Dataset metadata

9-1-2013

Sai Deng
University of Central Florida, sai.deng@ucf.edu

Find similar works at: https://stars.library.ucf.edu/ucfscholar
University of Central Florida Libraries http://library.ucf.edu

Part of the Cataloging and Metadata Commons, and the Scholarly Communication Commons

Original Citation


This Other Presentation is brought to you for free and open access by STARS. It has been accepted for inclusion in Faculty Scholarship and Creative Works by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
Dataset Metadata

Sai Deng
University of Central Florida Libraries
What will be covered?

- Dataset Metadata Service at UCF
- Understanding data, research data and datasets
- What do the funding agencies say: Data, metadata related requirements
- Data types, formats, and documentation
- ResearcherID, Scopus and ORCiD
- DOIs and data citation
- Dataset record examples, their associated standards, and data repositories
- Curation Tools for datasets
*Word cloud generated using Tagxedo.*
Dataset Metadata Service at UCF

○ “Data Set (also called ‘Dataset’) Metadata” provides researchers consultation on:
  ○ Project and dataset documentation;
  ○ Acquiring DOIs for your datasets;
  ○ Metadata standards (Common and Domain Specific);
  ○ Metadata schemas customization;
  ○ Controlled vocabularies and thesauri;
  ○ Data curation tools and practices.

○ Assists in describing basic properties of your data and enriching metadata for your datasets;
○ Supports applying controlled vocabularies or optimizing keywords to enhance the search of your datasets;
○ Helps to prepare your metadata and data for deposit and preservation.
Data

- Data are *numerical quantities* or other *factual attributes derived from observation, experiment or calculation.*

- Data are *facts, numbers, letters, and symbols* that describe an object, idea, condition, situation, or other factors. Data in a database may be characterized as *predominantly word oriented* (e.g., as in a text, bibliography, directory, dictionary), *numeric* (e.g., properties, statistics, experimental values), *image* (e.g., fixed or moving video, such as a film of microbes under magnification or time-lapse photography of a flower opening), or *sound* (e.g., a sound recording of a tornado or a fire)... Data can also be referred to as *raw, processed, or verified.*
In the context of these Principles and Guidelines [Principles and Guidelines for Access to Research Data from Public Funding], “research data” are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings.

Research Data (Cont.)

- Research data is often defined as the information (e.g. data sets, microarray, numerical data, clinical trial information, textual records, images, sound, etc.) generated or used as quantitative evidence in primary biomedical research. This research data is distinguished by the fact that it is accepted by the research community as a means to validate research findings, observations and hypotheses.

- Research data, unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results.
Dataset

○ A logically meaningful collection or grouping of similar or related data, usually assembled as a matter of record or for research, for example, the American FactFinder Data Sets provided online by the U.S. Census Bureau or the National Elevation Dataset available from the U.S. Geological Survey.

- Online dictionary for library and information science (ODLIS).
  http://www.abc-clio.com/ODLIS/odlis_A.aspx

○ A research data set constitutes a systematic, partial representation of the subject being investigated.

"Over the life course of a survey that results in a dataset - from initial conceptualization to data publication and beyond - a huge amount of metadata is typically produced. These metadata can be recorded in DDI format and re-used as the data collection, processing, tabulation, and reporting/dissemination take place."

The term “data” is used in this report to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment. -National Science Foundation (2005). Long-Lived digital data Collections: enabling Research and education in the 21st Century. P.9. Available at: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

As stated in NSF’s “Information about the Data Management Plan Required for all Proposals” for Biological Sciences, the Federal government defines data (OMB Circular A-110) as: “...the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.” This definition includes both original data (observations, measurements etc.) as well as metadata (e.g., experimental protocols, software code for statistical analysis etc.).
The NSF Grant Proposal Guide recommends the inclusion of a “data management plan” that explains how your proposal will comply with NSF’s data sharing policies. The data management plan may include:

- The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
- The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
- Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
- Policies and provisions for re-use, re-distribution, and the production of derivatives; and
- Plans for archiving data, samples, and other research products, and for preservation of access to them.

See NSF’s Grant Proposal Guide for more information.
Data Definitions and Specific Requirements from the Funding agency: National Institutes of Health (NIH)

○ NIH Data Sharing Policy and Implementation Guidance
  (http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm)

  ○ “Investigators seeking $500,000 or more in direct costs in any year should include a description of how final research data will be shared, or explain why data sharing is not possible…”

○ Data Sharing Plan (to follow immediately after the Research Plan Section)

  ○ “The precise content of the data-sharing plan will vary, depending on the data being collected and how the investigator is planning to share the data. Applicants who are planning to share data may wish to describe briefly the expected schedule for data sharing, the format of the final dataset, the documentation to be provided, whether or not any analytic tools also will be provided, whether or not a data-sharing agreement will be required…”
Data Definitions and Specific Requirements from the Funding agency: National Institutes of Health (NIH)

○ Data Documentation

○ “Regardless of the mechanism used to share data, each dataset will require documentation. (Some fields refer to data documentation by other terms, such as metadata or codebooks)… Documentation provides information about the methodology and procedures used to collect the data, details about codes, definitions of variables, variable field locations, frequencies, and the like. The precise content of documentation will vary by scientific area, study design, the type of data collected, and characteristics of the dataset.”

○ “It is appropriate for scientific authors to acknowledge the source of data upon which their manuscript is based. Many investigators include this information in the methods and/or reference sections of their manuscripts. Journals generally include an acknowledgement section… Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication.”
Changes to Public Access Policy Compliance Efforts Apply to All Awards with Anticipated Start Dates on or after July 1, 2013 (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-13-042.html)

For non-competing continuation grant awards with a start date of July 1, 2013 or beyond:

- NIH will delay processing of an award if publications arising from it are not in compliance with the NIH public access policy (which requires paper to be posted to PubMed within a year after publication; PMCID included in citations);

- Investigators will need to use My NCBI to enter papers onto progress reports. Papers can be associated electronically using the RPPR, or included in the PHS 2590 using the My NCBI generated PDF report.
Data Definitions and Specific Requirements from the Funding agency: National Endowment for the Humanities (NEH)

- Data Management Plans for NEH Office of Digital Humanities: Proposals and Awards
  - “‘Data’ is defined as materials generated or collected during the course of conducting research.
  - Examples of humanities data could include citations, software code, algorithms, digital tools, documentation, databases, geospatial coordinates (for example, from archaeological digs), reports, and articles.
  - Excluded, however, are things such as preliminary analyses, drafts of papers, plans for future research, peer-review assessments, communications with colleagues, materials that must remain confidential until they are published, and information whose release would result in an invasion of personal privacy (for example, information that could be used to identify a particular person who was one of the subjects of a research study).
  - Many variables govern what constitutes “data” and the management of data, and each discipline has its own culture regarding data...”
Data Definitions and Specific Requirements from the Funding agency: National Endowment for the Humanities (NEH) (Cont.)

○ Contents of the DMP

○ "Expected data. The DMP should describe the types of data, samples, physical collections, software, curriculum materials, or other materials to be produced in the course of the project. It should then describe the expected types of data to be retained..."

○ "Data formats and dissemination. The DMP should describe data formats, media, and dissemination approaches that will be used to make data and metadata available to others. Policies for public access and sharing should be described, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, and other rights or requirements..."

○ "Final Performance Reports. Final performance reports are required for all NEH awards. The final performance report must discuss the execution and any updating of the original DMP. This discussion should describe

  ○ data produced during the grant period;
  ○ data to be retained after the grant period expires;
  ○ verification that data will be available for sharing;
  ○ discussion of community standards for data format;
  ○ the plan to disseminate the data;
  ○ the format that will be used to make data available to others, including any metadata; and
  ○ the archival location of data."
Data Definitions and Specific Requirements from the Funding agency: National Aeronautics and Space Administration (NASA)

- **NASA Earth Science Data & Information Policy**
  - "...NASA has adopted the following data policy (in this context the term ‘data’ includes observation data, metadata, products, information, algorithms, including scientific source code, documentation, models, images, and research results):
    - NASA will plan and follow data acquisition policies that ensure the collection of long-term data sets needed to satisfy the research requirements of NASA's Earth science program.
    - NASA commits to the full and open sharing of Earth science data obtained from NASA Earth observing satellites, sub-orbital platforms and field campaigns with all users as soon as such data become available.
    - There will be no period of exclusive access to NASA Earth science data...
    - NASA will make available all NASA-generated standard products along with the source code for algorithm software, coefficients, and ancillary data used to generate these products.
    - All NASA Earth science missions, projects, and grants and cooperative agreements shall include data management plans to facilitate the implementation of these data principles.
    - [More on data access...]
    - ... Data archives will include easily accessible information about the data holdings, including quality assessments, supporting relevant information, and guidance for locating and obtaining data.
    - [More on partnerships and metrics...]

- **NASA Guidebook for Proposers Responding to a NASA Research Announcement (NRA) or Cooperative Agreement Notice (CAN)**
Data Definitions and Specific Requirements from the Funding agency: United States Department of Agriculture (USDA)


- USDA Forest Service
  - Forest Inventory & Analysis and National Research Data Archive
    (http://www.fs.fed.us/research/products/data/)

- The Forest Service Metadata Users Guide
  (http://www.fs.fed.us/gac/metadata/index.html)
  - "The steps to get from the REAL WORLD to a GIS product are detailed and many. With each step, information must be gathered in the form of METADATA or information about data. Metadata describes the overall history of our data, including content, quality, condition, and other characteristics…"

- 6 steps to create metadata: Gathering Metadata- Preparation-Creating FGDC Metadata-Publishing-Using Metadata-Maintaining Metadata"
  (http://www.fs.fed.us/gac/metadata/step1.html)
Data Definitions and Specific Requirements from the Funding agency: US Geological Survey (USGS)

- **USGS Data Management**
  

- **How do I Create Metadata?**
  

  - "Metadata describe information about a dataset, such that a dataset can be understood, re-used, and integrated with other datasets. Information described in a metadata record includes **where the data were collected**, **who is responsible for the dataset**, **why the dataset was created**, and **how the data are organized**. Metadata generally follow a standard format, making it easier to compare datasets and to transfer files electronically."

- **Standard**: "Make sure your record is compliant with **FGDC standards**."


- **Tools**: Online Metadata Editor (OME); Metavist; FGDC Metadata Editor for ArcGIS 10; Morpho...

- **USGS Core Science Metadata Clearinghouse** ([http://mercury.ornl.gov/clearinghouse](http://mercury.ornl.gov/clearinghouse))
Research Data Classification

- Research data can be generated for different purposes and through different processes. In general, it can include the following types of data:
  - **Observational**: data captured in real-time, usually irreplaceable. For example, sensor data, survey data, sample data, neuroimages.
  - **Experimental**: data from lab equipment, often reproducible, but can be expensive. For example, gene sequences, chromatograms, toroid magnetic field data.
  - **Simulation**: data generated from test models where model and metadata are more important than output data. For example, climate models, economic models.
  - **Derived or compiled**: data is reproducible but expensive. For example, text and data mining, compiled database, 3D models.
  - **Reference or canonical**: a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated. For example, gene sequence databanks, chemical structures, or spatial data portals.
Research Data Format (by Categories)

- **Text** - flat text files, Word, PDF, RTF, XML.
- **Numerical** - Statistical Package for the Social Sciences (SPSS), Stata, Excel.
- **Multimedia** - jpeg, tiff, dicom, mpeg, quicktime.
- **Models** - 3D, statistical.
- **Software** - Java, C programs.
- **Discipline specific** - Flexible Image Transport System (FITS) in astronomy, Crystallographic Information File (CIF) in chemistry.
- **Instrument specific** - Olympus Confocal Microscope Data Format, Carl Zeiss Digital Microscopic Image Format (ZVI).
DOE generates **scientific research data** in many forms, both text and non-text. Much of the Department's text-based **R&D results** are readily available via OSTI databases. OSTI has broadened efforts to make **non-text scientific and technical information (STI)** available as well, providing access to underlying non-text data such as numeric files, computer simulations and interactive maps, as well as **multimedia** and **scientific images**.

Research Data Examples

- Documents (text, Word), spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films
- Test responses
- Slides, artifacts, specimens, samples
- Collection of digital objects acquired and generated during the process of research
- Data files
- Database contents (video, audio, text, images)
- Models, algorithms, scripts
- Contents of an application (input, output, log files for analysis software, simulation software, schemas)
- Methodologies and workflows
- Standard operating procedures and protocols

Other research records:
- Correspondence
- Project files
- Grant applications
- Ethics applications
- Technical reports
- Research reports
- Master lists
- Signed consent forms
Data Documentation and Metadata

- To make your data easy to understand and analyze through your research lifecycle and in the long term, it is considered good practice to document your data. **Data documentation** is part of the data curation process.

- Research data can be documented at various levels: **Project level, File or database level** and **Variable or item level**.

- Documentation is **meant to be read by humans**; some metadata is designed more for machine processing than human readability.

- Documentation and metadata are different things. However, **metadata can be taken as a type of documentation**.

- This session will explain an important metadata element, **identifier**, including **identifier for researcher and for dataset**; it will cover some **dataset record examples, their related standards, and data repositories**.

- More metadata information will be covered in the Metadata Services session.
ResearcherID (Thomson Reuters)

- ResearcherID (www.researcherid.com)
  - ResearcherID provides a solution to the author ambiguity problem within the scholarly research community.
  - Each member is assigned a unique identifier to enable researchers to:
    - Manage their publication lists;
    - Track their times cited counts and h-index;
    - Identify potential collaborators and avoid author misidentification.

- ResearcherID information integrates with Thomson Reuters’ Web of Knowledge and is ORCID compliant, allowing you to:
  - Claim and showcase your publications from a single one account;
  - Search the registry to find collaborators;
  - Review publication lists and explore how research is used around the world.
ResearchersID Example

- Penny Beile
  - ResearcherID: I-5179-2013
  - Scopus Author ID: 6508098666

http://www.researcherid.com/rid/I-5179-2013
Scopus Author ID (Elsevier)

- Scopus is the largest abstract and citation database of peer-reviewed literature, features smart tools to track, analyze and visualize research.
- Scopus Author IDs are system generated.

- Penny Beile
  - ResearcherID: I-5179-2013
  - Scopus Author ID: 6508098666
  - ORCid: 0000-0003-3502-4865

http://www.scopus.com/authid/detail.url?authorId=6508098666
ResearchID Example (log in mode)

- Sai Deng
  - RID: C-3066-2013
  - Scopus Author ID: 23979305900
  - ORCID: 0000-0002-3681-4888
ResearcherID Example (log in mode) (Cont.)

- Exchange Data with ORCID

https://www.researcherid.com/DisplayOrcidMenu.action

RESEARCHERID

ORCID

You have an ORCID associated with your ResearcherID. Data can be exchanged between your ORCID and ResearcherID accounts. More information

What data would you like to exchange between ResearcherID and ORCID?

Profile data  Go

Send ResearcherID publications to my ORCID account  Go

Retrieve ORCID publications into my ResearcherID account  Go

Back to My Researcher Profile
ORCiD: Open Researcher & Contributor ID

- ORCiD: Combination of ResearcherID (Thomson Reuters) and Contributor ID (CrossRef);
- Independent of publishers.

Sai Deng
http://orcid.org/0000-0002-3681-4888

Keywords: digital libraries, metadata, data curation, scholarly communication
Other IDs: ResearcherID: C-3066-2013

http://orcid.org/0000-0002-3681-4888
ORCID Linked to Scopus and ResearcherID

- Penny Beile
  - ResearcherID: I-5179-2013
  - Scopus Author ID: 6508098666
  - ORCID: 0000-0003-3502-4865

http://orcid.org/0000-0003-3502-4865
Dataset Identifiers

- Digital Object Identifier (DOI)
  - e.g. http://dx.doi.org/10.3886/ICPSR20363.v1

- Archival Resource Keys (ARKs)
  - e.g. http://ark.cdlib.org/ark:/13030/tf5p30086k?

- Handles
  - e.g. http://soar.wichita.edu/handle/10057/3031

- Persistent URLs (PURLs)
  - All can be resolved to an internet location.
**DOI**

- **Digital Object Identifier (DOI):** an identifier scheme administered by the International DOI Foundation. It is built on the Handle System.

- **Example:**
  
  [http://dx.doi.org/10.3886/ICPSR20363.v1](http://dx.doi.org/10.3886/ICPSR20363.v1)

<table>
<thead>
<tr>
<th><a href="http://dx.doi.org/">http://dx.doi.org/</a></th>
<th>10.3886/</th>
<th>ICPSR20363.v1</th>
</tr>
</thead>
<tbody>
<tr>
<td>resolver service</td>
<td>prefix</td>
<td>suffix</td>
</tr>
<tr>
<td></td>
<td>(assigning body)</td>
<td>(resource)</td>
</tr>
</tbody>
</table>
Dataset Registration and DataCite

- **DataCite**: A global citations framework for data with member institutions offering services and advice to researchers.

- Individuals wishing to register a DOI for their dataset normally do so via their data repository, rather than directly through DataCite.

- Any repository wishing to register DOIs needs to obtain a username and password from DataCite to gain access to the registration service.

- Alternatively, the organization can manage its DOIs through a third-party service such as EZID.
  
  ([http://www.dcc.ac.uk/resources/how-guides/cite-datasets#x1-17000](http://www.dcc.ac.uk/resources/how-guides/cite-datasets#x1-17000))

- **DataCite Membership**
  - Affiliated membership fee: 1700 p.a
  - Doesn’t need to be a member, can work with a DataCite member or a third-party service to get DOIs for datasets
DataCite Metadata

- ICPSR (Interuniversity Consortium for Political and Social Research): an associate member of DataCite.

- ICPSR’s “How to prepare citation”:
  - Citation required basic elements:
    - Identifier
    - Creator
    - Title
    - Publisher
    - Publication Year

- For example:
  - Persistent URL: http://dx.doi.org/10.3886/ICPSR20363.v1

- Can be exported as RIS (generic format for RefWorks, EndNote, etc.) or EndNote XML (EndNote X4.0.1 or higher)
## DataCite Metadata Schema

- **DataCite Metadata Schema 3.0** (released 2013-07-24; preferred)

### Table 1: DataCite Mandatory Properties

<table>
<thead>
<tr>
<th>ID</th>
<th>Property</th>
<th>Obligation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identifier (with type sub-property)</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>Creator (with name identifier sub-properties)</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>Title (with optional type sub-properties)</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>Publisher</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>PublicationYear</td>
<td>M</td>
</tr>
</tbody>
</table>

### Table 2: DataCite Recommended and Optional Properties

<table>
<thead>
<tr>
<th>ID</th>
<th>Property</th>
<th>Obligation</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Subject (with scheme sub-property)</td>
<td>R</td>
</tr>
<tr>
<td>7</td>
<td>Contributor (with type and name identifier sub-properties)</td>
<td>R</td>
</tr>
<tr>
<td>8</td>
<td>Date (with type sub-property)</td>
<td>R</td>
</tr>
<tr>
<td>9</td>
<td>Language</td>
<td>O</td>
</tr>
<tr>
<td>10</td>
<td>ResourceType (with general type description sub-property)</td>
<td>R</td>
</tr>
<tr>
<td>11</td>
<td>AlternateIdentifier (with type sub-property)</td>
<td>O</td>
</tr>
<tr>
<td>12</td>
<td>RelatedIdentifier (with type and relation type sub-properties)</td>
<td>R</td>
</tr>
<tr>
<td>13</td>
<td>Size</td>
<td>O</td>
</tr>
<tr>
<td>14</td>
<td>Format</td>
<td>O</td>
</tr>
<tr>
<td>15</td>
<td>Version</td>
<td>O</td>
</tr>
<tr>
<td>16</td>
<td>Rights</td>
<td>O</td>
</tr>
<tr>
<td>17</td>
<td>Description (with type sub-property)</td>
<td>R</td>
</tr>
<tr>
<td>18</td>
<td>GeoLocation (with point and box sub-properties)</td>
<td>R</td>
</tr>
</tbody>
</table>
DataCite XML File

http://www.icpsr.umich.edu/icpsrweb/ICPSR/datacite/studies/20363
Dataset Record Examples, Their Associated Standards, and Data Repositories

- Social Science Dataset
- Humanities Dataset
- Biological Science Dataset
- Biotechnology Dataset
- Geospatial and Earth Science Dataset
Social Science Dataset

ICPSR: Interuniversity Consortium for Political and Social Research.

http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/20363/archives/NACJD&lq=%22university+of+central+florida%22&permitt5B0%5D=AVAILABLE&x=-999&y=-84

Field Labels:
Title
Principal investigator(s)
Summary
Access notes
Dataset(s)
Field Labels:

Study description
Citation
Funding
Scope of study
• Subject terms
• Smallest geographic unit
• Geographic coverage
• Time period
• Date of collection
• Unit of observation
• Universe
• Data types
• Data collection notes
Methodology
• Study purpose
• Study design
Sample:
To select women for Part 1 (Female Interviews) of the study, researchers entered into a cooperative agreement with a large, general-purpose shelter for the homeless in each of Jacksonville, Miami, Orlando and Tampa. All of the shelters where respondents were solicited were general-purpose homeless facilities, not battered-women's facilities, and not special-purpose facilities devoted exclusively to teens, to the addicted, or to the mentally ill.

Researchers attempted to interview the first 200 women who came through the door of the participating facilities during the data collection period. Recognizing the logistical difficulties of implementing any specific sampling plan in a social service context often characterized by crisis and relative chaos, researchers allowed for some deviation from the prescriptum. Efforts were made to interview every woman who sought services at the respective facilities until the goal of 200 interviews per site was reached.

Realizing also that interviewing each woman that came through the door would not always be possible, researchers tried to interview everyone who could be included from the facilities.

For Part 2 (Male Interviews), men were selected in similar fashion as females but only at the Orlando facility and only for the quota of 100 and was met.

You can find more information via the sample characteristics utility.

Mode of Data Collection: face-to-face interview

Description of Variables:
For Part 1 (Female Interviews), the data include information related to the respondent’s living conditions in the past month, lifetime experience with homelessness, the respondent’s partner, and living conditions as a child. There are also variables related to childhood experience with violence, experience with forced sexual situations, adult experience with violence, and basic demographic information. As well, there is information covering such areas as experience with stalking, self-image, use of alcohol and drugs, current health, and financial status. There is also a self-report of criminal history, information related to how the respondent spent her days and evenings, and the physical environment surrounding the respondent during the day and evening. Finally, there are a small number of questions answered by the respondent respecting the interview itself.

For Part 2 (Male Interviews), the data include much of the same information as was collected in Part 1. Information from Part 1 not included in Part 2 primarily includes questions pertaining to experience with forced sexual situations and questions related to pregnancy and children.

Response Rates: Not applicable.

Presence of Common Scales: The Conflict Tactics Scale (CTS) The Personal History Form (PHF) The Addiction Severity Index (ASI)

Extent of Processing: ICPSR data undergo a confidentiality review and are altered when necessary to limit the risk of disclosure. ICPSR also routinely creates ready-to-go data files along with setups in the major statistical software formats as well as standard codebooks to accompany the data. In addition to these procedures, ICPSR performed the following processing steps for this data collection:
- Created variable labels and/or value labels.
- Standardized missing values.
- Checked for undocumented or out-of-range codes.
Social Science Dataset in DDI Format

- **Data Documentation Initiative (DDI):** A metadata specification for the social and behavioral sciences. It is an XML metadata standard for documenting numeric data. Detailed information is available at: [http://www.ddialliance.org/](http://www.ddialliance.org/)

- **Projects using the DDI** ([http://www.ddialliance.org/ddi-at-work/projects](http://www.ddialliance.org/ddi-at-work/projects))

- **DDI-compliant data repository:**
  - **ICPSR**
    - Data deposit form: [https://www.icpsr.umich.edu/cgi-bin/ddf2](https://www.icpsr.umich.edu/cgi-bin/ddf2)
    - UCF is a member of ICPSR.
Social Science Dataset in DDI Format (Cont.)

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<codeBook version="1.2.2" ID="ICPSR20363">
  <docDscr>
    <citation>
      + <titStmt></titStmt>
    </citation>
    <prodStmt>
      + <producer abbr="ICPSR">
        <ExtLink URI="http://www.icpsr.umich.edu/images/icpsr-logo.gif" title="ICPSR Logo" role="image"/>
        Inter-university Consortium for Political and Social Research
        <ExtLink URI="http://www.icpsr.umich.edu/ICPSR/" title="URL of ICPSR Web Site"/>
      </producer>
    </prodStmt>
    + <verStmt>
      <holdings URI="http://www.icpsr.umich.edu/icpsrweb/ICPSR/ddi2/studies/20363"/>
    </verStmt>
  </docDscr>
  <stdyDscr>
    + <fileDscr ID="F1"/>
      - <fileTxT ID="Part1">
        <fileName>Female Interviews</fileName>
      </fileTxT>
    </fileDscr>
    + <fileDscr ID="F2"/>
  </fileDscr>
</codeBook>
```

Included Fields:
- citation
- titleStmt
- prodStmt
- verStmt
- holdings

http://www.icpsr.umich.edu/icpsrweb/ICPSR/ddi2/studies/20363
The Study Description consists of information about the data collection, study, or compilation that the DDI-compliant documentation file describes. This section includes information about how the study should be cited, who collected or compiled the data, who distributes the data, keywords about the content of the data, summary (abstract) of the content of the data, data collection methods and processing, etc.
Social Science Dataset in DDI Format (Cont.)

Included Fields:
- fileDscr
- fileTxt
- fileName

**fileDscr**
Data Files Description
Information about the data file(s) that comprises a collection. This section can be repeated for collections with multiple files.
Social Science Dataset in OAI_DC Format

OAI_DC
OAI DC is an XML format for the serialisation of Simple Dublin Core metadata descriptions. It is used within the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH requires that data providers support the oai_dc metadata format.

Fields:
- oai_dc:dc
dc:title
dc:creator
dc:subject
dc:description
dc:date
dc:type
dc:identifier
dc:coverage
dc:rights

Wright, James D
Jasinski, Jana L.
Mustaine, Elizabeth
Wesely, Jennifer
abuse
2010-11-22
Survey data
20363
10.3886/ICPSR20363.v1
Florida
Jacksonville
Miami
Orlando
Tampa
United States
2003--2004
http://www.icpsr.umich.edu/icpsrweb/ICPSR/dc/studies/20363
Social Science Dataset in MARC21 XML Format

MARC21 XML
MARC21
encoded in XML.

Fields:
collection
Record
Leader
controlfield
datafield
subfield

http://www.icpsr.umich.edu/icpsrweb/ICPSR/marc/studies/20363
Charles Brockden Brown Electronic Archive and Scholarly Edition

“The Charles Brockden Brown Electronic Archive and Scholarly Edition is an editorial collective that is preparing an MLA-CSE approved print edition of Charles Brockden Brown’s (1771-1810) personal letters, poetry, and selected periodical writings to be published by Bucknell University Press (7 volumes). A searchable archive of all of Brown’s works (unedited) is also being developed.” –From project website.

This archive uses “The Text Encoding Initiative’s Guidelines for Electronic Text Encoding and Interchange” (TEI P5, http://www.tei-c.org/Guidelines/P5/) for text encoding;

It uses eXtensible Text Framework (XTF) to index bibliographical and primary-document materials.

All information is from the project website: http://www.brockdenbrown.ucf.edu
Humanities Data: UCF Project (Cont.)

- Charles Brockden Brown Electronic Archive and Scholarly Edition

*Shown here are the documents display front-end. Don’t see TEI XML files for download.*


ENRICH: European Networking Resources and Information concerning Cultural Heritage.

- Below is the overall structure of an ENRICH-conformant XML document.

```
<TEI>
  <teiHeader>
    <![-- ... metadata describing the manuscript -->]>
  </teiHeader>
  <facsimile>
    <![-- ... metadata describing the digital images -->]>
  </facsimile>
  <text>
    <![-- (optional) transcription of the manuscript -->]>
  </text>
</TEI>
```

The minimal required structure for `<teiHeader>`:

```
<TEI>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>[Title of manuscript]</title>
      </titleStmt>
      <publicationStmt>
        <distributor>[name of data provider]</distributor>
        <idno>[project-specific identifier]</idno>
      </publicationStmt>
      <sourceDesc>
        <msDesc xml:id="ex5" xml:lang="en">
          <![-- [full manuscript description ]-->]
        </msDesc>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
</TEI>
```

<teiHeader> (TEI header) supplies the descriptive and declarative information making up an electronic title page prefixed to every TEI-conformant text.

msDesc (manuscript description) provides detailed information about a single manuscript.

Examples from ENRICH
(http://projects.oucs.ox.ac.uk/ENRICH/Deliverables/referenceManual_en.html)

Fields:
msDesc
  msIdentifier
    Settlement
    repository
    Idno
    altIdentifier
  msContents
    p
      quote
      title
  physDesc
    p
      material
  History
    p
      origPlace
      origDate
      quote

The official TEI P5 guideline is at:

More TEI projects and examples are available at the TEI website: http://www.tei-c.org/Activities/Projects/
More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biology Letters 8(5): 783-786.

We present the first genomic-scale analysis addressing the phylogenetic position of turtles, using over 1,000 loci from representatives of all major reptile lineages including tuatara…

A digital repository for data underlying the international scientific and medical literature.
## Biological Science Dataset

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.relation.isreferencedby</td>
<td>PMID:22593086</td>
</tr>
<tr>
<td>dc.subject</td>
<td>ultraconserved elements</td>
</tr>
<tr>
<td>dc.subject</td>
<td>phylogenomic</td>
</tr>
<tr>
<td>dc.subject</td>
<td>phylogenetics</td>
</tr>
<tr>
<td>dc.subject</td>
<td>reptiles</td>
</tr>
<tr>
<td>dc.subject</td>
<td>turtles</td>
</tr>
<tr>
<td>dc.subject</td>
<td>evolution</td>
</tr>
<tr>
<td>dc.subject</td>
<td>archosaurs</td>
</tr>
<tr>
<td>dc.title</td>
<td>Data from: More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs</td>
</tr>
<tr>
<td>dc.type</td>
<td>Article</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Pantherophis guttata</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Pelomedusa subrufa</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Chrysemys picta</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Alligator mississippiensis</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Crocodylus porosus</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Chrysemys picta</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Sphenodon tuatara</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Gallus gallus</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Taeniopygia guttata</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Anolis carolinensis</td>
</tr>
<tr>
<td>dwc.ScientificName</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>dc.contributor.correspondingAuthor</td>
<td>Faircloth, Brant C.</td>
</tr>
<tr>
<td>prism.publicationName</td>
<td>Biology Letters</td>
</tr>
</tbody>
</table>

**Dryad**

(https://datadryad.org/)

- It is built upon the open-source **DSpace** repository software;

- It utilizes a combination of **Dublin Core (DC)** and **Darwin Core (DwC)** metadata standards.

- Digital Object Identifiers (**DOIs**) provided by DataCite through EZID.
<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Download</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>turtles-all-probes.fasta</td>
<td>Fasta file of probe sequences commercially synthesized to target UCE loci in reptiles. Probes are identical to those used in Faircloth et al. 2012 (Systematic Biology; doi: 10.1093/sysbio/sys004)</td>
<td>turtles-all-probes.fasta (805.4Kb)</td>
<td>View File Details</td>
</tr>
<tr>
<td>turtles-contigs-enriched-from-taxa.fasta</td>
<td>FASTA file of contigs assembled from raw reads generated by sequencing enriched libraries of several reptile species.</td>
<td>turtles-contigs-enriched-from-taxa.fasta (11.31M)</td>
<td>View File Details</td>
</tr>
<tr>
<td>turtles-lastz-matches-to-reptiles.tar.bz2</td>
<td>LASTZ matches of probes to species-specific contigs in turtle-contigs-enriched-from-taxa.</td>
<td>turtles-lastz-matches-to-reptiles.tar.bz2 (380.4Kb)</td>
<td>View File Details</td>
</tr>
<tr>
<td>turtles-probe-matches-to-reptiles.sqlite</td>
<td>SQLite database of LASTZ matches (from turtles-lastz-matches-to-reptiles.tar.bz2) between assembled contigs in turtle-contigs-enriched-from-taxa fasta and the UCE probes in turtles-all-probes.fasta. This database is used to construct data sets containing loci that are shared across taxa and prep data for alignment.</td>
<td>turtles-probe-matches-to-reptiles.sqlite (496.6Kb)</td>
<td>View File Details</td>
</tr>
</tbody>
</table>
Biological Science Dataset (Cont.)

- If clicking View File Details, it displays:

  **Simple View**

  **turtles-all-probes.fasta**

  When using this data, please cite the original article:

  Additionally, please cite the Dryad data package:
  Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC (2012) Data from: More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Dryad Digital Repository. [doi: 10.5061/dryad.75nv22qj](https://doi.org/10.5061/dryad.75nv22qj)

  **Cite | Share**

  **Files in this item**

  **Name:** turtles-all-probes.fasta
  **Size:** 845.1Kb
  **Format:** application/x-fasta
  **Description:** dataset-file
  **Checksum (MD5):** d90214f9786ef085d83a7eab10d435b03

  **Description**
  Fasta file of probe sequences commercially synthesized to target UCE loci in reptiles. Probes are identical to those used in Faircloth et al. 2012 (Systematic Biology; doi: 10.1093/sysbio/sys004)

  To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data. [Learn more about data licensing](https://creativecommons.org/licenses/by/4.0/).
Deposit in Dryad

**Fees:** Starting Sept. 1, 2013, Dryad will charge fees upon submission. The submitter is asked to pay this fee at the time of submission unless:

- the associated journal, or another organization, has already contracted with Dryad to cover the submission fee, or
- the submitter is based in a country classified by the World Bank as a low-income or lower-middle-income economy.
- Additional submission fees will apply to data packages in excess of 10 GB and from journals without integrated submission.

**Pricing Plan Comparison Tool** ([http://www.datadryad.org/pages/pricing](http://www.datadryad.org/pages/pricing))

<table>
<thead>
<tr>
<th>Payment Plan</th>
<th>Member</th>
<th>Non-member</th>
<th>Minimum purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Voucher Plan</td>
<td>USD$65 per data package</td>
<td><strong>USD$70 per data package</strong></td>
<td>25 vouchers</td>
</tr>
<tr>
<td>2. Deferred Payment Plan</td>
<td>USD$70 per data package</td>
<td>USD$75 per data package</td>
<td>1 yr contract</td>
</tr>
<tr>
<td>3. Subscription Plan</td>
<td>annual fee based on USD$25 per published research article</td>
<td>annual fee based on USD$30 per published research article</td>
<td>2 yr contract</td>
</tr>
<tr>
<td>For individuals: Pay on submission</td>
<td>NA</td>
<td>USD$80 per data package, payable by the submitter</td>
<td>1 data package</td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Pluenneke, David [Collector]</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>dc.contributor.author</td>
<td>Pluenneke, David [Cataloger]</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.date.accessioned</td>
<td>2010-09-14T19:48:27Z</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.date.available</td>
<td>2010-09-14T19:48:27Z</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.date.created</td>
<td>1987-07-15</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.date.issued</td>
<td>1987-07-15</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.identifier</td>
<td>3647</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.identifier.uri</td>
<td><a href="http://library.wichita.edu/techserv/herbarium/browse.asp?id=3647">http://library.wichita.edu/techserv/herbarium/browse.asp?id=3647</a></td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.identifier.uri</td>
<td><a href="http://hdl.handle.net/10057/3031">http://hdl.handle.net/10057/3031</a></td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.format.extent</td>
<td>1034821 bytes</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.format.mimetype</td>
<td>image/jpeg</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.language.iso</td>
<td>en_US</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dc.publisher</td>
<td>Wichita State University. Dept. of Biological Sciences</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.relation</td>
<td>U.S. Dept. of Agriculture Natural Resources Conservation Service</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.rights</td>
<td>Copyright Wichita State University, 2010</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.source</td>
<td>WSU herbarium</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.subject</td>
<td>Asteraceae</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.subject</td>
<td>Achillea lanulosa Nutt.</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.subject</td>
<td>Yarrow</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.subject</td>
<td>United States -- Colorado -- Teller Co.</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.title</td>
<td>Achillea lanulosa Nutt. [Taxon]</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.title.alternative</td>
<td>Yarrow</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.type</td>
<td>Image</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.coverage.spatial</td>
<td>North America</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.coverage.spatial</td>
<td>United States</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.coverage.spatial</td>
<td>Colorado</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.coverage.spatial</td>
<td>Teller Co.</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dc.coverage.spatial</td>
<td>NW1/4, SE1/4, Sec 32, T13S, R69W; Woodland Pk Quad</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dwc.family</td>
<td>Asteraceae</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dwcverbatimElevation</td>
<td>11600'</td>
<td>en_US</td>
<td></td>
</tr>
<tr>
<td>dwc.habitat</td>
<td>Alpine</td>
<td>en_US</td>
<td></td>
</tr>
</tbody>
</table>

**DSpace:** A sample herbarium record from “Virtual Herbarium Collection” in Shocker Open Access Repository (SOAR)

**Files in this item**

Name: HERBARIUM_58.jpg
Size: 1010.Kb
Format: image/jpeg

[View/Open](http://soar.wichita.edu/handle/10057/3031?show=full)

[View in Browser](http://soar.wichita.edu/handle/10057/3031?show=full)
GenBank

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2011 Jan;39(Database issue):D32-7).

Biotechnology Dataset @ PubChem

PubChem
The leading freely-available, small compound database, part of the National Center for Biotechnology Information/National Library of Medicine’s Entrez suite of databases.

MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed.
Content Standard for Digital Geospatial Metadata (CSDGM)
(http://www.fgdc.gov/metadata/geospatial-metadata-standards)
It is maintained by the Federal Geographic Data Committee (FGDC).
Often referred to as the “FGDC Metadata Standard.”

Geospatial Platform: An Internet-based capability providing shared and trusted geospatial data, services, and applications for use by the public and by government agencies and partners to meet their mission needs.

Web display:
Data and Resources
- Web Page
- XML File
- Web Page
- Web Page
- Web Page
- Web Page

Metadata Source
- ISO-19239 Metadata
- Original FGDC Metadata

http://www.geoplatform.gov/node/243/bf5a5c64-085e-4c68-a489-93e8608d3ad1
Biological data of field activity 08CRD01 (B-1-08-VI) in U.S. Virgin Islands from 05/30/2008 to 06/13/2008

Metadata

File Identifier:

Metadata Language: eng; USA: utf8

Resource Type: Dataset

Responsible Party:

Individual Name: Clint Steele <http://walrus.wr.usgs.gov/staff/csteele.html>


Role: Point Of Contact

Contact Info: ...

Metadata Date: 2013-03-03

Metadata Standard Name: ISO 19115-2 Geographic Information - Metadata - Part 2: Extensions for Imagery and Gridded Data


http://catalog.data.gov/harvest/object/dfe4a33f-0fd2-4135-ae33-a1c526bd7a73/html
# Geospatial Dataset (Cont.)

**Data Identification**

**Abstract:** United States Geological Survey, Saint Petersburg, Florida, Center for Coastal and Watershed Studies...

**Purpose:** These data and information are intended for science researchers, students...

**Language:** eng; USA

**Citation:**

**Title:** Biological data of field activity 08CRD01 (B-1-08-VI) in U.S. Virgin Islands from 05/30/2008 to 06/13/2008

**Date:**

**Date:** 2013-03-03

**Date Type:** Publication Date


**Role:** Publisher

**Contact Info:** ...

**Point Of Contact:** ...

**Representation Type:** Vector

**Topic Category:**

**Keyword Collection:**

**Keyword:** EARTH SCIENCE > OCEANS

**Associated Thesaurus:** Global Change Master Directory (GCMD)

**Keyword:** Marine Geology

**Associated Thesaurus:** USGS CMG InfoBank

**Spatial Extent:**

- **West Bounding Longitude:** -65.75000
- **East Bounding Longitude:** -63.25000
- **North Bounding Latitude:** 18.75000
- **South Bounding Latitude:** 17.25000
Geospatial Dataset (Cont.)

Constraints: Please recognize the U.S. Geological Survey (USGS) as the source of this information. Physical materials are under controlled on-site access. Some USGS information accessed through this means may be preliminary in nature and presented without the approval of the Director of the USGS...
Legal Constraints:
Use Constraints: Other Restrictions
Other Constraints: Use Constraints: Please recognize the U.S. Geological Survey (USGS) as the source of this information. Physical materials are under controlled on-site access...

... Distribution
Distribution Format:
Format Name: ASCII
Format Version:
File Decompression Technique: No compression applied
Transfer Options:
Distributor:
Distributor Contact: ...

Quality
Scope: Dataset
Geospatial Dataset in XML View

CSDGM Fields (under idinfo):

- Citation (under citeinfo):
  - Origin
  - Pubdate
  - Title
  - Pubinfo
  - Onlink
- Descript
- Abstract
- Purpose
- Supplinf
- Timeperd
- Status
- Spdom
- Keywords
- Accconst
- Useconst
- Ptcontac
- Native
- Crossref

Top level elements:

- idinfo: Identification Information;
- dataqual: Data Quality Information;
- spdoinfo: Spatial Data Organization Information;
- spref: Spatial Reference Information;
- eainfo: Entity and Attribute Information;
- distinfo: Distribution Information;
- metainfo: Metadata Reference Information.

Content Standard for Digital Geospatial Metadata (CSDGM) Record in XML View

http://catalog.data.gov/harvest/object/dfe4a33f-0fd2-4135-ae33-a1c526bd7a73/original
Earth Science Dataset

NASA Atmospheric Science Data Center (ASDC)

Labels:
Summary
Related URL
Geographic Coverage
Spatial coordinates
Temporal Coverage

http://gcmd.gsfc.nasa.gov/KeywordSearch/KeywordSearch/Metadata.do?Portal=langley&KeywordPath=Parameters%7CATMOSPHERE%7CAIR+QUALITY%7CCARBON+MONOXIDE&OrigMetadataNode=GCMD&EntryId=MOP034&MetadataView=Full&MetadataDataType=0&lbnode=mdlb1
Earth Science Dataset (Cont.)

Directory Interchange Format (DIF): a descriptive and standardized format for exchanging information about scientific data sets.


Origin: DIF was the product of an Earth Science and Applications Data Systems Workshop (ESADS) held February 24-26, 1987 on catalog interoperability (CI). (http://gcmd.gsfc.nasa.gov/add/difguide/whatisadif.html)
<table>
<thead>
<tr>
<th>Dublin Core Metadata</th>
<th>Standard DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td><strong>Entry_Title</strong></td>
</tr>
<tr>
<td><strong>Creator</strong></td>
<td><strong>Data_Set_Citation: Dataset_Creator</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Personnel: Role: Investigator: Last_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Personnel: Role: Investigator: First_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Personnel: Role: Investigator: Middle_Name</strong></td>
</tr>
<tr>
<td><strong>Subject and Keywords</strong></td>
<td><strong>Keyword</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Parameters: Category</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Parameters: Topic</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Parameters: Term</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Parameters: Variable</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Parameters: Detailed_Variable</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Source_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Sensor_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Project</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Location</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td><strong>Summary</strong></td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td><strong>Data_Set_Citation: Dataset_Publisher</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Data_Center: Data_Center_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Data_Center: Data_Center_URL</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Data_Center: Data Center Contact, Last_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Data_Center: Data Center Contact, First_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Data_Center: Data Center Contact, Middle_Name</strong></td>
</tr>
<tr>
<td><strong>Contributor:</strong></td>
<td><strong>Personnel: Role:</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Personnel: Last_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Personnel: First_Name</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Personnel: Middle_Name</strong></td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td><strong>Data_Set_Citation: Dataset_Release_Date</strong></td>
</tr>
<tr>
<td><strong>Resource Type</strong></td>
<td><strong>Data_Set_Citation: Data_Presentation_Form</strong></td>
</tr>
</tbody>
</table>

**Scientific Dataset: Mapping of Dublin Core to DIF**
<table>
<thead>
<tr>
<th>Format</th>
<th>Group: Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution_Media</td>
<td></td>
</tr>
<tr>
<td>Distribution_Size</td>
<td></td>
</tr>
<tr>
<td>Distribution_Format</td>
<td></td>
</tr>
<tr>
<td>Fees</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource Identifier</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Center: Data_Set_ID</td>
<td></td>
</tr>
<tr>
<td>Data_Set_Citation: Online_Resource</td>
<td></td>
</tr>
<tr>
<td>Related_URL: URL_Content_Type</td>
<td></td>
</tr>
<tr>
<td>Related_URL: URL</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Related_URL: URL_Content_Type</td>
<td></td>
</tr>
<tr>
<td>Related_URL: URL</td>
<td></td>
</tr>
<tr>
<td>Source_Name</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data_Set_Language</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent_DIF</td>
<td></td>
</tr>
<tr>
<td>Data_Set_Citation: Online_Resource</td>
<td></td>
</tr>
<tr>
<td>Related_URL: URL_Content_Type</td>
<td></td>
</tr>
<tr>
<td>Related_URL: URL</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial_Coverage: Southernmost_Latitude</td>
<td></td>
</tr>
<tr>
<td>Spatial_Coverage: Northernmost_Latitude</td>
<td></td>
</tr>
<tr>
<td>Spatial_Coverage: Easternmost_Longitude</td>
<td></td>
</tr>
<tr>
<td>Spatial_Coverage: Westernmost_Longitude</td>
<td></td>
</tr>
<tr>
<td>Temporal_Coverage: Start_Date</td>
<td></td>
</tr>
<tr>
<td>Temporal_Coverage: Stop_Date</td>
<td></td>
</tr>
<tr>
<td>Paleo_Temporal_Coverage:</td>
<td></td>
</tr>
<tr>
<td>Paleo_Start_Date</td>
<td></td>
</tr>
<tr>
<td>Paleo_Temporal_Coverage:</td>
<td></td>
</tr>
<tr>
<td>Paleo_Stop_Date</td>
<td></td>
</tr>
<tr>
<td>Paleo_Temporal_Coverage:</td>
<td></td>
</tr>
<tr>
<td>Chronostratigraphic_Unit</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rights Management</th>
<th>Use_Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access_Constraints</td>
<td></td>
</tr>
</tbody>
</table>
Data Imaging Example

Screen snap shot from the CDF Windows Imaging Tool (CWIT). From Examples of CDF Applications (http://cdf.gsfc.nasa.gov/html/examples.html)

National Space Science Data Center (NSSDC) NASA’s permanent archive for space science mission data.

“Many applications have been created to support the SPDF Common Data Format (CDF) data standard.”
Curation Tools: Data Documentation and Management with Excel

DataUp: An open source tool helping researchers document, manage, and archive their tabular data. DataUp operates within the scientist's workflow and integrates with Microsoft® Excel.
http://dataup.cdlib.org/

Colectica for Microsoft Excel
A free tool to document your spreadsheet data using the Data Documentation Initiative (DDI) metadata format, the open standard for data documentation.
http://www.colectica.com/software/colecticaforexcel
Curation Tools: Data Preparation

**QualAnon: DSDR Qualitative Data Anonymizer**
This free transcript anonymization tool is designed solely to de-identify qualitative interview transcripts.
https://www.icpsr.umich.edu//icpsrweb/DSDR/tools/anonymize.jsp

**OpenRefine (ex-Google Refine)** is a powerful tool for working with messy data, cleaning it, transforming it from one format into another, extending it with web services, and linking it to databases like Freebase.
http://openrefine.org/

**Nesstar Publisher** is a free advanced data management program. It can be used for the preparation of data and metadata. It's DDI compliant.
http://www.nesstar.com/software/publisher.html
Curation Tools: XML Editing

**<oXygen/> XML Editor** is an XML tool that supports all the XML schema languages. The XSLT and XQuery support is enhanced with powerful debuggers and performance profilers. You can use `<oXygen/>` XML Editor to work with all XML-based technologies including XML databases, XProc pipelines, and web services. [http://www.oxygenxml.com/](http://www.oxygenxml.com/)

**Schematron** is a rule-based validation language for making assertions about the presence or absence of patterns in XML trees. It is a structural schema language expressed in XML using a small number of elements and XPath. [http://xml.ascc.net/resource/schematron/schematron.html](http://xml.ascc.net/resource/schematron/schematron.html)

**Altova XMLSpy** is an advanced XML editor for modeling, editing, transforming, and debugging XML-related technologies. [http://www.altova.com/xmlspy.html](http://www.altova.com/xmlspy.html)
LabTrove is a free blogging platform specifically designed for use in a research environment. It aims to serve as a highly flexible electronic notebook and data management system by integrating with a lab’s data-producing instruments; researchers can describe an experiment and associate it with its data output at the time of capture, rather than annotating after the fact. [http://www.labtrove.org/](http://www.labtrove.org/)

Kepler is a scientific workflow modeling and management system that enables users, regardless of programming experience, to set up data analysis pipelines. The software will assemble, execute, and document the services and scripts that scientists use to execute research. [https://kepler-project.org/](https://kepler-project.org/)
Data Curation Tools: Data Management Plan and Persistent ID Assignment

**DMPTool** is an online service to enable researchers to create data management plans now required by many funding agencies, and to receive tailored institutional guidance to help them in the process.

https://dmp.cdlib.org/

**DataCite**

The DataCite Consortium provides a number of services to support efforts at increasing the ease and prevalence of data citation.

http://www.datacite.org
Data Publication: Data and Digital Repository Directories

DataBib: Databib is a community-driven, annotated bibliography of research data repositories.
http://databib.org/

OpenDOAR: An authoritative worldwide directory of academic open access repositories.
http://www.opendoar.org/countrylist.php

Open Access Directory: Data Repositories A list of repositories and databases for open data. It is part of the Open Access Directory maintained by Simmons College.
http://oad.simmons.edu/oadwiki/Data_repositories
A life cycle approach

CDL Curation and Publishing Services
http://www.cdlib.org

Create, edit, share, and save data management plans

Open source add-in for Microsoft Excel as a data collection tool

Create and manage persistent identifiers

Curation repository: store, manage, and share research data

Open access scholarly publishing services: papers, journals, books, seminars & more

An infrastructure to publish and get credit for sharing research data

* This slide is by Joan Starr, California Digital Library.
http://www.slideshare.net/joanstarr/dataset-metadata-tools-approaches-for-access-preservation?from_search=1
Research Lifecycle at UCF

Data Set Related Services

Data Set Metadata Service

- **Data Set (Dataset) Metadata Service:**
  - Provides researchers consultation on:
    - Project and dataset documentation;
    - Acquiring DOIs for your datasets;
    - Metadata standards (Common and Domain Specific);
    - Metadata schemas customization;
    - Controlled vocabularies and thesauri;
    - Data curation tools and practices.

- Assists in describing basic properties of your data and enriching metadata for your datasets;
- Supports applying controlled vocabularies or optimizing keywords to enhance the search of your datasets;
- Helps to prepare your metadata and data for deposit and preservation.

- Will work with the library Scholarly Communication team and subject librarians to:
  - Introduce ORCid to researchers;
  - Promote data curation tools;
  - Provide data repositories information.
Related Library Resources


- UCF Library Research Guides [http://guides.ucf.edu](http://guides.ucf.edu)


- Research and Information Services [http://library.ucf.edu/Reference/](http://library.ucf.edu/Reference/)
  - Subject Librarians [http://library.ucf.edu/SubjectLibrarians/](http://library.ucf.edu/SubjectLibrarians/)

- More information on metadata standards, controlled vocabularies, data curation tools and repositories will be covered in the Metadata Services session.
Contact:
Sai Deng, Metadata Librarian
sai.deng@ucf.edu
407-823-4312 (Office)

Thank you!