Scholarly Communication Meeting  
Pittsburgh, January 14-15, 2013

Participants

Ron Larsen, U Pittsburgh  
Steve Griffin, U Pittsburgh  
Bill Arms, Cornell U  
Johan Bollen, Indiana U  
Fran Berman, Rensselaer Poly  
Bob Pego, Carnegie Mellon U  
Micah Altman, MIT Libraries  
Greg Crane, Tufts U

Spencer Keralis, U North Texas  
Josh Greenberg, Sloan Foundation  
Victoria Stodden, Columbia U  
Tom Moritz, consultant  
Ed Fox, VPI  
Chuck Henry, CLIR  
Carole Goble, U Manchester

Unable to attend due to travel/other circumstances:

Lewis Lancaster, U Cal, Berkeley  
Don Waters, Mellon Foundation

Carl Lagoze, U Michigan  
Sandy Payette, Cornell U  
John Unsworth, Brandeis U
Meeting Goals

The meeting goals were to identify new means and opportunities for enhancing scholarly communication across disciplines and to explore new models for documenting and disseminating a comprehensive record of computational and data-centered research.
New Research and Study Approaches Associated with Data and Computation

- **theoretical/analytical** (new theories are formulated and proven using a priori axioms and definitions)
- **empirical/observational** (inquiry based on detectable and measurable evidence; hypotheses driven and often aimed at theory building which in turn can yield new hypotheses and identify potential new theories)
- **computational** (typically, large-scale computation applied to mathematical models using high performance computers to produce simulations of physical phenomena often displayed in the form of scientific visualizations [e-Science])
- **data-driven** (analysis of very large numerical or textual data sets with the goal of elucidating patterns or discovering new correlations or relationships from which hypotheses might be constructed)
- **data-intensive** (inquiry involving the exploration and use of large, diverse data stores containing complex data objects with the goal of advancing interdisciplinary and domain scholarship)

- Recent developments that can also be considered as extensions or new forms of theoretical and empirical methods.

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
New computation and data-centered modes of inquiry and experimentation are often referred to as cyberscholarship or digital scholarship
Digital Scholarship

Exploring New Modes Of Inquiry In The Digital Era

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
Repository Development Trends

Very large repositories and global data infrastructures (2010) federation at data level via semantic web technologies, linked open data principles:
interlinked data over the web using URIs, RDF, links, vocabularies, relations; abstractions; graphs ...

Institutional and disciplinary repositories (2000) federated across repositories [DSpace; Fedora; ePrints]:
interoperability across repositories via OAI-PMH compound object packaging formats, etc..

Digital repositories (1990s): functional individual repositories; metadata catalogues and diverse information objects ...

Middleware services layer: software between the network and the applications providing authentication, identification, authorization, directories, security ...

ITC Infrastructure: processors, networks, storage, codes, compilers, tools, algorithms, software libraries ...

increasing capacities and capabilities
Progress is Being Made in Developing Very Large Scale, Highly Functional Global Data Infrastructures Containing a Rich Diversity of Information Objects

- Repositories and global data infrastructures (2010) federation at data level via semantic web technologies, linked open data principles

- Institutional repositories 2000 federated across repositories (DSpace; Fedora; ePrints)

- Digital repositories (1990s)

- Middleware = services layer

- ITC Infrastructure = Processors, memory, network

Merging to become universally accessible, carefully maintained knowledge infrastructure based on common principles and practices.

Clear evidence suggests that these can now support a full range of scholarly communication activities and develop into a persistent scholarly communication infrastructure.

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
Fishes of Texas

The Fishes of Texas Project is a multi-year research project that attempts to consolidate, standardize and fully georeference all of the known scientific information on the freshwater fishes of Texas. This package is the Linked Data version of the project's published data.

Resources

- Example (RDF/XML) (future permanent URI scheme) © 15 example/rdf+xml
- Example (RDF/XML) (temporary URI scheme) © 13 example/rdf+xml

Additional Information

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td><a href="http://data.fishesoftexas.org">http://data.fishesoftexas.org</a></td>
</tr>
<tr>
<td>Author</td>
<td>Juan F. Sequeda</td>
</tr>
<tr>
<td>Maintainer</td>
<td>Juan F. Sequeda</td>
</tr>
<tr>
<td>links:dbpedia</td>
<td>15241</td>
</tr>
<tr>
<td>links:taxonconcept</td>
<td>228</td>
</tr>
<tr>
<td>namespace</td>
<td><a href="http://data.fishesoftexas.org:8080/fishy/">http://data.fishesoftexas.org:8080/fishy/</a></td>
</tr>
<tr>
<td>triples</td>
<td>528887</td>
</tr>
</tbody>
</table>

Cite this

http://datahub.io/dataset/fishes-of-texas
Potential Application Areas Increase With Data Infrastructure Scope and Functionality

Repositories and global data infrastructures (2010) federation at data level via semantic web technologies, linked open data principles

Institutional repositories 2000 federated across repositories [DSpace; Fedora; ePrints]

Digital repositories (1990s) [Digital Libraries and other repositories, Scientific DBs]

Middleware services layer

ITC Infrastructure

Processors, memory, network

Applications Space

humanities

social sciences

natural sciences (experimental methods)

formal sciences

natural sciences (computational simulations)

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
Research and Scholarship Funding is Not Seen as Equitably Distributed

Repositories and global data infrastructures (2010) federation at data level via semantic web technologies, linked open data principles

Institutional repositories 2000 federated across repositories [DSpace, Fedora, ePrints]

Digital repositories (1990s)

Middleware services layer

ITC Infrastructure

Levels of Activity and Investment

humanities

social sciences

big data

eScience

natural sciences (computational simulations)

natural sciences (experimental methods)

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
Meeting Presentations and Group Discussion Foci

• new methodologies, reach and affordances of digital scholarship

• technologies and activities to capture of a more complete record of stages in the scholarly research workflow

• effective frameworks (existing and proposed) to accelerate the repurposing and reuse of open data resulting from scholarly work and research

• robust document models for presenting and "bundling" the processes, resources, outputs and potential impacts of scholarly work

• new means for dissemination to increase the diffusion and reach of new concepts and findings

• accurate measures to ensure appropriate and fair attribution, acknowledgement, credit and reward for those involved in carrying out the work
A Persistent and Recurring Theme
- The Burden of Evidence -

Defining features of science include **repeatability** and **reproducibility**. **Repeatability** refers to the ability to duplicate an experiment under the same conditions many times and obtain the same result. **Reproducibility** refers to the ability for others to replicate the work in different environments and obtain the same results, setting the stage for extending the work in new directions. These requirements hold for theoretical and empirical research and apply to the formal, natural and social sciences. Replication of results using proven, rigorous methodologies confirms the veracity of a research process and outcome.

Carole Goble
Victoria Stodden
Tom Moritz
Some necessary conditions for reproducibility ...

- access to a comprehensive record of the research process and scholarly workflow including:
  - process records: algorithms, software pipelines and versioning, datasets and transformations, storage formats and protocols, event tracing, ...
  - resource descriptions: journals, logs, tools, methods, dialog, collaborative activities and external contributions, ...
  - intermediate forms: temporary models, concept changes, recursion points, software versions, external dialogs and contributions ...
  - workflow artifacts: transcriptions, translations, annotations, steps taken to acknowledge distribution of effort, attribution and credit, ...
Working Group Assignments

Charge to Break Out Group #1
What can be done to effectively capture, document, and prepare the information flow associated with each stage of a research project or scholarly work so that they can become part of a larger, global knowledge and scholarly communications infrastructure.

Charge to Break Out Group #2
Digital scholarship often involves new types of information objects, data analytic processes, resources, tools and heuristic representation of findings that cannot be accurately or completely described or communicated in traditional print or in print + electronic venues. Semantic access is also inherently limited. What new expressive forms, document models, practices and venues might help remedy this situation.
Charge to Break Out Group #1

What can be done to capture, document, and prepare elements of research and scholarly workflows so that essential components can become part of the overall reporting of the effort and eventually be integrated into global knowledge and scholarly communications infrastructure.

Points to consider:

The primary researchers will be disinclined to do this for multiple reasons. Is there a possibility of automating this? If so, would this be most tractable during the individual stages of the research workflow or after the project is complete. What might this entail?

Should there be “meta-research” activities as part of major research projects? The purpose would be to document the research process and prepare resources and artifacts to enable reproducibility.

Is this already being done to a certain degree in some disciplinary areas? Why? What has been the benefit to the larger scholarly communities involved?
A Very General Scholarly Research Workflow Model
- example of simple and traditional form -

mix of dialog, data and resources from individuals, the web, libraries, archives, etc.

primarily informal processes

Information flows into and out of the project at each stage

primarily formal processes

- inspiration, explore, discover area of interest
- formulate problem, design research, collect data
- conduct research, analyze results
- prepare findings, disseminate results

data
discovered, referenced, accessed, gathered, transformed, analyzed, presented

high low

\[ \Delta t \]

Libraries, Academic Departments, Individuals, ...

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
Current Scholarly Research Workflow and Communication Model

- **recently emerging global data and resources infrastructures**
  - data and research cyberinfrastructure: digital libraries, scientific databases, reports, publications, ETDs, software & code libraries, executable documents, 1st and 2nd generation repositories (linked open data; semantic web technologies ...), processing, storage and grid services
  - conversant/discursive web: social media, blogs, chat rooms, project sites, commentaries, ...

- **Information flows into and out of project**
  - **activity:**
    - inspiration, explore, discover area of interest
    - formulate problem, design research, collect data
    - conduct research, analyze results
    - prepare findings, disseminate results
  - **data:**
    - discovered, studied
    - accessed, collected
    - transformed, analyzed
    - prepared, presented

- **hosting institutions**
  - (libraries, archives, other content and service providers)

- **subscription & open access journals, self-published documents & pre-prints, hybrid dissemination models**

- **loosely organized activities to collect and prepare artifacts for future repurposing and reuse by others [event tracing, versioning, logs, journals, data documentation, intermediate forms, temporary models, concept changes, recursion points, transcription, translation, annotation, ...]**

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
A New General Model for Scholarly Communication Infrastructure Based on Scholarly Workflow

- The scholarly communications layer: dynamic research reports with detailed descriptive information of the methods and concepts as well as access to software, data, and other experimental assets, provenance and citation linkages, etc. meeting community-adopted practices for presentation, access, preservation and archiving.

- The global data and research cyberinfrastructure: research data infrastructures, digital libraries, scientific databases, reports, publications, ETDs, software & code libraries, executable documents, 1st and 2nd generation repositories (linked open data; semantic web technologies ...), processing, storage, cloud and grid services.

- The conversant/discursive web: social media, blogs, chat rooms, project sites, commentaries, ...

---

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
New Roles for Libraries, Archives and Service Providers

A New General Model for Scholarly Communication Infrastructure Based on Scholarly Workflow

In this model the role of Libraries evolves from one of holders and providers of knowledge resources to one of being an active partner in the research process. Libraries and librarians provide tools and expertise that expedite research and scholarship. Libraries have the institutional structure and many of the resources needed to advance and sustain scholarly workflows.

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
Summary Recommendations from Breakout Group #1

• capture a comprehensive record of research process and scholarly production to support verification and reproducibility of results

• create full research process record: logs, applications, methods, datasets, dialog, collaborative activities

• prepare workflow artifacts for repurposing and reuse

• develop a protocol model for scholarly output that allows for modularity, distributes effort and credit, and facilitates democratic access

• develop methods for managing release of components of scholarly output from all stages of the scholarly workflow

meta-research process necessary to accomplish this
Charge to Break Out Group #2

*Digital scholarship often involves new types of information objects, data analytic processes, resources, tools and heuristic representation of findings that cannot be accurately or completely described or communicated in traditional print or print + electronic venues. Semantic access is also inherently limited. What new expressive forms, document models, practices and venues might help remedy this situation.*

**Points to consider:**

The primary researchers may be disinclined to do this for multiple reasons. For example: extra work in preparation of their project results; institutional pushback from “publishing” in non-conformist or new experimental venues; complexities introduced into review processes.
Challenges for Document Models: Describing Complex Projects

Figure 3:
Information flow map for epidemiology of zoonotic diseases (case study 3)

Research Information Network and British library “Patterns of information use and exchange: case studies of researchers in the life sciences” [http://www.rin.ac.uk/system/files/attachments/Pattems_information_use-REPORT_Nov09.pdf](http://www.rin.ac.uk/system/files/attachments/Pattems_information_use-REPORT_Nov09.pdf)

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
A Richer Document Model: One Example
New Tools Can Help

**RightField: Spreadsheet Annotation by Stealth**

RightField is an open source desktop tool for embedding ontology term selection into Excel spreadsheets to create templates. Vanilla Excel spreadsheets are produced but do not contain macros or require special plugins.

- RightField enables ranges of cells to be marked up so that they can only be selected from specific ontology terms.
- It works with OWL and OBO ontologies and RDF Schema vocabularies.
- Provenance information such as complete term IRIs, the source ontologies they were taken from and ontology version information is embedded within the Excel spreadsheet itself for later use. However, all embedded ontology information is hidden from the spreadsheet user.

Excel workbook with marked up cells:
Marked up cells are shown with green overlays in Excel; values for these cells must come from the specified list of ontology terms.

Term list for selected cells:
The values of the selected marked up cells must come from the list of ontology terms. In this case, those terms represent subclasses of the ActivityType from the ERM ontology.

When a spreadsheet template with embedded terms and marked up cells is opened in Excel, drop down lists allow users to choose values to enter for that cell.

Once filled out a spreadsheet can be saved and processed so that an RDF Graph, which relates back to the original ontology can be produced.

Biolists, who are intimately familiar with Excel, produce semantic annotations as part of their daily workflow, without any need for specialised ontology editing tools or knowledge of ontologies.

Contact us and find out more at: [http://www.rightfield.org.uk](http://www.rightfield.org.uk)
Dynamic Project Sites are Complex Scholarly Documents

Religious Atlas of China and Himalaya

More: about the China Religious Atlas

Home | About the Site | Collaborators | Contact Us

© 2006 Electronic cultural Atlas Initiative

University of California, Berkeley

www.ecai.org

Visualizing Cultures — Image-Driven Scholarship

Visualizing Cultures was launched at MIT in 2002 to explore the potential of the Web for developing innovative, image-driven scholarship and teaching. The VC mission is to use new technology and infinite accessible visual materials to reconstruct the past as people of the time visualized the world (or imagined it to be).

Topics under active focus on Japan in the modern world and early modern China. The thrust of these explorations extends beyond Asia per se. However, to address "cultural" in much broader ways — culture of modernization, war and peace, consumerism, images of "Self" and "Others", and so on.

Images of every sort are not only documents and examined here — in partnership with contributing institutions and collections, and with the collaboration of experts devoted to preserving the printed word and rare books.

The Visualizing Cultural Curriculum offers a full complement of standards-compliant lessons, providing a pathway by teachers and students become active historians and knowledgeable readers of images.


S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
Summary Recommendations from Breakout Group #2

Need a new “bundled” modular research document model in which elements are linked semantically, released when ready and capable of being recombined at any time and in different environments

• Provides a variety of presentation forms to accommodate disciplinary domain requiring different expressive forms

• Facilitates its own automated retrieval

• Gives direct access to datasets, tools and other workflow elements

• Anticipates future needs for storage, access and use (curation, stewardship, provenance issues)

• Capable of aggregation at the component level with other research documents

• Annotation and relationship friendly; indefinite versioning

Greg Crane
Carole Goble
Benefits from Richer Document Models
One Example: Cultural Historical Research

“Cultural historical research means understanding 'possible pasts', the facts, events, material, social and psychological influences and motivations. It lives from understanding contexts, by pulling together bits and pieces of related facts from disparate resources, which can typically not be classified under subjects in an obvious way. It lives from taking into account all known facts.

... Under these conditions, the global network of knowledge can reveal deep “stories” built out of an immense number of concatenated primary facts, and a thing impossible for a traditional library.”

Martin Doerr - Principal Researcher, Forth - Hellas

S. Griffin/U Pittsburgh/CNI Spring 2013 Meeting
Next Meeting TBA: Focus on the Humanities

Thank You

Meeting Web Site is Under Construction: Comments and suggestions are welcome! Contact Steve Griffin at sgriffin@pitt.edu