Agenda

• Rationale and Motivations for Data Management Plans
• Data and data structures
• Metadata and provenance
• Provisions for privacy, confidentiality and licensing
• Data sharing/management Storage and long-term preservation
Background and motivations

• Why DMPs?
  – Data approach to research and education
  – Data collections as the library of the future
  – Funding agencies and publishing requirements, accountability and new discoveries

• Does my proposal need a DMP?
  – Will not generate data
  – Will generate software
  – It is a simulation project
  – Will reuse published data
    • Still need a DMP

Some journals have a policy that require authors to make data available to readers via public repositories as a condition of publication. DOIs for data
Considerations

- Consider the elements that will allow reproducing your research
- Integrate the data processes to the research stages
- Domain science driven
- Throughout the research process, data and technology change
- Data may have a longer lifespan than the research process
- Data management plans and their evaluation will evolve as well

Consider your data management plan as a narrative of your research process in which data is the unit of analysis and you need to document each step so that the research can be reproduced.
Resources for writing DMP

• Major University Libraries have a DMP page to provide guidance
• Funding Agencies have different requirements
• DMP templates to guide you through the process
• Particularly useful:
  – MIT Data Management Guide
  – DCC Curation Reference Manual

The point is not the template but the plan
DATA and DATA STRUCTURES
Reproducibility value and data retention

- Experimental data
  - From labs and equipment (S – C)
- Observational data (E)
  - Captured in real time
- Derived data (S – C)
  - After data mining and statistical processing
- Simulation data (S – C?)
  - Data generated from modeling processes
- Peer reviewed data (S – C)
  - Genome banks
- Software (S – C)

STABLE: Derives from simulations, reductions, measurements, experiments

EPHEMERAL: Cannot be reproduced or reconstructed as it is time-sensitive

COSTLY: Stable but costly to regenerate

Assessment of the reproducibility value of your data in relation to the goals of your research during the early research stages will aid in scheduling your data and shaping your data management activities.
Data structures, types, sizes and needs

- Data structure
  - Structured data
    - Relational Data Base Management System
    - Schema
  - Unstructured data
    - Simple or complex data objects
    - Record-keeping system
- File formats
  - Library of Congress sustainability evaluation
    - Self documenting, ubiquitous, non-proprietary, etc.
- Data growth and content
- Data origin, mode of capture, transmission, workflows
- Platform
- Results as software, publications, demos, videos

Each research project has specific data types and a software architecture to render data functionalities that will meet the research goals
METADATA AND PROVENANCE
Keywords for this requirement

• Integration
  • XML standards, ontologies, ETL
• Interoperability
  • Metadata mapping
• Access
  • Metadata aggregators, catalogues, semantic web
• Reproducibility
  • Documentation of relationships between data and the research processes
• Quality
  • Audit, veracity, integrity
Metadata standards

- Description of the collection
  - Data Documentation Initiative
  - Core Scientific Metadata Model
  - Dublin Core and other descriptive standards
    - Different levels of granularity

- Data object level metadata
  - Darwin Core for biodiversity data
  - VRA Core for visual resources data
  - CIDOC CRM for cultural heritage documentation
  - GRIB, HDF5, NetCDF, DICOM, etc.

- Preservation and technical metadata
  - PREMIS, MIX, EXIF, etc.

- Metadata exchange and documentation of relationships between objects
  - METS

CSMD, 2010
Provenance

- Reproducibility, validity, audit of your research
- Documentation of scientific workflows
  - Ancestral data – transformations/analysis/curation – derived data
  - XML to capture and preserve relationships
  - Difficulties to track provenance across systems
  - Specific and domain science tools
    - Protégé
    - EU Provenance project
    - Vistrails
- Store and disseminate provenance data
  - Provides context to the interpretation

Simmhan et al, 2005

Figure 1 Taxonomy of Provenance
<table>
<thead>
<tr>
<th>Condition</th>
<th>Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribution</td>
<td>You must give the original author credit.</td>
</tr>
<tr>
<td>Non-Commercial</td>
<td>You may not use this work for commercial purposes.</td>
</tr>
<tr>
<td>No Derivative Works</td>
<td>You may not alter, transform, or build upon this work.</td>
</tr>
</tbody>
</table>

**CONFIDENTIALITY, PRIVACY, LICENCING**

**CONFIDENTIAL**

You are free:
- To Share: To copy, distribute and use the database.
- To Create: To produce works from the database.
- To Adapt: To modify, transform and build upon the database.

As long as you:
- Blank: This section is intentionally left blank. The PDCL imposes no restrictions on your use of the PDCL licensed database.
Confidentiality

- Confidentiality of human subjects
  - Institutional Review Board (IRB)
- Data sensitivity
  - Content linked to identification
- Access restrictions
  - Data
  - Computer equipment
  - Scheduling and or redacting data points as needed
- Family Educational Rights and Privacy Act – FERPA
- Health research data – HIPAA

*Confidentiality and intellectual property issues impact the possibilities to share, analyze, and archive your data and thus on the types of systems the data will be stored in.*
IPR and licensing

- Content such as images, video, software, articles, etc. are subjected to IP Rights
  - Provide usage under different Creative Commons Licenses
- Data/facts are not copyrightable
- Database design/schema is copyrightable
- Databases may include
  - Content (subject to IPR)
  - Data (not subjected to IPR)
  - Different options under the Open Data Commons
- Databases may have thousands of data sources
  - Hard to comply with attribution
    - Hard to comply with attribution
    - You may use CC0 from Creative Commons
      - To waive all rights
SHARING AND MANAGING DATA

Transfer time vs. file size for all connections

- TACCingest Wireless
- TACCingest Wired
- iDrop Wireless
- iDrop Wired

Transfer Time (Minutes)

File Size (GB)
Sharing/managing your data

• Throughout the research process:
  – 15 specialists in remote locations need to contribute their interpretation to a database
  – Data needs to be accessed only by a number of researchers
  – Data needs to be curated before it will be publicly released
  – Data will be transferred from other repositories

• Technical Issues
  – Timely and reliable transfers
  – Finding what you need
  – Integration/interoperability
  – Organized workflows
  – From storage to compute
  – Where is everything?
High-Performance compute workflow

• Data may be gathered from sensors or public repositories and is moved to temporary storage on HPC system
• Computational output is written to temporary storage on HPC system
• Analysis is done on the same or a different system
• Output data (reduced?) is transferred to archive or to external system
HPC Data Management Plan

• Document data flow
  – Which systems? What data?

• Document data components
  – Input vs. Output
  – Temporary vs permanent
  – Complete output vs. analyzed results

• Establish good/consistent practices
  – Naming conventions
  – Locations of data components
“Citizen Science” workflow

• Website is created for user submission of data
• Database used to hold user results
• Research team reviews input for data quality
• Create derived/aggregated data from vetted input data
• Create automated processes for aggregating data
• Data may be “created” on an ongoing basis
“Citizen Science” DMP

• Document database structure
• Provide for/document plans for backup of database
• Document plans for ongoing operation of data collection website
• Will data be disseminated from the website? In raw or derived form? Under what license?
Bioinformatics workflow

• Collect data from public and/or private sources (GenBank, other labs)
• Create data using next-gen sequencing facility
• Assemble and analyze created data in comparison with gathered data
• Visualize/summarize results
• Submit new sequence data to GenBank or other repository
Bioinformatics DMP

• Document data sources, license, types, and quality

• Define practices for gathering data and for generating new data
  – Provenance? Other metadata?

• Document plans for releasing data
  – Pre- or post-publication?
  – Which repositories are appropriate?
STORAGE and PRESERVATION
Preservation

• File formats & technologies
  – Open source, documented
  – Unencrypted and uncompressed

• Gather useful information for preservation decision making
  – Technical metadata
  – PREMIS preservation event metadata
  – Characterization information= file formats in relation to their risk.

• No individual storage device is safe
  – Safe data storage and replication

• Reliable data transfer
Long-Term Preservation?

• Important to show that you have considered the full lifecycle of your data
• Few “permanent” storage options available for large/complex research data
• Plan for the lifetime of your grant, and years beyond when needed and possible
• TACC and others will work with you to make these kinds of arrangements
Options for long term retention of research data

• Repositories (centralized model)
  – Domain specific and institutional
  – Static data
  – Collection size and functionalities limitations

• Researcher gives access (decentralized model)
  – Issues with maintenance and permanence

• Cloud storage
  – With preservation interface
  – Selected functionalities
  – Contracts

• Evolving data storage (semi-centralized)
  – Flexibility to configure/curate your collection
  – Seamless transition between research stages
  – Long-term agreements
Institutional options

- University of Texas and many other institutions have repositories for research data and/or publication materials
- Depending on the size and the nature of data they may be available at low or no cost
- Some services handle long-term preservation
- Libraries or IT Department staff are prepared to help
Data infrastructure in DMPs

• Data Management helps keep research effort focused on research goals
• DMP should reflect the research workflow
• Research workflow will reflect the resources available/used
• XSEDE, institutional, and local resources should be part of your DMP
XSEDE and other shared resources

• If computation on XSEDE or another shared resource is part of your research plan, explain how the data will be:
  – Moved to shared resources
  – Retired from shared resources
  – Managed within those resources (archives vs scratch space)
Budget and human resources

• Having budget devoted to data management adds credibility to the plan
• Infrastructure could be a component
• Human resources are important
  – Could be as simple as designating a portion of already-allocated time to data management
  – Important to have clearly-defined roles in data management
Concluding remarks

• We have tried to provide some guidance for working through the issues
• Examples and templates are available and more are being posted all the time
• Library and other local resources may be available to you
• TACC Data Management group can work with you to develop and execute a DMP
Contact Information

Maria Esteva
Chris Jordan
data@tacc.utexas.edu
Our reach will forever exceed our grasp, but, in stretching our horizon, we forever improve our world.