Bringing Research Data into the Library: MIT's Experience

Mackenzie Smith
Massachusetts Institute of Technology

Follow this and additional works at: https://escholarship.umassmed.edu/escience_symposium

Part of the Library and Information Science Commons

This work is licensed under a Creative Commons Attribution 4.0 License.

Repository Citation

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License.
This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in University of Massachusetts and New England Area Librarian e-Science Symposium by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Bringing Research Data into the Library: MIT’s Experience
Case studies for

- Biological Oceanography
- Climate Change
- Chemistry and Chemical Engineering
- Materials Science and Engineering
- Neuroimaging
- High Energy Physics

http://hmi.ucsd.edu/howmuchinfo.php
Neuroimaging Case Study
**Neuroimaging Case Study**

- Sources: Brain & Cognitive Science Department; McGovern Institute for Brain Research; Martinos Imaging Center; Research Lab of Electronics

- Need to *combine* digital images (MRIs, DTIs, VBM, etc.) *with* protocol data, phenotype and genomic data, EEGs, etc. for the same subject

- LARGE-SCALE
  >10Tb per year for one group of 4 faculty

- EXPENSIVE
  each subject ~$1000 (1500/year, per machine)

- HARD TO LOCATE, INTERPRET
  no standard way to annotate images for sharing, reuse
Temperature versus salinity (T-S) relations for the North Pacific Subtropical Gyre at station ALOHA
Sources: Earth, Atmospheric and Planetary Sciences; Civil Engineering; Biological Engineering

Need to *combine* metagenomic data *with* biochemical sensor data, including water chemistry, optical properties, physical data (e.g. sample location)

LARGE-SCALE
Solexa sequencer produces 1Tb per run X 2-3 runs/week

IRREPLACEABLE
time dependent, not fully analyzable today

HARD TO LOCATE
no integrated DB exists (e.g. GenBank only takes sequence data)
MIT Libraries and Data
Building on *already established* data curation for

**statistical** (Harvard-MIT Data Center)
**geospatial** (Geodata Repository)
**bioinformatics** (via NCBI)
**digital library collections** (e.g. images, videos)
**general datasets** (DSpace digital archive)
Welcome to Social Science Data Services

Social Science Data Services provides assistance with finding, understanding, and managing statistics or numeric or tabular data in the social sciences, management, and related areas. Browse the left-hand navigation bar for information about accessing data, training, software and hardware, and archiving and disseminating your data. See also How Social Science Data Services relates to the Geographic Information Systems (GIS).

If you're searching for data, see our resources by subject and consider searching in one of several data centers such as the Harvard-MIT Data Center and ICPSR.

News

- IAP Workshop: Finding Research Datasets
- ICPSR Summer Internship Program
- ICPSR Undergraduate Research Paper Competition
- New Office Hours for Statistical Consultant: Mondays 9am-12pm; NOTE: No office hours held in December

This page was last updated on Thursday, 11-Dec-2008 18:04:00 EST
GIS

Geographic Information Systems (GIS) enable users to visualize and analyze spatial information in a dynamic, digital environment. It provides tools for integrating, querying and analyzing a wide variety of data types, such as scientific and cultural data, satellite imagery and aerial photography, as well as data collected by individuals, into projects, with geographic locations providing the integral link between all the data.

Mission

The GIS Laboratory is a collaboration between the MIT Libraries, Information Services & Technology (IS&T), and the Office of Educational Innovation and Technology (OEIT). Our goal is to support the MIT community’s research and academic activities, as they relate to Geographic Information Systems (GIS). This support not only includes the collection of literature, data, software, and hardware; but also, training and assistance in the use of GIS. We are available to work with faculty to implement and assist in teaching GIS-based labs in individual subjects.

MIT does not offer a GIS certificate or degree program.

Hot Topics:

- New Workshop: Introduction to GIS for Architecture
- How do I get started working with GIS at MIT?
- How do I obtain ESRI GIS software for my personal or lab computer?
- What GIS data is available at MIT?
- ESRI Virtual Campus Classes
- Data updates:
  - Manhattan, NY 3D buildings from 2009 in MIT Geodata Repository - (MIT web certificates required for viewing and accessing)
  - New administrative boundaries for countries all around the world
  - New data on the web by state webpage update: NYC Data Mine
  - Energy data from Flatts in the MIT Geodata Repository - (MIT web certificates required for viewing and accessing)
  - City of Boston Portal - (MIT web certificates required for viewing and accessing)
  - NYC City Tax Lots - (MIT web certificates required for viewing and accessing)
  - New Jersey Tax Lots - (MIT web certificates required for viewing and accessing)
  - New York City Taxi Lots - (MIT web certificates required for viewing and accessing)
  - South Africa: Cape Town | Eastern Cape | Western Cape
  - India GIS datasets
  - China GIS datasets
  - data on the web by state webpage update: NYC Data Mine
- Find out what’s new in the MIT libraries through RSS feeds

- Services updates:
  - Download the latest version of the MIT Geodata Repository Search Tool - released October 2009
  - Learn more about it with the new MIT Geodata Repository Search Tool for ArcGIS Help Guide
  - ESRI ArcGIS 9.3.1 available to the MIT community
  - MIT Geoplot - GIS data access with a web browser
  - GIS units available for checkout from Rotch Library

Hours

The GIS laboratory, housed in Rotch Library, is available for use during Rotch operating hours.

GIS Lab Assistance: September 14 - December 10

A GIS specialist will be available in the GIS lab for walk-in assistance from 12:30-4:00 pm, Monday-Thursdays.

If you have a GIS question and are not available during walk-in lab hours please email gishelp@mit.edu to check if a GIS specialist can meet with you at a different time A brief description of the issue you are having. Please indicate a time that you would like the meeting to occur.

GIS Home
- Hours
- Spatial Data
- GeoWeb
- Hardware + Software
- Teaching + Learning
- Examples

Maps + Mapping
- Social Science Data Services
- US Census
- GIS Lab
- Rotch Library, building 3-238
- Contact Us
gishelp@mit.edu

Search GIS Services
Library Roles in Data Curation

- Acquire and catalog faculty-authored and externally-published datasets
- Consultation and training services
- Liaise with national archives (e.g. ICPSR)
- Develop standards (e.g. DDI, FGDC)
- Maintain website, data repositories
Data Management and Publishing

Manage Your Data

Managing your data before you begin your research and throughout its life cycle is essential to ensure its current usability and long-run preservation and access. Data management activities include:

A Data Planning Checklist | Create a Data Management Plan | Guides to Data Management

A Data Planning Checklist:

1. What type of data will be produced? Will it be reproducible? What would happen if it got lost or became unusable later?
2. How much data will it be, and at what growth rate? How often will it change?
3. Who will use it now, and later?
4. Who controls it (PI, student, lab, MIT, funder)?
5. How long should it be retained? e.g. 3-5 years, 10-20 years, permanently
6. Are there tools or software needed to create/process/visualize the data?
7. Any special privacy or security requirements? e.g. personal data, high-security data
8. Any sharing requirements? e.g. funder data sharing policy
9. Any other funder requirements? e.g. data management plan in proposal
10. Is there good project and data documentation?
11. What directory and file naming convention will be used?
12. What project and data identifiers will be assigned?
13. What file formats? Are they long-lived?
14. Storage and backup strategy?
15. When will I publish it and where?
16. Is there an ontology or other community standard for data sharing/integration?

Create a Data Management Plan

Planning for your data management needs ahead of time will save you time and resources in the long run and ensure that your data will be usable in the future. A formal plan can be valuable to you and may be required by your funding agency. Topics to cover in a written plan include:

- name of the person responsible for data management within your research project
- description of data to be collected and the methodology
- how data will be documented throughout the research project
- data quality issues
- backup procedures
- how data will be made available for public use and potential secondary uses
- preservation plans
- any exceptional arrangements that might be needed to protect participant confidentiality or intellectual property

For tips on creating a data sharing or management plan, see the:

- Australian National University Data Management Manual
- NIH examples of data sharing plans
- UCL-OSS Data Management Plan

Guides to Data Management
General Data Management

• Website averages 650 hits/month

• Team of Public Services librarians
  • Led by liaisons for Civil and Environmental Engineering, Computer Science, GIS, Social Science
  • Broader group includes Bioinformatics, Chemistry, Humanities, others

• Consultation from Metadata Services specialists, IT specialists
Data Archiving in DSpace@MIT

- Excel spreadsheets [microarray data, survey data]
- Zip files [gene sequences in .gb, code in .py]
- XML files [chemical molecules in CML]
- Media files [images, audio/video in .wav]
- Software-specific [Sequest DTA format for spectral data]
- Java Web Start files
- Lab e-notebooks
- Thesis supplementary data

April, 2010 E-Science Symposium ©MIT
Case Study: RADISH
Robotics Datasets in DSpace@MIT

http://dspace-demo.mit.edu/handle/1234567890/41939
What we did:

- Defined local taxonomy for metadata values
- Changed item record display
- Changed submission form, simplified workflow
- Loaded data from previous repository and added CC.o licenses
- Non-MIT submissions reviewed by community owner or library staff
Radish: The Robotics Data Set Repository

The Robotics Data Set Repository (Radish for short) provides a collection of standard robotics data sets. Contained here-in you will find:

- Logs of odometry, laser and sonar data taken from real robots.
- Logs of all sorts of sensor data taken from simulated robots.
- Environment maps generated by robots.
- Environment maps generated by hand (i.e., re-touched floor-plans).

By making these data sets available to the community, we aim to facilitate the development, evaluation and comparison of robotics algorithms. While the current focus is clearly on localization and mapping, we expect that Radish will ultimately expand to reflect the interests of the broader robotics community.

Radish is a community effort. Researchers are invited to download and make use of the data sets contained here-in, and, in return, to make their own contributions to the repository.

Good data sets are too precious to keep to ourselves!

Submitting to Radish and FAQ:
To register as a new data contributor, email radish@mit.edu.
FAQ: [link]
Software:
Player/Stage: [link]
Carmen: [link]
OpenSLAM: [link]

Please note: When you put your data in Radish, you are putting it in the public domain. Please contact radish@mit.edu if you have questions about this.

Recent Submissions

- [uliberta.csc-flir3-vision](#) by Krippenstein, Jonathan (2009-04-07)
- [comparison of self-localization methods, continued](#) by Gutmann, Steffen (2009-04-07)
- [albert-b-laser-vision](#) by Stachniss, Cyrill (2009-04-07)
- [ir-fctuc_irf2_cmu_mun_carmen_dataset](#) by Davim, Luiz; Dias, Jorge Manuel Miranda; Ferreira, Filipe; Prado, Jose (2009-04-07)
- [ir101-explored](#) by Stachniss, Cyrill (2009-03-11)
Data Archiving in DSpace@MIT

Abdul Latif Jameel Poverty Action Lab (J-PAL) J-PAL Datasets

- stata files (.dta)
- associated codebooks (.doc)

E-Science Symposium
©MIT
Research Laboratory for Electronics (RLE)

Speech Communication Group: MIT American English Map Task

- sound files (.wav) and scanned map files (.pdf)
- README file
Synthetic Biology: software accompanying Master's Thesis
"Analysis of Targeted and Combinatorial Approaches to Phage T7 Genome Generation"

- Software files (.py, .h, .lib, etc.)
- Input data files (.txt)
- README (.doc)
Data Issues to Think About
Challenges of Access and Preservation for Research Data

- Persistent identifiers
- Storage space
- Metadata and documentation
- Long-term preservation
- Data sharing
Identifiers for Datasets

Citable data identifiers enable linking data to publications

- e.g. DSpace uses Handles
- New DOI registration agency
  - German National Library of Science and Technology
  - For research datasets from technology/science and medicine
  - Also, the British Library, the Library of the ETH Zurich, the French Institute for Scientific and Technical Information (INIST), the Technical Information Center of Denmark and the Dutch TU Delft Library
Dataset Storage

- Local (library, campus, consortia) or
- Cloud (Amazon S3, Azure, Atmos, etc.)

- Also for backup and replication (for digital preservation)

- Ideally policy-driven (e.g. # copies, locations, access guarantees)
## Metadata and Documentation

**DSpace Metadata Schema for Edinburgh DataShare**

Robin Rice, Stuart Macdonald & George Hamilton  
Ver.1 (3. 7. 2008)

<table>
<thead>
<tr>
<th>DSpace Form</th>
<th>DSpace ID</th>
<th>Dublin Core Element</th>
<th>DC Qualifier</th>
<th>Field Label</th>
<th>Input Type</th>
<th>Mandatory?</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>1</td>
<td>contributor</td>
<td></td>
<td>Depositor</td>
<td>name</td>
<td>false</td>
<td>Hint: Enter the name of the person entering this record.</td>
</tr>
</tbody>
</table>
| 1.2         | 9         | creator             |              | Data Creator| onebox     | true       | Required: You must enter a Data Creator for this dataset.  
             |           |                     |             |             |            | Repeatable: true |
| 1.3         | 64        | title               |              | Title       | onebox     | true       | Hint: Enter the main title of the dataset.  
             |           |                     |             |             |            | Required: You must enter a main title for this dataset.  
             |           |                     |             |             |            | Repeatable: false |
| 1.4         | 65        | title               |              | Alternative Title | onebox     | false      | Hint: If the dataset has any alternative titles, please enter them here.  
             |           |                     |             |             |            | Repeatable: true |
| 1.5         | 27        | description         |              | Dataset Description (abstract) | textArea | false | Hint: Please provide a summary description of the study in which the data was generated.  
             |           |                     |             |             |            | Repeatable: false |
| 1.6         | 66        | type                |              | Type        | list       | true       | Hint: Select the type(s) of content of the item. To select more than one value in the list, you may have to hold down the "CTRL" or "Shift" key.  
             |           |                     |             |             |            | Repeatable: true |

Preserving Data

- Similar to other digital archives

Neutral, standard formats are best; keep lots of copies, lots of documentation

- Special concern for software
  - e.g. SPSS, netCDF, CATIA
  - Archiving software is difficult
    (need source code, compilers, sometimes hardware)
Data Sharing

**IPR and data licenses**

- Most data NOT copyrightable in the U.S.  
  *facts cannot be copyrighted*

- Licenses (e.g. CC licenses) usually DO NOT APPLY and are not enforceable

- Should be placed in the *public domain* or not shared at all
No Copyright

The person who associated a work with this document has dedicated the work to the Commons by waiving all of his or her rights to the work worldwide under copyright law and all related or neighboring legal rights he or she had in the work, to the extent allowable by law.

Works under CC0 do not require attribution. When citing the work, you should not imply endorsement by the author.

Other Rights — in no way are any of the following rights affected by CC0:

- Patent or trademark rights held by the person who associated this document with a work.
- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

This is a human-readable summary of the Legal Code (read the full text).

Disclaimer

Use this tool for your own work.
The Library’s Role

Summing up

Data organization and annotation
  e.g. ontologies and metadata

Data curation
  e.g. long-term storage, preservation, publishing

*Outreach and support to local researchers*
<table>
<thead>
<tr>
<th>Distributed high performance computing; analysis tools applied to data over secure international network; NJM interfaces</th>
<th>Data Grid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer review of datasets; seamless link to publications; role-based layers of access; data overlay journals</td>
<td>Data publishing</td>
</tr>
<tr>
<td>Graphs, charts, maps configurable online</td>
<td>Data visualisation</td>
</tr>
<tr>
<td>“Actionable” marked up dataset installed in a data browser tool subsetting capability</td>
<td>Data manipulation online</td>
</tr>
<tr>
<td>Original format plus XML markup of data or XML database; open standards used appropriate to domain; metadata or setup files may be bundled with dataset for importing elsewhere</td>
<td>Data enhanced for re-use</td>
</tr>
<tr>
<td>Quality assured metadata; guidance available for depositors; suitably anonymized/consent for sharing obtained from subjects; thorough documentation about data creation and methodology included; permanent IDs; formats validated and suitable for distribution; migration-based preservation commitment</td>
<td>Network of distributed repositories/subject and/or institutionally based</td>
</tr>
<tr>
<td>Data files with minimal documentation (e.g. readme file describing each data file) downloadable from Internet</td>
<td>Zip and ship</td>
</tr>
<tr>
<td>Metadata record of dataset on website or in repository; possibly with embargo and contact information to request access</td>
<td>Open access</td>
</tr>
<tr>
<td>Networked drive, available to research group, version control</td>
<td>Search and discovery enabled; restricted access</td>
</tr>
<tr>
<td>Password protected, networked drive (backup procedures)</td>
<td>Email dissemination by request</td>
</tr>
<tr>
<td>Personal hard drives un-networked</td>
<td>Privileged access</td>
</tr>
<tr>
<td>Simple data storage</td>
<td>Typical status quo</td>
</tr>
</tbody>
</table>

Robin Rice, September 2007