DMP: Archiving data products

Bob Cook
Environmental Sciences Division
Oak Ridge National Laboratory
Email: cookrb@ornl.gov
Data Management Plan

4. Plan for archiving data and providing long-term access to data

When preparing a data management plan:
“Begin with the end in mind”

- Identify Data Center
  - *May be institutional resource (library) or discipline data center*
- Collaborative effort between data center and project
  - *Add letter of collaboration to proposal*
Selecting a Data Center

1. Scientific discipline
   • Keep discipline data together

2. Functionality
   • Discovery, access, and visualization
   • Specialized data types (geospatial data, genetic sequences, etc.)

3. Requirements
   • Data center’s and project’s requirements
Data Center: Stewardship and Archive Functions

- **Acquisition**
  - identify how best to serve the scientific community
  - establish how and when to receive data

- **Ingest**
  - perform QA checks
  - compile project-provided metadata
  - convert to archivable file formats

- **Enhance** (as requested)
  - convert to standard formats & units
  - aggregate files

- **Metadata / Documentation**
  - Prepare final metadata record and documentation

- **Archive / Publish**
  - generate citation

- **Exploration and Distribution**
  - provide tools to explore, access, and extract data for users worldwide

- **Post-Project Data Support**
  - serve as a buffer between end users and PIs
  - provide usage statistics

- **Stewardship**
  - provide long-term secure archiving of the data
  - security, disaster recovery
  - migration to new computer systems
Ecological Archive Data Papers:
• compilations and syntheses of important data sets
• peer-reviewed
• abstracts published in ESA journals
• formal citation
• published in digital, Internet-accessible form

---

Appendix A. Mean biomass (g dry mass m$^{-2}$; SE in parentheses) of major foods (as defined by Randolph et al., 1991) during four seasons in three patch types occupied by cotton rats. Values are for foods only, not total plant biomass. Abbreviations are: MO = monocot, MI = mixed, DI = dicot patches types; lgs = lower green stems, ugs = upper green stems, uds = upper dry stems, lds = lower dry stems, gl = green leaves, and f = fruits.

<table>
<thead>
<tr>
<th>Plant Species</th>
<th>Plant Part</th>
<th>Spring</th>
<th>Summer</th>
<th>Autumn</th>
<th>Winter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MO</td>
<td>MI</td>
<td>DI</td>
<td>MO</td>
</tr>
<tr>
<td>Schizachyrium scoparium</td>
<td>lgs</td>
<td>302.7</td>
<td>62.8</td>
<td>33.9</td>
<td>142.2</td>
</tr>
<tr>
<td></td>
<td>ugs</td>
<td>(21.7)</td>
<td>(12.2)</td>
<td>(9.2)</td>
<td>(20.4)</td>
</tr>
<tr>
<td></td>
<td>uds</td>
<td>218.4</td>
<td>45.3</td>
<td>24.5</td>
<td>146.8</td>
</tr>
<tr>
<td></td>
<td>lds</td>
<td>(15.6)</td>
<td>(8.8)</td>
<td>(16.6)</td>
<td>(21.1)</td>
</tr>
</tbody>
</table>
Dryad (~1,000 data products)

Coordinated submission of articles and underlying data

Handshaking with specialized repositories

Promotion of reuse and incentives for deposit

ESA 2011: Data Management Plans
Knowledge Network for Biocomplexity
(> 24,000 data packages)

Contributors
- Individual investigators
- Field stations and networks
- Government agencies
- Non-profit partnerships
- Synthesis centers

Data Types
- Ecological
- Environmental
- Demographic
- Social/Legal/Economic

Data Sizes

<table>
<thead>
<tr>
<th>Size</th>
<th>Data Sizes</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>^1</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>1-10</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>10-200</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>&gt;200</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

ESD 2011: Data Management Plans
ORNL DAAC
• Part of NASA’s Earth Observing System Data & Information System
• Terrestrial Ecology and Biogeochemical Dynamics Data

Data Holdings (>900 products)
• NASA Field Campaigns
• Land Validation (remote sensing)
• Global and Regional Spatial Data
• Terrestrial Biogeochem. Model Code

Tools for Discovery, Access, Extraction, and Visualization
Additional Discipline Data Centers

Compilation of Earth Observation data centers will be available at DataONE

ESA 2011: Data Management Plans
Additional Resources

Ecological Archives Data Papers
   http://www.esapubs.org/archive/

Dryad
   http://datadryad.org/

Knowledge Network for Biodiversity
   http://knb.ecoinformatics.org/

ORNL DAAC
   http://daac.ornl.gov