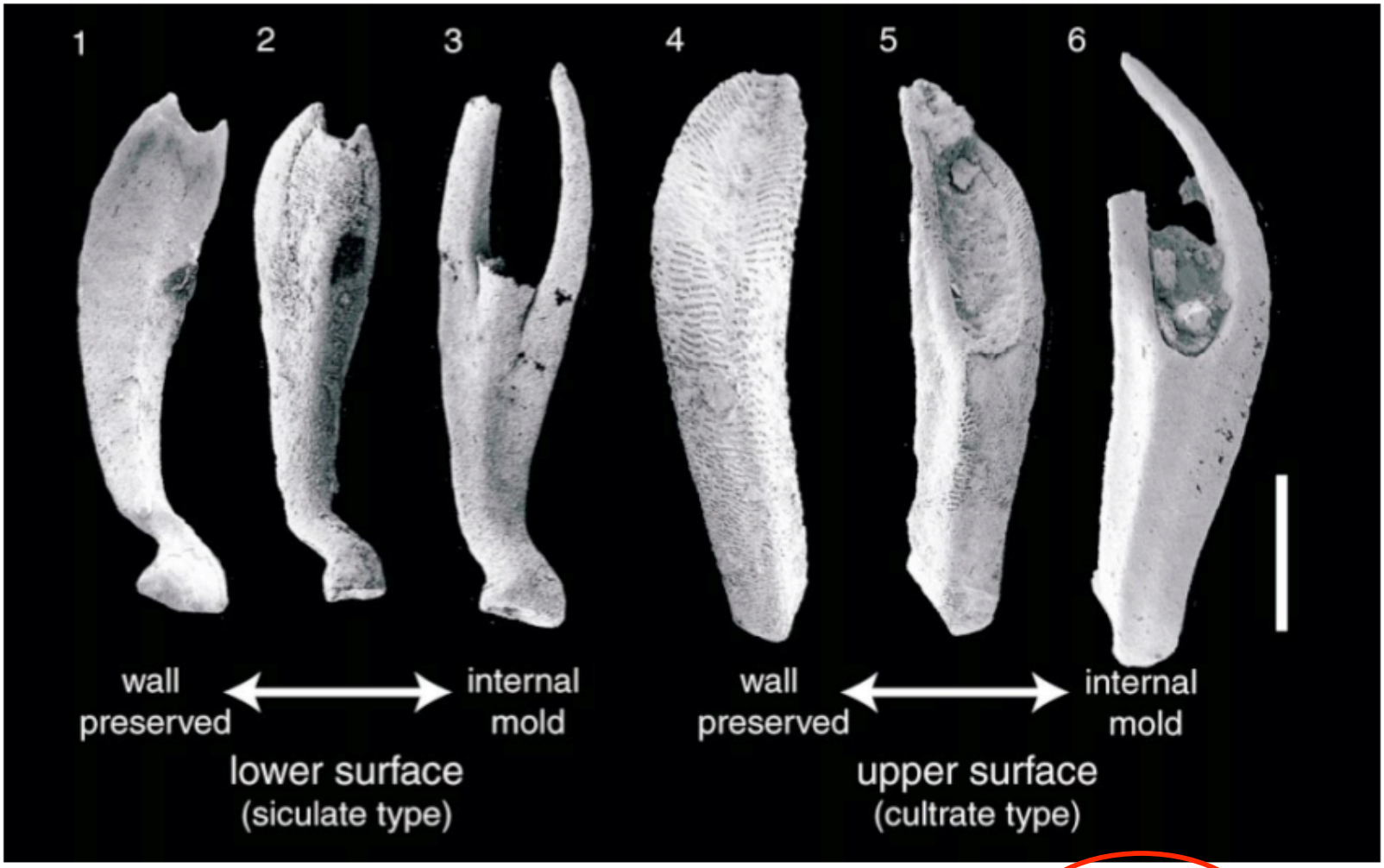
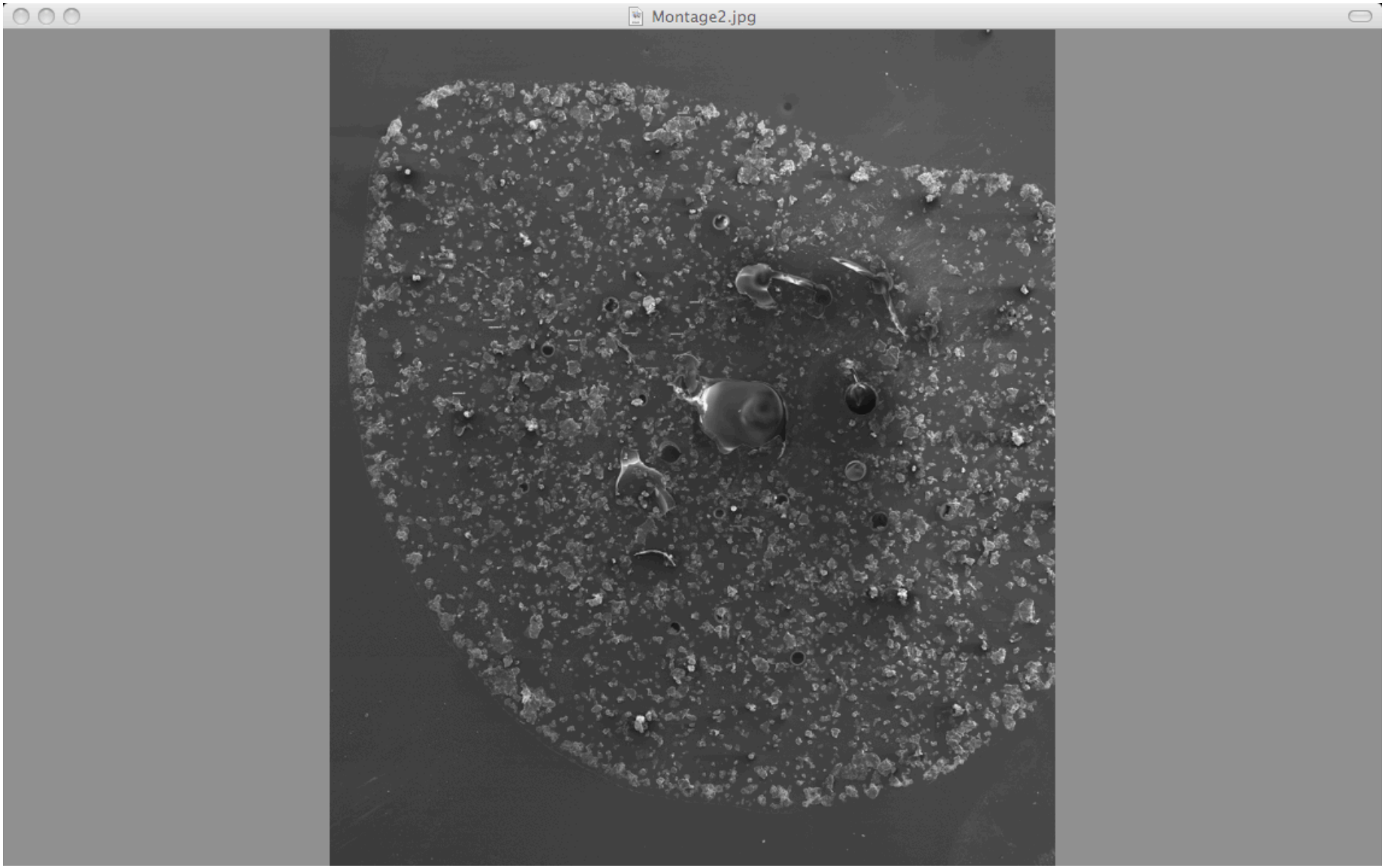
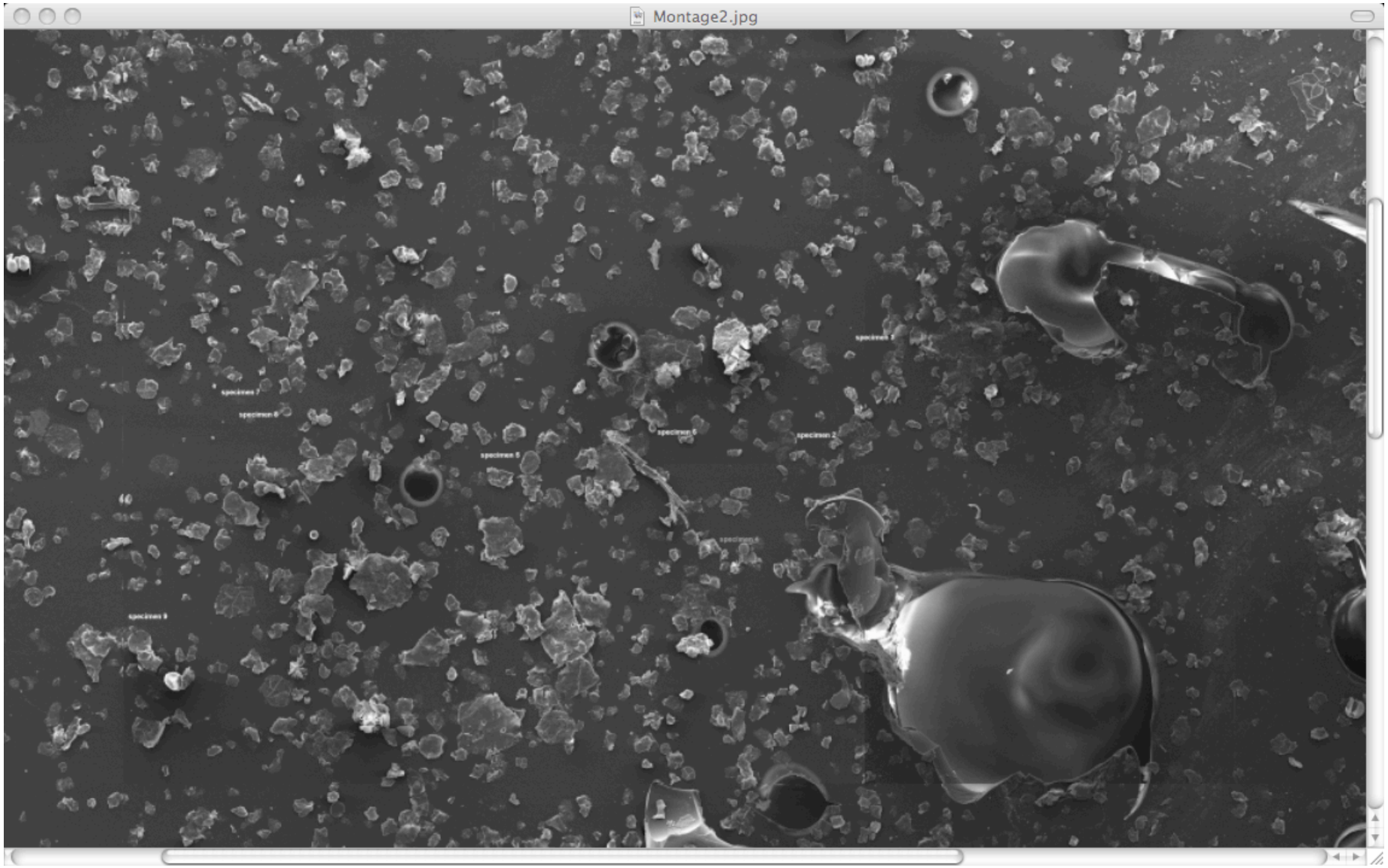
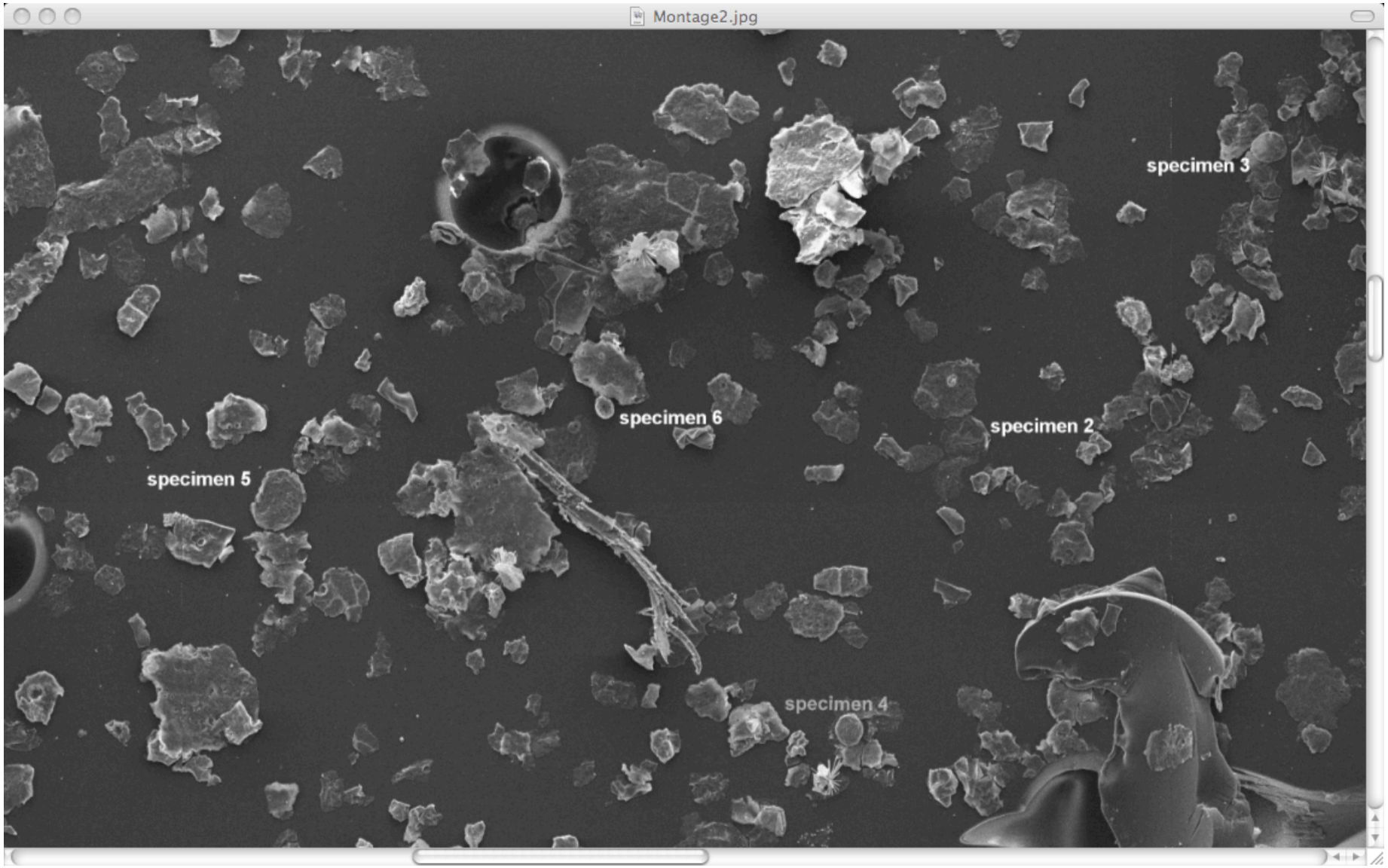
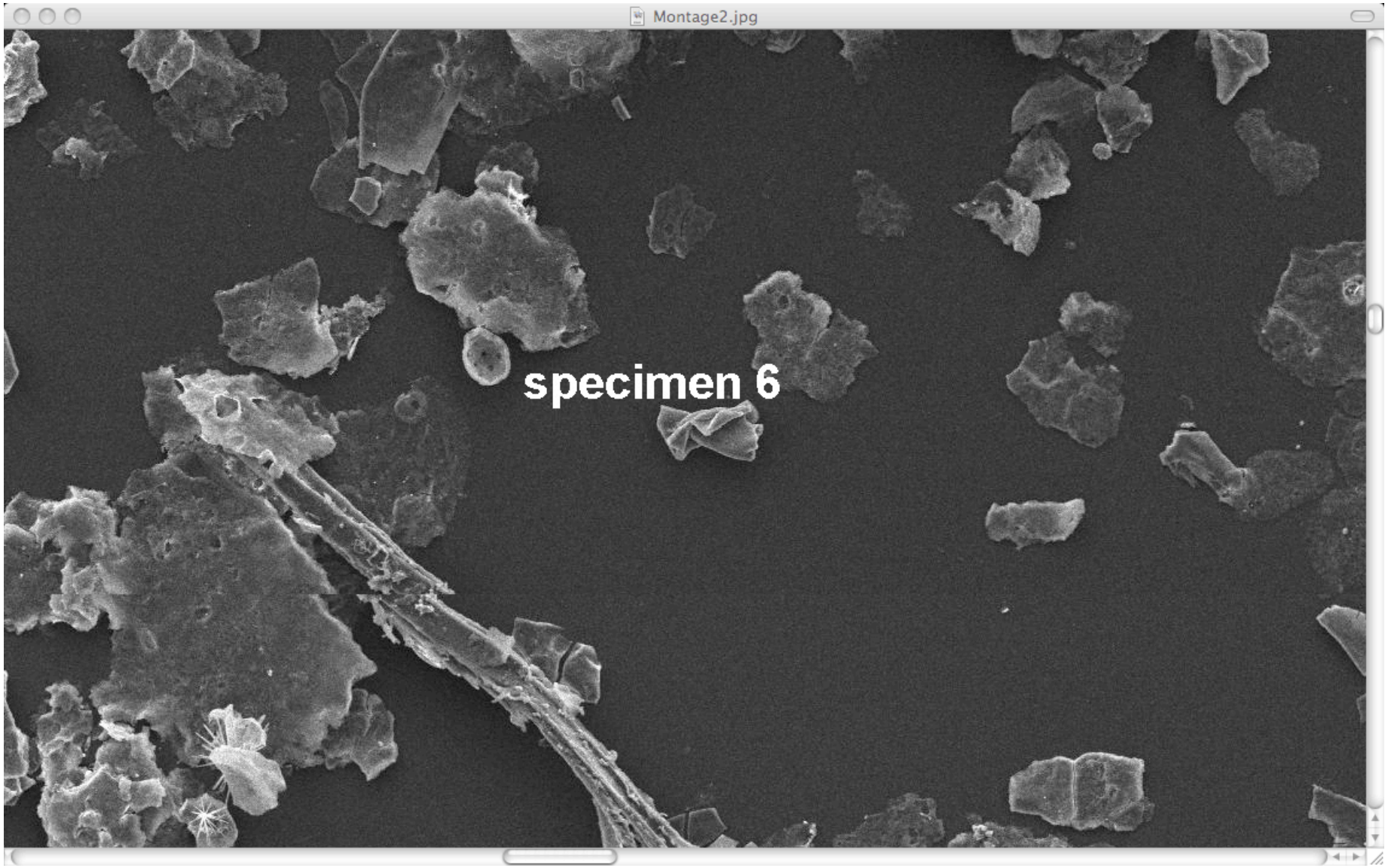


FIGURE 4—Examples of *A. superstes* blade structure. *1–3*, views of the lower surface of siculate sclerites with varying degree of wall preservation; *1*, CPC 37211; *2*, CPC 37212; *3*, CPC 37170. *4–6*, as in *1–3*, but for the upper surface of cultrate sclerites; *4*, CPC 37189; *5*, CPC 37190; *6*, CPC 37171. Scale bar represents 150 μm for *1*, *2*, *4*, *5*; 110 μm for *3*; and 90 μm for *6*.


Montage2.jpg


Montage2.jpg


Montage2.jpg
specimen 3
specimen 6
specimen 2
specimen 5
specimen 4


Montage2.jpg
specimen 6

Deeper complexities

- *Specimen map documents*
- *Consistent, unambiguous, persistent identification*
- *Image-specimen-map-slide linkages*
- *And:*
  - Additional, unpublished images
  - Intellectual privacy, embargoes
  - Copyright, use in textbooks
  - Lack of metadata (of course)
  - Lots of contextual semantics known only to researcher

Observation: increased publicness

- *Technological changes, concerns over integrity are driving new openness in how science is performed*

- *PLOS "Ten Simple Rules" paper:*
  - "…scientific papers commonly leave out experimental details essential for reproduction … difficulties with replicating published experimental results … increase in retracted papers … high number of failing clinical trials…"
  - "This has led to discussions on how individual researchers, institutions, funding bodies, and journals can establish routines that increase transparency and reproducibility. In order to foster such aspects, it has been suggested that the scientific community needs to develop a 'culture of reproducibility' for computational science, and to require it for published claims."

Traditional split

private | public

http://thevintagestandard.com/?p=1296

Zur affinen Feldtheorie.

Von A. Einstein.

Weiteres Nachdenken führte mich zu einer Vervollkommnung der in zwei früheren Mitteilungen behandelten Theorie des Feldes der Gravitation und Elektrizität. Ich will die Theorie in ihrer neuen Form im folgenden kurz darstellen.

Der affine Zusammenhang sei durch die 40 Funktionen $\Gamma^{\nu}_{\mu\nu}$ dargestellt. Der Riemannsche Krümmungstensor zweiten Ranges $R_{\mu\nu}$ sei in einen symmetrischen Teil $\gamma_{\mu\nu}$ und einen antisymmetrischen Teil $\phi_{\mu\nu}$ gespalten, so daß man hat

$$\gamma_{\mu\nu} = -\frac{\partial \Gamma^{\alpha}_{\mu\nu}}{\partial x_\alpha} + \Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\alpha} + \frac{1}{2}\left(\frac{\partial \Gamma^{\alpha}_{\mu\alpha}}{\partial x_\nu} + \frac{\partial \Gamma^{\alpha}_{\nu\alpha}}{\partial x_\mu}\right) - \Gamma^{\alpha}_{\mu\nu}\Gamma^{\beta}_{\alpha\beta}, \quad (1)$$

$$\phi_{\mu\nu} = \frac{1}{2}\left(\frac{\partial \Gamma^{\alpha}_{\mu\alpha}}{\partial x_\nu} - \frac{\partial \Gamma^{\alpha}_{\nu\alpha}}{\partial x_\mu}\right). \quad (2)$$

Die Hamiltonsche Funktion $\mathfrak{H}$ (skalare Dichte) sei eine vorläufig unbekannte Funktion[1] der $\gamma_{\mu\nu}$ und $\phi_{\mu\nu}$. Das Hamiltonsche Integral ist zu variieren nach den $\Gamma^{\nu}_{\mu\nu}(= \Gamma^{\nu}_{\nu\mu})$. Man erhält zunächst

$$\int (\mathfrak{g}^{\mu\nu}\delta\gamma_{\mu\nu} + \mathfrak{f}^{\mu\nu}\delta\phi_{\mu\nu})d\tau = 0, \quad (3)$$

wobei zur Abkürzung gesetzt ist

$$\left.\begin{aligned} \frac{\partial \mathfrak{H}}{\partial \gamma_{\mu\nu}} &= \mathfrak{g}^{\mu\nu} \\ \frac{\partial \mathfrak{H}}{\partial \phi_{\mu\nu}} &= \mathfrak{f}^{\mu\nu}. \end{aligned}\right\} \quad (3a)$$

$\mathfrak{g}^{\mu\nu}$ und $\mathfrak{f}^{\mu\nu}$ seien als Tensordichten des metrischen und elektrischen Feldes aufgefaßt. Setzt man die durch (1) und (2) gegebenen Ausdrücke von $\gamma_{\mu\nu}$ und $\phi_{\mu\nu}$ in (3) ein, so erhält man durch Variation die Gleichungen

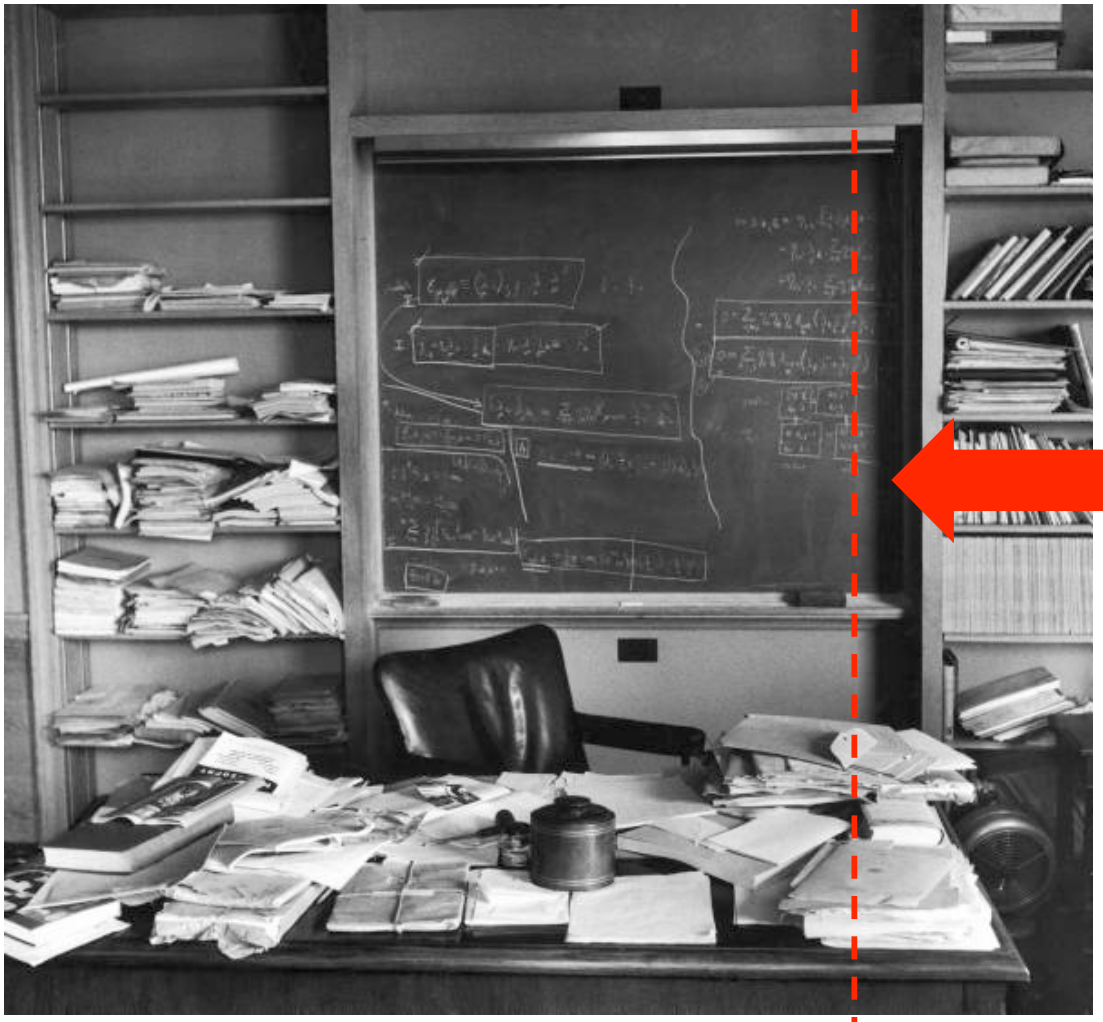
$$\mathfrak{g}^{\mu\nu}{}_{;\alpha} - \frac{1}{2}\mathfrak{g}^{\mu\tau}{}_{;}\delta^{\nu}_{\alpha} - \frac{1}{2}\mathfrak{g}^{\nu\tau}{}_{;}\delta^{\mu}_{\alpha} - \frac{1}{2}\mathfrak{f}^{\mu\tau}\delta^{\nu}_{\alpha} - \frac{1}{2}\mathfrak{f}^{\nu\tau}\delta^{\mu}_{\alpha} = 0, \quad (4)$$

[1] Die Voraussetzung, daß $\mathfrak{H}$ nur von $\gamma_{\mu\nu}+\phi_{\mu\nu}$ abhänge, wird aufgegeben. Ich möchte ferner bemerken, daß Hr. Donder in Leiden schon vor zwei Jahren ähnliche Überlegungen angestellt aber nicht publiziert hat.

http://www.aip.org/history/einstein/images/ae70.jpg

Today

private | public



http://thevintagestandard.com/?p=1296

Zur affinen Feldtheorie.

Von A. Einstein.

Weiteres Nachdenken führte mich zu einer Vervollkommnung der in zwei früheren Mitteilungen behandelten Theorie des Feldes der Gravitation und Elektrizität. Ich will die Theorie in ihrer neuen Form im folgenden kurz darstellen.

Der affine Zusammenhang sei durch die 40 Funktionen $\Gamma^{\nu}_{\mu\sigma}$ dargestellt. Der Riemannsche Krümmungstensor zweiten Ranges $R_{\mu\nu}$ sei in einen symmetrischen Teil $\gamma_{\mu\nu}$ und einen antisymmetrischen Teil $\phi_{\mu\nu}$ gespalten, so daß man hat

$$\gamma_{\mu\nu} = -\frac{\partial \Gamma^{\alpha}_{\mu\nu}}{\partial x_\alpha} + \Gamma^{\alpha}_{\mu\beta}\Gamma^{\beta}_{\nu\alpha} + \frac{1}{2}\left(\frac{\partial \Gamma^{\alpha}_{\mu\alpha}}{\partial x_\nu} + \frac{\partial \Gamma^{\alpha}_{\nu\alpha}}{\partial x_\mu}\right) - \Gamma^{\alpha}_{\mu\nu}\Gamma^{\beta}_{\alpha\beta}, \qquad (1)$$

$$\phi_{\mu\nu} = \frac{1}{2}\left(\frac{\partial \Gamma^{\alpha}_{\mu\alpha}}{\partial x_\nu} - \frac{\partial \Gamma^{\alpha}_{\nu\alpha}}{\partial x_\mu}\right). \qquad (2)$$

Die Hamiltonsche Funktion $\mathfrak{H}$ (skalare Dichte) sei eine vorläufig unbekannte Funktion[1] der $\gamma_{\mu\nu}$ und $\phi_{\mu\nu}$. Das Hamiltonsche Integral ist zu variieren nach den $\Gamma^{\nu}_{\mu\sigma}(= \Gamma^{\nu}_{\sigma\mu})$. Man erhält zunächst

$$\int (\mathfrak{g}^{\mu\nu}\delta\gamma_{\mu\nu} + \mathfrak{f}^{\mu\nu}\delta\phi_{\mu\nu})\,d\tau = 0, \qquad (3)$$

wobei zur Abkürzung gesetzt ist

$$\left.\begin{aligned} \frac{\partial \mathfrak{H}}{\partial \gamma_{\mu\nu}} &= \mathfrak{g}^{\mu\nu} \\ \frac{\partial \mathfrak{H}}{\partial \phi_{\mu\nu}} &= \mathfrak{f}^{\mu\nu}. \end{aligned}\right\} \qquad (3a)$$

$\mathfrak{g}^{\mu\nu}$ und $\mathfrak{f}^{\mu\nu}$ seien als Tensordichten des metrischen und elektrischen Feldes aufgefaßt. Setzt man die durch (1) und (2) gegebenen Ausdrücke von $\gamma_{\mu\nu}$ und $\phi_{\mu\nu}$ in (3) ein, so erhält man durch Variation die Gleichungen

$$\mathfrak{g}^{\mu\nu}{}_{;\alpha} - \frac{1}{2}\mathfrak{g}^{\mu\tau}{}_{;}\delta^{\nu}_{\alpha} - \frac{1}{2}\mathfrak{g}^{\nu\tau}{}_{;}\delta^{\mu}_{\alpha} - \frac{1}{2}\mathfrak{f}^{\mu\tau}\delta^{\nu}_{\alpha} - \frac{1}{2}\mathfrak{f}^{\nu\tau}\delta^{\mu}_{\alpha} = 0, \qquad (4)$$

[1] Die Voraussetzung, daß $\mathfrak{H}$ nur von $\gamma_{\mu\nu}+\phi_{\mu\nu}$ abhänge, wird aufgegeben. Ich möchte ferner bemerken, daß Hr. Donder in Leiden schon vor zwei Jahren ähnliche Überlegungen angestellt aber nicht publiziert hat.

http://www.aip.org/history/einstein/images/ae70.jpg

Observation: ownership

- *Researchers want to keep a local copy of their data*
  - Deeper motivations than redundancy

Recommendation 1

- *Develop repository/curation services supporting small-scale data producers*
  - Such data is often uncurated, unsupported
  - Images, spreadsheets, tabular data, GIS data

- *Lots of models:*
  - Berkeley Research Hub
  - Purdue University Research Repository (PURR)
  - CDL/UCSF DataShare
  - DataUp for Excel spreadsheets

- *Underlying repository not important; service is*

Implementation challenges

- *Success =*
  - If service is used, and if use of service results in data being curated
- *How to motivate researchers to use new tools?*
  - Guilt or shame is unlikely to work
    - Though DMPs may provide motivation
  - Outreach required
  - Answer: offer compelling functionality and value

Purdue University Research Repository

- *Create:*
  - any Purdue faculty, staff, or graduate student researcher can create projects
  - describe the project
  - disclaim use of sensitive or restricted data
  - receive a default allocation of storage
  - register a grant award to increase allocation
  - invite collaborators to join project
- *Collaborate:*
  - git repository to share and version files (Google Drive integration)
  - wiki
  - blog
  - to-do list management and project notes
  - newsfeed
  - stage data publications

Michael Witt, Purdue University
DataCite Summer Meeting 2013
http://www.slideshare.net/datacite/2013-datacite-summer-meeting-california-digital-library-joan-starr-california-digital-library

Recommendation 2

- *Establish a curation unit*
  - Support campus researchers in curating their data
  - Office curation expertise
  - Recommend repositories, services
  - Document, promote best practices
  - Disseminate tutorial materials
  - Host events (lecture series, etc.)
  - Provide a “home” for campus data curators

Target audience: data curators

- *Support staff*
  - "Data managers," "data librarians," "information managers," "scientific programmers"; official or de facto
  - Work under or for PIs
  - Develop software, run software, handle and manage data
  - Domain specialists
    - Degrees, training in relevant disciplines

- *Shortcomings*
  - Little large-scale software engineering experience
    - Historically, level of expertise has been sufficient
  - Low use of standard software development tools, techniques
  - Education needed

- *Challenge*
  - How to improve existing practices?

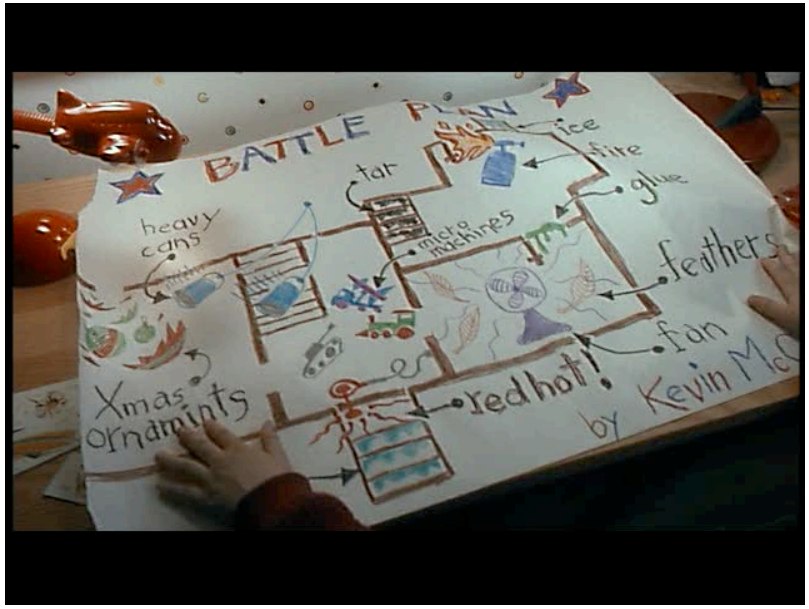
Implementation challenges

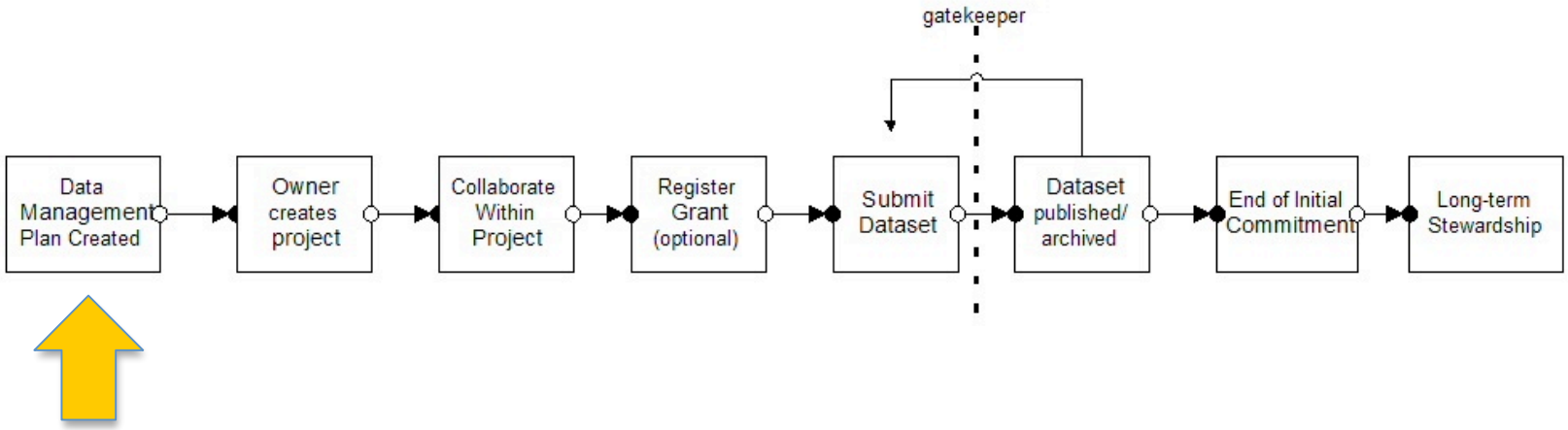
- *Success =*
  - If curation practices change for the better

- *Model: CSF mailing list*
  - Around ~30 years
  - Every campus sysadmin is on it
  - Ask questions, get answers
  - Buy/sell/trade equipment
  - Advertise training events
  - Annual beer bash
  - “Home” for sysadmins

Recommendation 3

- *Support creation, execution of data management plans*
  - DMPs are first, best, most well-defined point at which curation of a dataset is considered

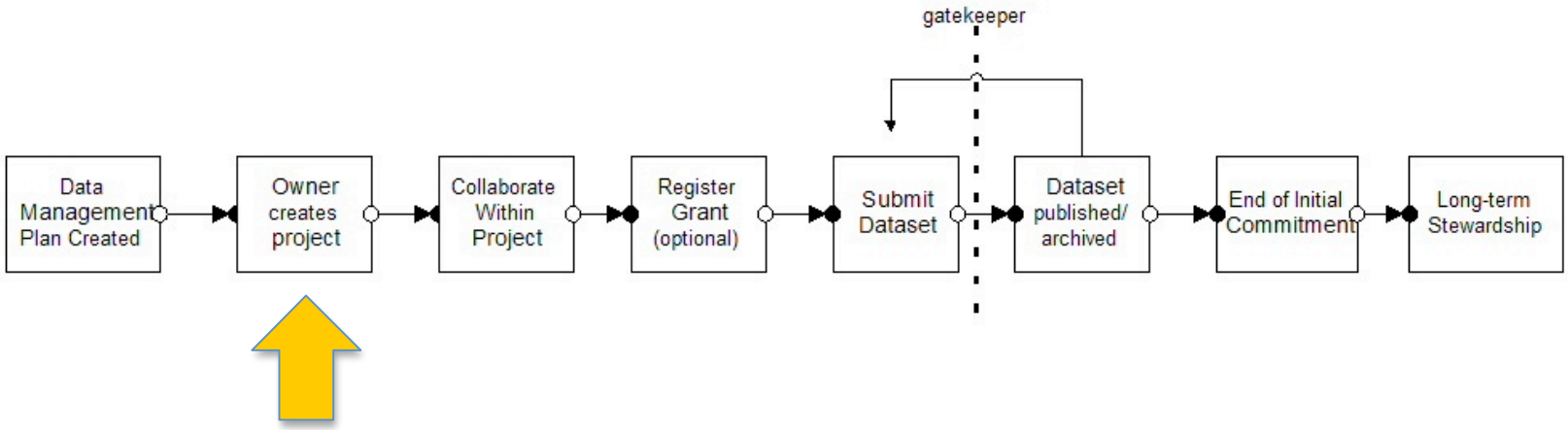




Librarians consult on data management plans in their subject areas.

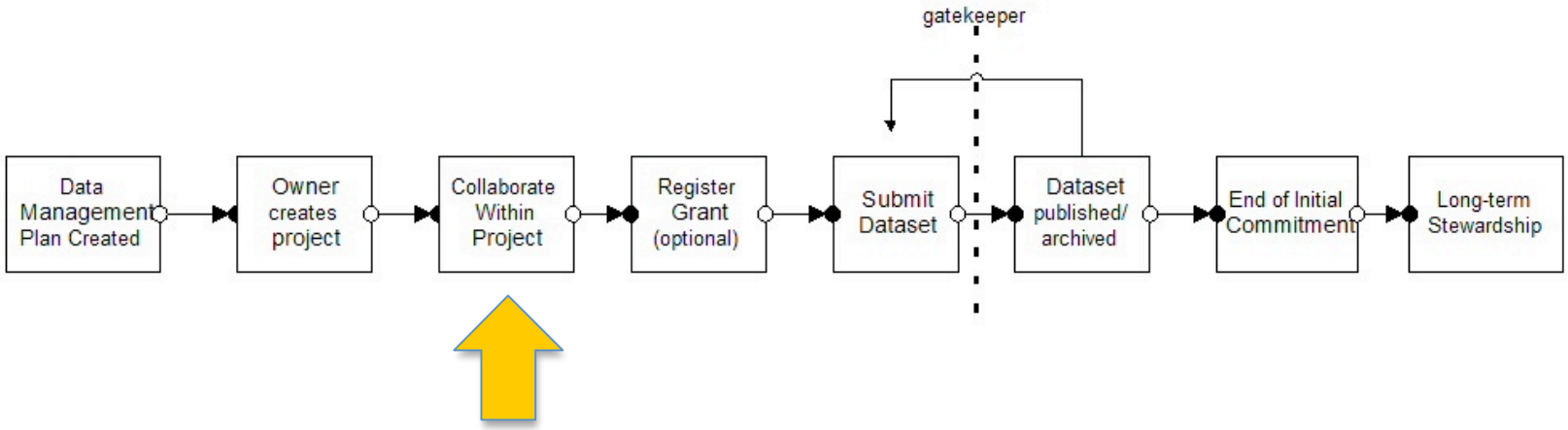
Michael Witt, Purdue University
DataCite Summer Meeting 2013
http://www.slideshare.net/datacite/2013-datacite-summer-meeting-california-digital-library-joan-starr-california-digital-library

Creating opportunities for librarians to interact with researchers about data…



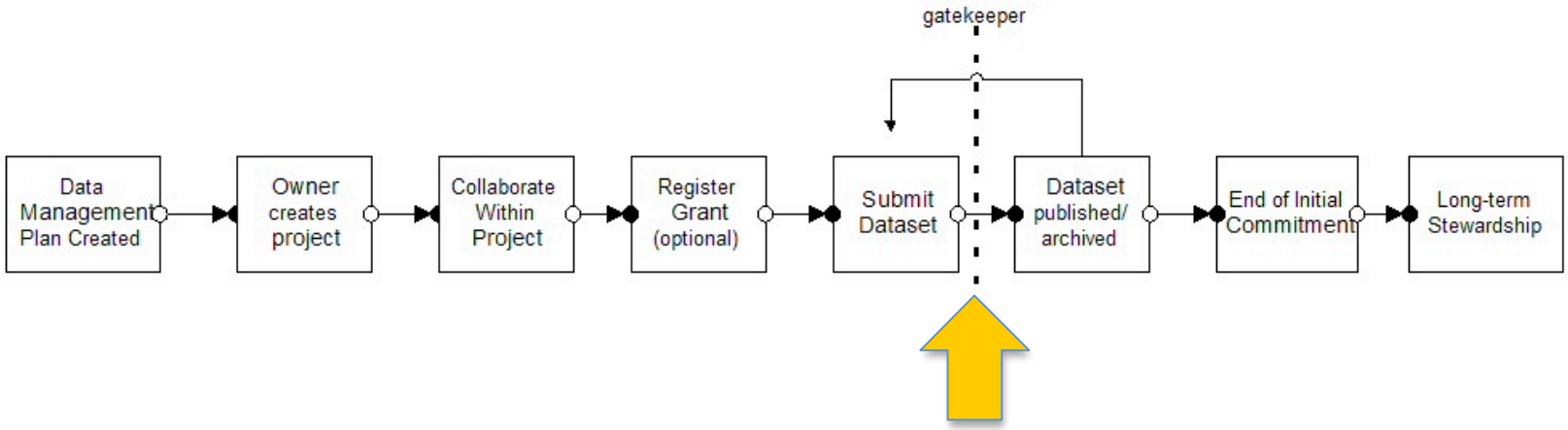
Librarian is notified by e-mail when a new project is created or a grant is awarded, based on department affiliation of Purdue project owner.

Creating opportunities for librarians to interact with researchers about data…



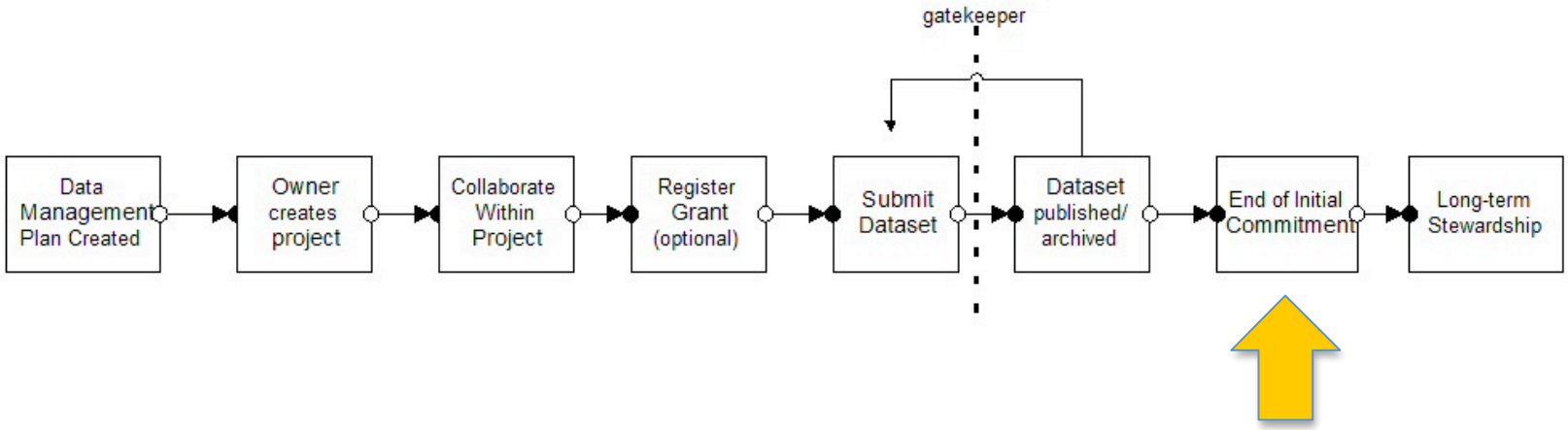
Librarian may consult or collaborate on project if needed.

Creating opportunities for librarians to interact with researchers about data…



Librarians review and post submitted datasets.

Creating opportunities for librarians to interact with researchers about data…



At the end of initial commitment (10 years), archived and published datasets are remanded to the Libraries' collection. A librarian working with the digital archivist selects (or not) the dataset for the collection.

Creating opportunities for librarians to interact with researchers about data…

Next steps

- *Continue in-depth studies*
  - Especially in the humanities
- *Scope, define proposed curation unit*
  - Staffing, leadership, resources, functions
- *Further flesh out curation service requirements*
- *Explore possibilities for DMP tie-ins*