11-19-2014

Data documentation & metadata

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

Original Citation

Deng, S. (2014). Data documentation and metadata. University of Central Florida graduate students library research workshop: Publishing in the Academy.

UCF Libraries Research Workshop

Data Documentation &Metadata

Sai Deng, Metadata Librarian

University of Central Florida Libraries

What will be covered?

Part I: The Survey and Some Data Basics

- The UCF Research Data Management Survey: Data Recording and Analysis Section Results (Q, D)
- Understanding Data, Research Data and Datasets
- Why data documentation (Q)

Part II: Data Documentation ABC

- Data Documentation: Study-level (E)
- Data Documentation: Data-level (Structured tabular data, Qualitative data) (E)

Part III: Dataset Metadata

- Dataset record examples, their associated standards, and data repositories (E, D)
- Data: DOIs and Data Citation
- Controlled Vocabularies and Thesauri (Q)
- Curation Tools for Datasets

Part IV: Thoughts and Services

- A Researcher's View vs. A Curator or Librarian's Perspective on Data Documentation (D)
- Dataset and Metadata Services at UCF

Q: w/ question. E: w/ examples. D: w/ discussion.

- Data
- Research data
- Dataset
- Data documentation
- Data types
- Data formats
- Project level
- File level
- Variable level
- Label
- Code
- Derived data
- Data list
- SPSS
- SAS
- R
- Access
- Spreadsheet
- Curation tool
- Metadata
- Metadata standards
- Metadata schemas
- Controlled vocabularies
- Thesauri
- Funding agencies
- Research data management
- DataCite
- DOI
- Data citation
- Data repository
- Dataset Metadata Service

** Word cloud generated using Tagxedo.*

Part I:
The Survey and Some Data Basics

Data Practices at UCF

- **The UCF Research Data Management (RDM) Survey**
  - The UCF Research Data Management Survey, November 2013
  - Results delivered on Research Computing Day at Institute for Simulation and Training by Dr. Penny Beile on February 11, 2014
    - http://www.ist.ucf.edu/hpc/rcd/Beile_datahandout.pdf

- **Data Recording and Analysis Section Questions and Results**
  - 17. Provide any technical details about the tools that you use or would like to be able to easily use for your work or research. These can be name or vendor of the software product, technical requirements of the software, special accelerators like graphical processor units (GPU), etc.

Before exposing the RDM survey results, let's see…

- Provide any technical details about the tools that you use or would like to be able to easily use for your work or research.

- If applicable, how are you recording lab data? Please check all that apply.
  - Lab notebooks in paper
  - Excel (or other) files on computers in the lab
  - Electronic lab notebook (ELN) tool. Please specify which one.

- Do you document or record any metadata for your data or dataset?
  - Yes
  - No

- If you record metadata for your dataset, do you use any local, agency-specific, or national standards or guidelines?
  - Yes
  - No
  - Not sure

UCF RDM Survey Results: Data Recording and Analysis Section

Thirty-nine (39) respondents listed a variety of technical tools used or needed to perform their research.

More popular tools:
SAS/SAS Enterprise version (6)
MatLab (5)
SPSS (5)
R-project programs (4)
NVivo (3)
SigmaPlot (3)
…

Processing, analysis, and writing software and databases	**Processing, backup, and storage network, server and cloud space**
AMOS	Automated backup internal to UCF system (2)
Ansys/Fluent (2)	Black Armor RAID backup system
ArcGIS/GIS ((2)	**Cloud storage/backup** (Dropbox and HIPAA-compliant cloudspace specifically mentioned) **(4)**
AspenTech	DSpace
CST Microwave Studio	Personal drives
Database with graphical viewing capabilities, basic statistics, filtering, custom output of datasets	Replication
DTreg	STOKES
EndNote	
FACTSAGE	
GPower	**Hardware**
Gephi	EPSON Workforce Pro GT-550 scanner
Git/GitHub (2)	Tablets
Interactive Data Language	
LimeSurvey	
Lumerical FDTD	
MathCad (Vensim) (2)	
MatLab (5)	
MS Office (2)	
NVivo (3)	
Origin	
RedCap	
REMARK'S OMR software	
R-project programs (4)	
SAS/SAS Enterprise version (6)	
SciFinder Scholar	
SigmaPlot (3)	
SPSS (5)	
SQL	
Stata (2)	
Video performance analysis software	

Source:
http://www.ist.ucf.edu/hpc/rcd/Beile_datahandout.pdf

UCF RDM Survey Results: Data Recording and Analysis Section

- 18. If applicable, how are you recording lab data? Please check all that apply.
  - The 49 respondents selected multiple answers, with **Excel** (or other) files on computers in the lab the most popular choice with 48 responses (98%). This was followed by **Lab notebooks in paper** (n=29, 59%) and **Electronic lab notebook tool** (n=3, 6%).

Lab notebooks in paper		29	59%
Excel (or other) files on computers in the lab		**48**	**98%**
Electronic lab notebook (ELN) tool. Please specify which one.		3	6%

  - If respondents indicated that they used an Electronic lab notebook they were asked to specify which one. The two ELNs identified were **Google Docs** and **Word** with embedded images storing NMR and other equipment data in a digital format.

Source:
http://www.ist.ucf.edu/hpc/rcd/Beile_datahandout.pdf

UCF RDM Survey Results: Data Recording and Analysis Section

- 19. Do you document or record any metadata for your data or dataset?
  - Of the 62 people who responded, 41 (66%) indicated that they do not add metadata to their datasets while 21 (34%) noted that they do. If respondents replied to the affirmative, they were asked about specific standards or guidelines. Those responses are reported in question 20.

Yes	21	34%
No	**41**	**66%**
Total	62	100%

Source:
http://www.ist.ucf.edu/hpc/rcd/Beile_datahandout.pdf

UCF RDM Survey Results: Data Recording and Analysis Section

- 20. If you record metadata for your dataset, do you use any local, agency-specific, or national standards or guidelines?
  - Twenty-one (21) respondents indicated that they assigned metadata to their data or dataset in question 19. Each of the respondents also answered the follow up question as to the type of standard or guideline applied. Of the responses, 15 (71%) do not use any specific standards or guidelines, five (24%) use identified standards, and one (5%) was not sure.

Yes (please specify)	5	24%
No	**15**	**71%**
I'm not sure	1	5%
Total	21	

  - The five who use standards or guidelines provided the following types: HIPAA/FERPA, FITS standard, program specific, librarians are helping us with this, and all of the above.

Source:
http://www.ist.ucf.edu/hpc/rcd/Beile_datahandout.pdf

Reflections from the UCF RDM Survey Results on Data Recording and Analysis

- After all, is data recording and documentation needed or important in your research lifecycle?
- What are the various ways to do data recording, documentation or analysis?
- Will you consider any standard for data documentation in your research process (e.g. local, agency-specific, or national standards or guidelines)? Is it necessary? What are these standards and where to find them?
- What are the typical tools out there that can help with data recording and analysis?

Refreshing the Concepts: What are data, research data and datasets?

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

  – National Research Council, 1992a. "Setting priorities for space research: Opportunities and imperatives."

- 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**.

  - Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, National Research Council. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases* (1999). Available at: http://www.nap.edu/openbook.php?record_id=9692&page=15

Research Data

- 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.**

– Organisation for Economic Co-operation and Development (OECD, 2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. P.13. Available at: http://www.oecd.org/science/sci-tech/38500813.pdf

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.

  - HLWIKI Canada (2011). http://hlwiki.slais.ubc.ca/index.php/Data_curation

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

  - Edinburgh University Data Library Research Data Management Handbook. http://www.docs.is.ed.ac.uk/docs/data-library/EUDL_RDM_Handbook.pdf

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.

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.

  - Organisation for Economic Co-operation and Development (OECD, 2007). http://www.oecd.org/science/sci-tech/38500813.pdf

Data Documentation

- "Data documentation explains how data were created or digitised, what data mean, what their content and structure are, and any manipulations that may have taken place." - UK Data Archive

- The term 'documentation' encompasses all the information necessary to interpret, understand and use a given dataset or set of documents. - Cambridge University Library

- "…a minimum requirement for closing the gap between the data producer and the secondary analyst is a high standard of data documentation." (note: the secondary analyst refers to the data user)
  - Nielsen, Per: How to teach data producers "the noble art" of data documentation. In: Clubb, Jerome M. (Ed.) ; Scheuch, Erwin K.(Ed.): Historical social research : the use of historical and process-produced data. Stuttgart : Klett-Cotta, 1980 (Historisch-Sozialwissenschaftliche Forschungen : quantitative sozialwissenschaftliche Analysen von historischen und prozeß-produzierten Daten 6). - ISBN 3-12-911060-7, pp. 477-487. URN: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-326298

Data Documentation and Metadata

- What is Metadata
  - Meta: Greek prefix. Means "after, behind or beyond." Data: Latin word. Factual information used for calculating, reasoning or measuring.
  - Metadata means something behind or beyond data itself, and it includes data about its content, containers and contextual information.
  - A formal definition: Metadata is data about data, data associated with an object, a document, or a dataset for purposes of description, administration, technical functionality and preservation.
- Can be embedded in the data files/documents themselves
- How is metadata relevant in the research data cycle? For example,

"Over the life course of a survey that results in a data set – 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."

- Arofan Gregory, Open Data Foundation (2011). The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes. Available at: http://odaf.org/papers/DDI_Intro_forNSIs.pdf

Data Documentation and Metadata (Cont.)

- Documentation and metadata are different things. However, metadata can be taken as a type of documentation.
- Documentation is meant to be read by humans; some metadata is designed more for machine processing than human readability.
- Research data can be documented at various levels: Project level, File or database level and Variable or item level.
- 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.

Why Data Documentation

- **Why data documentation?** (from Nielsen, Per: How to teach data producers "the noble art" of data documentation)
  - **Reliability aspect:** in hard sciences, research results are verified by repetition of the experiment; in social sciences, measuring unique phenomena, control of results and conclusions are possible only if data and full documentation are available
  - **Methodological aspect:** "we ask that all methodological considerations and decisions be reported at the time and place they are relevant"
  - **Economical aspect:** it can be "cheaper to clean and document data files for general use before the primary analysis is started," "reports on new issues can be based on existing well-documented files"
  - **Historical aspect:** archive and preserve information for future generations
- Additional aspect: to meet funder requirements

Data Definitions and Specific Requirements from the Funding Agency: National Science Foundation (NSF)

- 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.).

Data Definitions and Specific Requirements from the Funding Agency: National Science Foundation (NSF) (Cont.)

- 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;
  - 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.
- **Search Data Management Plan requirements of different funders at DMPTool** (https://dmptool.org/guidance).

Part II:
Data Documentation ABC

Data Documentation ABC

- Ensure that all data collected and generated through your research lifecycle is documented.

- At the beginning of your research, check what kind of documentation is available or necessary, and identify needed documentations which will enable data preservation and reuse in the future.
  - The various kinds of documentation may include:
    - **Embedded documentation** (included within the data, e.g., code, field and label descriptions, descriptive headers or summaries, transcripts, in document properties)
    - **Supporting documentation** (in separate file, e.g., working papers, lab books, questionnaires or interview guides, project reports, publications)
    - **Catalog Metadata** (for data archiving, identification and locating)

Data Documentation ABC (Cont.)

- **The different types of documentations may include:**
  - Laboratory notebooks & experimental protocols
  - Questionnaires, code books with full variable and value labels & data dictionaries
  - Information about equipment settings & instrument calibration
  - Software syntax & output files
  - Database schema
  - Methodology reports
  - Assumptions made during analysis
  - Provenance information about sources of derived data, different versions of the dataset

Data Documentation ABC (Cont.)

- During your research, document all **research data formats** utilized by your project. Research data comes in many varied formats, such as (by broad 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).

Research Data Types and Formats (in detail)

Type of data	Acceptable formats for sharing, reuse and preservation	Other acceptable formats for data preservation
Quantitative tabular data with extensive metadata a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information some structured text or mark-up file containing metadata information, e.g. DDI XML file	proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta) MS Access (.mdb/.accdb)
Quantitative tabular data with minimal metadata a matrix of data with or without column headings or variable names, but no other metadata or labelling	comma-separated values (CSV) file (.csv) tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate	delimited text of given character set - only characters not present in the data should be used as delimiters (.txt) widely-used formats, e.g. MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf) and OpenDocument Spreadsheet (.ods)
Geospatial data vector and raster data	ESRI Shapefile (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data	ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mif) for vector data Keyhole Mark-up Language (KML) (.kml) Adobe Illustrator (.ai), CAD data (.dxf or .svg) binary formats of GIS and CAD packages
Qualitative data textual	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml) Rich Text Format (.rtf) plain text data, ASCII (.txt)	Hypertext Mark-up Language (HTML) (.html) widely-used proprietary formats, e.g. MS Word (.doc/.docx) some proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti

Research Data Types and Formats (in detail, cont.)

Type of data	Acceptable formats for sharing, reuse and preservation	Other acceptable formats for data preservation
Digital image data	TIFF version 6 uncompressed **(.tif)**	JPEG (.jpeg, .jpg) but only if created in this format TIFF (other versions) **(.tif, .tiff)** Adobe Portable Document Format **(PDF/A, PDF) (.pdf)** standard applicable RAW image format **(.raw)** Photoshop files **(.psd)**
Digital audio data	Free Lossless Audio Codec (FLAC) **(.flac)**	MPEG-1 Audio Layer 3 (.mp3) but only if created in this format Audio Interchange File Format (AIFF) **(.aif)** Waveform Audio Format (WAV) **(.wav)**
Digital video data	MPEG-4 **(.mp4)** motion JPEG 2000 **(.mj2)**	
Documentation and scripts	Rich Text Format **(.rtf)** PDF/A or PDF **(.pdf)** HTML **(.htm)** OpenDocument Text **(.odt)**	plain text **(.txt)** some widely-used proprietary formats, e.g. MS Word **(.doc/.docx)** or MS Excel **(.xls/.xlsx)** XML marked-up text **(.xml)** according to an appropriate DTD or schema, e.g. XHMTL 1.0

Source: http://www.data-archive.ac.uk/create-manage/format/formats-table

Data Documentation ABC (Cont.)

- **Keep the wide variety of materials that are generated or collected in your research. Research data** (traditional and electronic research) may include all of the following:
  - 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

Source: How to manage research data, Research Support Services, University of Edinburgh Information Services

Data Documentation ABC (Cont.)

- **Document research data at different levels:**
  - **Study-level**
  - **Data-level**
    - Structured tabular data
    - Qualitative data

- Utilize software to create embedded documentation for the data (if applicable), and make separate supporting documentation (e.g. readme text files) to describe the list of files and documentations in a folder.

- In addition, provide unique identifier for the dataset (e.g. doi, purl, handle…);

- Further, make sure that your data meets citation requirement (if applicable), and discuss with relevant personnel on how data can be archived and shared in a data center, or a library digital repository for others to search, locate and reuse.

Data Documentation: Study-level

- Information in the Data Documentation Study-level and Data-level section is from UK Data Archive (http://www.data-archive.ac.uk/create-manage/document)

- **Study-level information:** the research context and design, data collection methods, data preparation and results or findings
  - the context of data collection: project history, aims, objectives and hypotheses
  - data collection methods: data collection protocols, sampling design, instruments used, hardware and software used, data scale and resolution, temporal coverage and geographic coverage, and digitization or transcription methods
  - structure of data files, number of cases, records, variables and relationships between files
  - data sources used and provenance of materials, e.g. for transcribed or derived data
  - data validation, checking, proofing, cleaning and other quality assurance procedures carried out, such as checking for equipment and transcription errors, calibration procedures, data capture resolution and repetitions, or editing, proofing or quality control of materials
  - modifications made to data over time since their original creation and identification of different versions of datasets
  - for time series or longitudinal surveys, changes made to methodology, variable content, question text, variable labelling, measurements or sampling
  - information on data confidentiality, access and use conditions, where applicable

Data Documentation: Data-level

- Descriptions and annotations at the variable, data item or data file level.
  - names, labels and descriptions for variables, records and their values
  - explanation of codes and classification schemes used
  - codes of, and reasons for, missing values
  - derived data created after collection, with code, algorithm or command file used to create them
  - weighting and grossing variables created and how they should be used
  - data list describing cases, individuals or items studied, for example for logging qualitative interviews

Data Documentation: Structured tabular data

- Structured, tabular data should have **cases or records and variables** adequately documented with:
  - **Names, labels and descriptions** for all variables, fields, records and their values. **Variable labels** should:
    - be brief with a maximum of 80 characters
    - indicate the unit of measurement, where applicable
    - reference the question number of a survey or questionnaire, where applicable

How to name the variable to document the survey result for "Q11: hours spent taking physical exercise in a typical week"?
For example: q11hexw

Data Documentation: Structured tabular data

- Code labels

  How to name the variable for female respondents?

  For example: p1sex (with codes '1=female ', '2=male', '-8=don't know', '-9=not answered‘)

- Coding or classification schemes used, ideally with a bibliographic reference

  Where to find a list of codes to classify respondents' jobs?

  Reference: Standard Occupational Classification 2000

  Where to get the country codes?

  Reference: ISO 3166 alpha-2 country codes

- Codes of, and reasons for, missing data

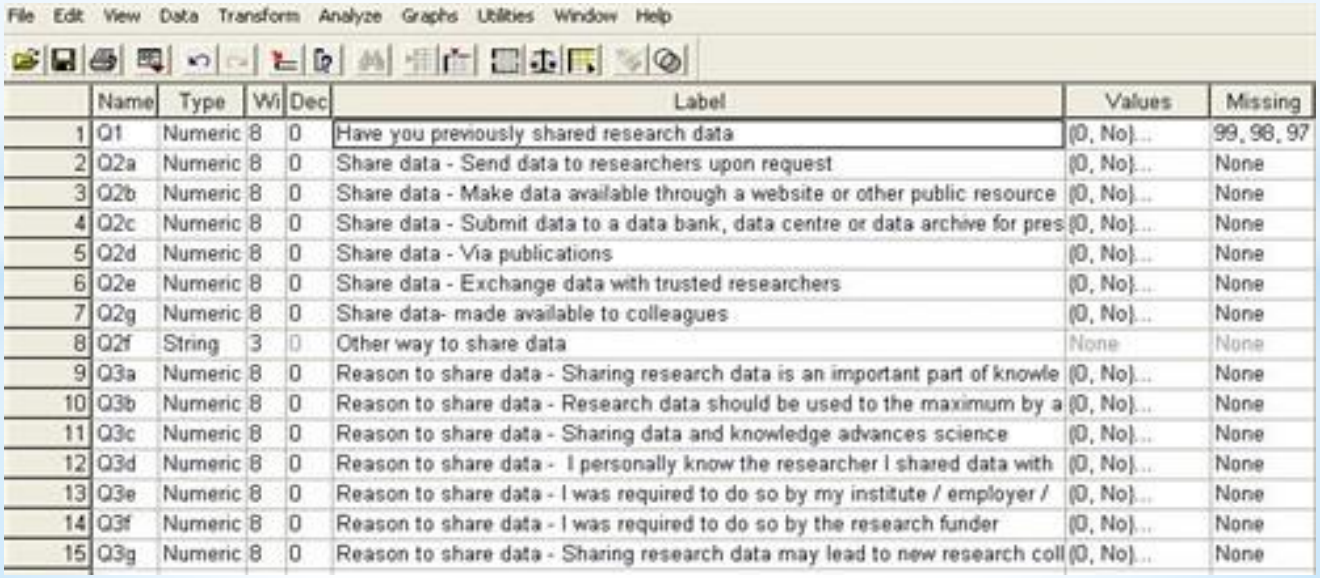
  How to document missing data?

  For example: '99=not recorded', '98=not provided (no answer)', '97=not applicable', '96=not known', '95=error'

Source: http://ukdataservice.ac.uk/manage-data/document/data-level.aspx

Data Documentation: Structured tabular data (cont.)

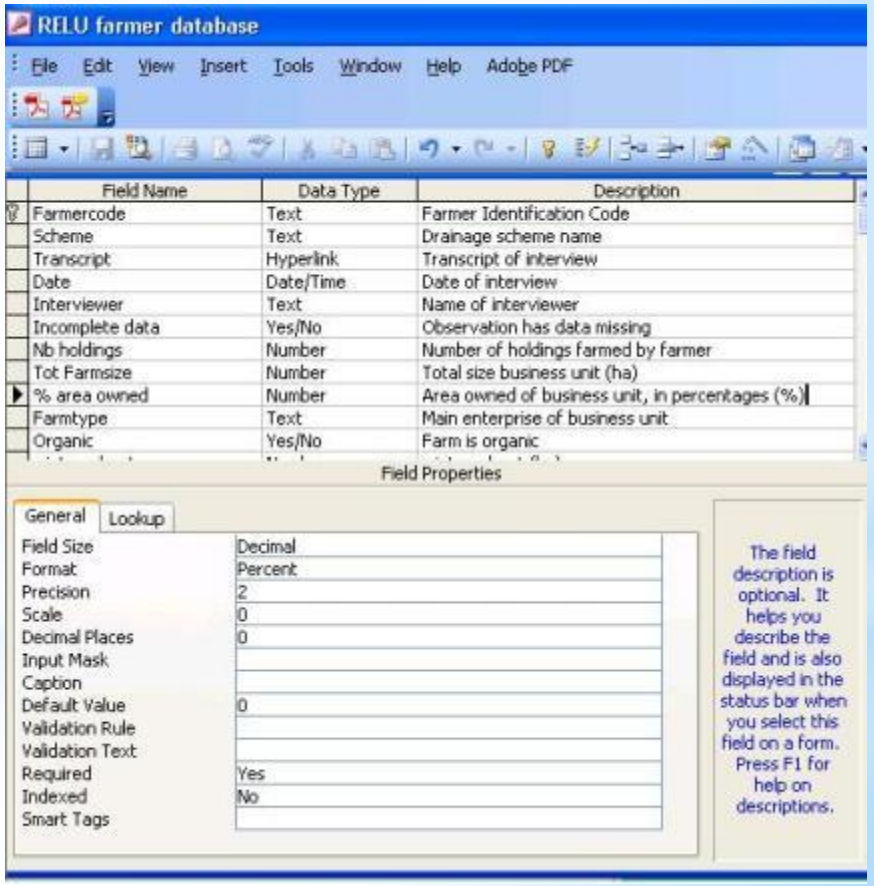
- Data-level descriptions can be **embedded within a data file**
  - **Statistical e.g. SPSS**
    - variable descriptions and attributes (codes, data type, missing values) of each variable in the data file can be documented in 'Variable View' or via syntax, whereby embedded data documentation is then contained in the SPSS command file



	Name	Type	Wi	Dec	Label	Values	Missing
1	Q1	Numeric	8	0	Have you previously shared research data	{0, No}...	99, 98, 97
2	Q2a	Numeric	8	0	Share data - Send data to researchers upon request	{0, No}...	None
3	Q2b	Numeric	8	0	Share data - Make data available through a website or other public resource	{0, No}...	None
4	Q2c	Numeric	8	0	Share data - Submit data to a data bank, data centre or data archive for pres	{0, No}...	None
5	Q2d	Numeric	8	0	Share data - Via publications	{0, No}...	None
6	Q2e	Numeric	8	0	Share data - Exchange data with trusted researchers	{0, No}...	None
7	Q2g	Numeric	8	0	Share data- made available to colleagues	{0, No}...	None
8	Q2f	String	3	0	Other way to share data	None	None
9	Q3a	Numeric	8	0	Reason to share data - Sharing research data is an important part of knowle	{0, No}...	None
10	Q3b	Numeric	8	0	Reason to share data - Research data should be used to the maximum by a	{0, No}...	None
11	Q3c	Numeric	8	0	Reason to share data - Sharing data and knowledge advances science	{0, No}...	None
12	Q3d	Numeric	8	0	Reason to share data - I personally know the researcher I shared data with	{0, No}...	None
13	Q3e	Numeric	8	0	Reason to share data - I was required to do so by my institute / employer /	{0, No}...	None
14	Q3f	Numeric	8	0	Reason to share data - I was required to do so by the research funder	{0, No}...	None
15	Q3g	Numeric	8	0	Reason to share data - Sharing research data may lead to new research coll	{0, No}...	None

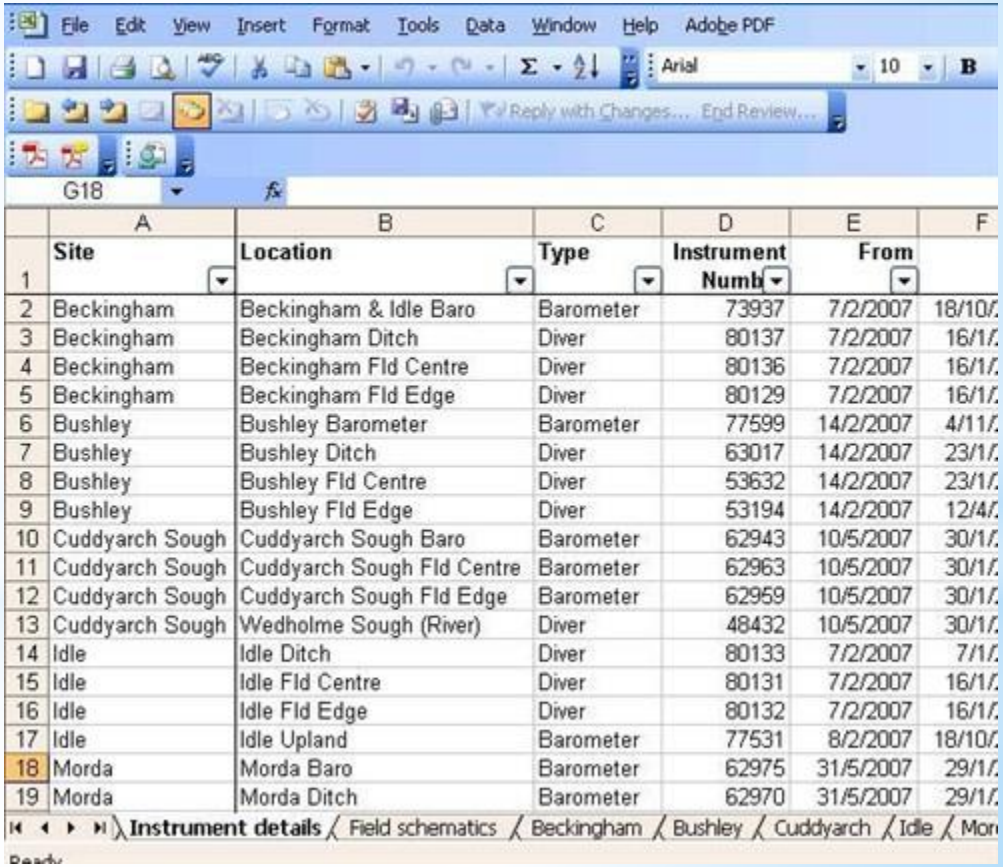
Data Documentation: Structured tabular data (cont.)

- Data-level descriptions can be **embedded within a data file**
  - **Databases e.g. MS Access**
    - variable descriptions and attributes can be documented in 'Design View' and relationships between tables and files can be created



Data Documentation: Structured tabular data (cont.)

- Data-level descriptions can be **embedded within a data file**
  - **Spreadsheets e.g. MS Excel**
    - an additional worksheet within the data file can contain data-related documentation



	A	B	C	D	E	F
1	Site	Location	Type	Instrument Numb	From	
2	Beckingham	Beckingham & Idle Baro	Barometer	73937	7/2/2007	18/10/
3	Beckingham	Beckingham Ditch	Diver	80137	7/2/2007	16/1/
4	Beckingham	Beckingham Fld Centre	Diver	80136	7/2/2007	16/1/
5	Beckingham	Beckingham Fld Edge	Diver	80129	7/2/2007	16/1/
6	Bushley	Bushley Barometer	Barometer	77599	14/2/2007	4/11/
7	Bushley	Bushley Ditch	Diver	63017	14/2/2007	23/1/
8	Bushley	Bushley Fld Centre	Diver	53632	14/2/2007	23/1/
9	Bushley	Bushley Fld Edge	Diver	53194	14/2/2007	12/4/
10	Cuddyarch Sough	Cuddyarch Sough Baro	Barometer	62943	10/5/2007	30/1/
11	Cuddyarch Sough	Cuddyarch Sough Fld Centre	Barometer	62963	10/5/2007	30/1/
12	Cuddyarch Sough	Cuddyarch Sough Fld Edge	Barometer	62959	10/5/2007	30/1/
13	Cuddyarch Sough	Wedholme Sough (River)	Diver	48432	10/5/2007	30/1/
14	Idle	Idle Ditch	Diver	80133	7/2/2007	7/1/
15	Idle	Idle Fld Centre	Diver	80131	7/2/2007	16/1/
16	Idle	Idle Fld Edge	Diver	80132	7/2/2007	16/1/
17	Idle	Idle Upland	Barometer	77531	8/2/2007	18/10/
18	Morda	Morda Baro	Barometer	62975	31/5/2007	29/1/
19	Morda	Morda Ditch	Barometer	62970	31/5/2007	29/1/

Data Documentation: Structured tabular data (cont.)

- Data-level descriptions can be **embedded within a data file**
  - **GIS e.g ArcGIS**
    - shapefiles (layers) and tables can be organised in a geo-database with rich metadata created in ArcCatalog
- A dataset may also be accompanied with a **Codebook** detailing all variables and their values
- Variable naming
  - Full variable name
  - meaningful abbreviations (e.g. oz%=percentage ozone; moocc=mother occupation)
  - question number system (Q1a, Q1b, Q2, Q3a,...)
  - numerical order system (V1, V2, V3,...)

Source:
http://ukdataservice.ac.uk/manage-data/document/data-level.aspx

Structured Metadata: XML Schemas

- XML schema brings documentation into a single document; creates structured content about the data and allows data interoperability and sharing.
- It can document comprehensive variable level information, such as basic data dictionary, question text and question routing instructions.
- **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/
- Projects using the DDI (http://www.ddialliance.org/ddi-at-work/projects)
- DDI-compliant data repository:
  - ICPSR - Inter-university Consortium for Political and Social Research
    - Data deposit form: https://www.icpsr.umich.edu/cgi-bin/ddf2
    - UCF is a member of ICPSR.
  - UKDA - UK Data Archive

Social Science Dataset Display

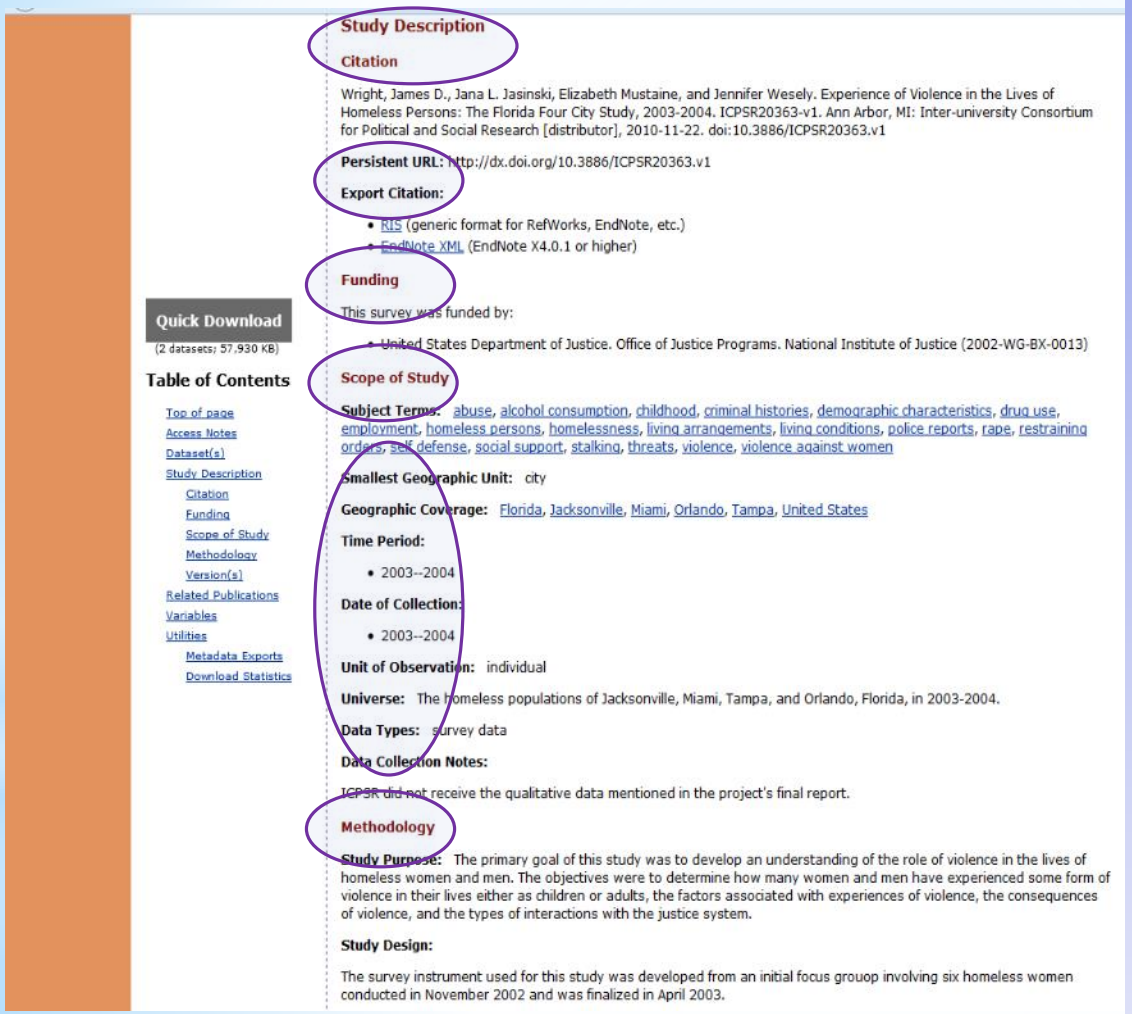
Field Labels:
Title
Principal investigator(s)
Summary
Access notes
Dataset(s)

ICPSR:
Interuniversity Consortium for Political and Social Research.

Dataset(s)
DSO: Study-Level Files
Documentation:
Questionnaire.pdf
User guide.pdf
DS1: Female Interviews
Documentation:
Codebook.pdf
…

http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/20363?archive=NACJD&q=%22university+of+central+florida%22&permit%5B0%5D=AVAILABLE&x=-999&y=-84

Social Science Dataset Display (Cont.)



Quick Download
(2 datasets; 57.930 KB)

Table of Contents

- Top of page
- Access Notes
- Dataset(s)
- Study Description
  - Citation
  - Funding
  - Scope of Study
  - Methodology
  - Version(s)
- Related Publications
- Variables
- Utilities
  - Metadata Exports
  - Download Statistics

Study Description

Citation

Wright, James D., Jana L. Jasinski, Elizabeth Mustaine, and Jennifer Wesely. Experience of Violence in the Lives of Homeless Persons: The Florida Four City Study, 2003-2004. ICPSR20363-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-11-22. doi:10.3886/ICPSR20363.v1

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

Export Citation:

- RIS (generic format for RefWorks, EndNote, etc.)
- EndNote XML (EndNote X4.0.1 or higher)

Funding

This survey was funded by:

- United States Department of Justice. Office of Justice Programs. National Institute of Justice (2002-WG-BX-0013)

Scope of Study

Subject Terms: abuse, alcohol consumption, childhood, criminal histories, demographic characteristics, drug use, employment, homeless persons, homelessness, living arrangements, living conditions, police reports, rape, restraining orders, self defense, social support, stalking, threats, violence, violence against women

Smallest Geographic Unit: city

Geographic Coverage: Florida, Jacksonville, Miami, Orlando, Tampa, United States

Time Period:

- 2003--2004

Date of Collection:

- 2003--2004

Unit of Observation: individual

Universe: The homeless populations of Jacksonville, Miami, Tampa, and Orlando, Florida, in 2003-2004.

Data Types: survey data

Data Collection Notes:

ICPSR did not receive the qualitative data mentioned in the project's final report.

Methodology

Study Purpose: The primary goal of this study was to develop an understanding of the role of violence in the lives of homeless women and men. The objectives were to determine how many women and men have experienced some form of violence in their lives either as children or adults, the factors associated with experiences of violence, the consequences of violence, and the types of interactions with the justice system.

Study Design:

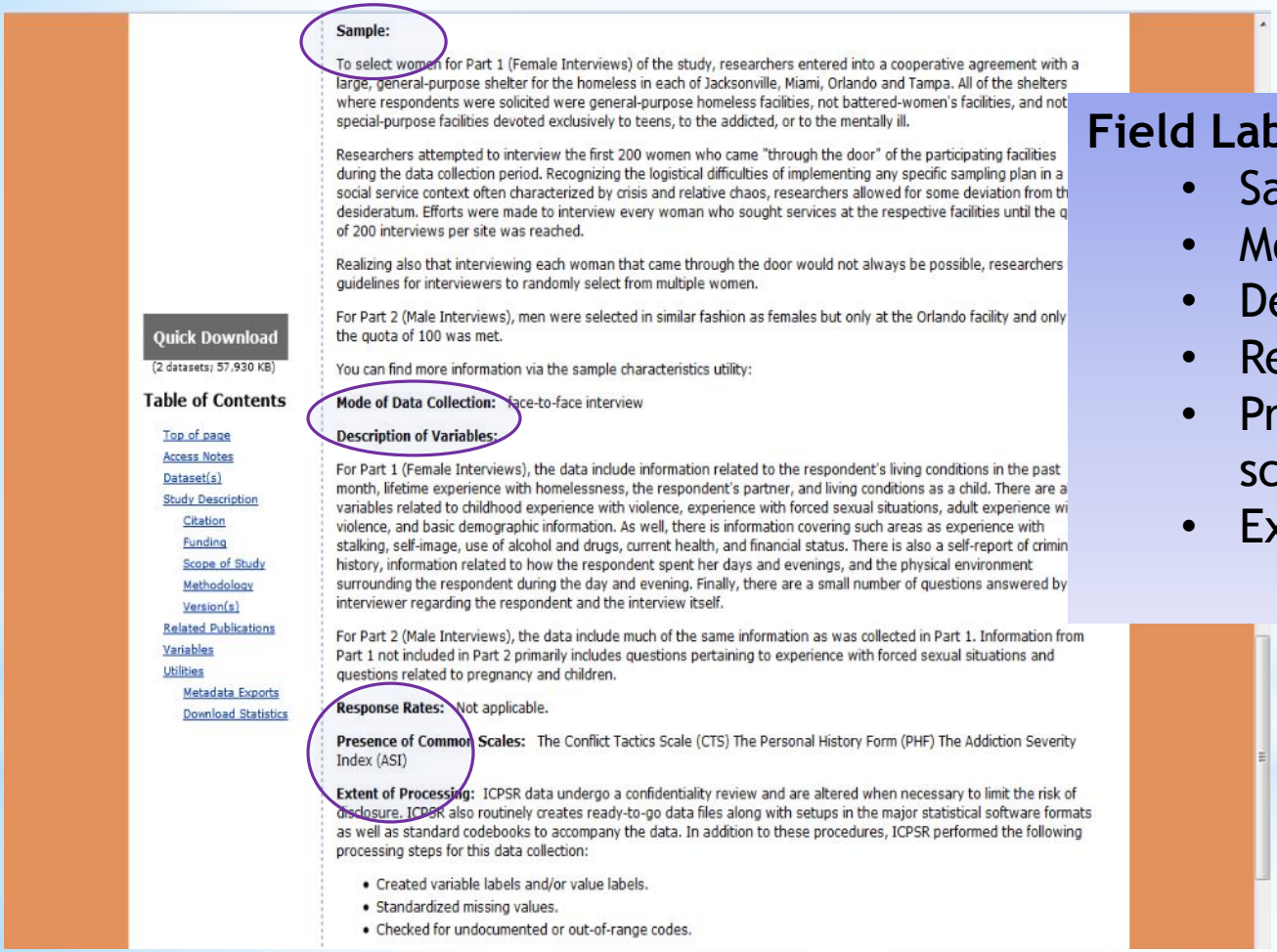
The survey instrument used for this study was developed from an initial focus grouop involving six homeless women conducted in November 2002 and was finalized in April 2003.

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

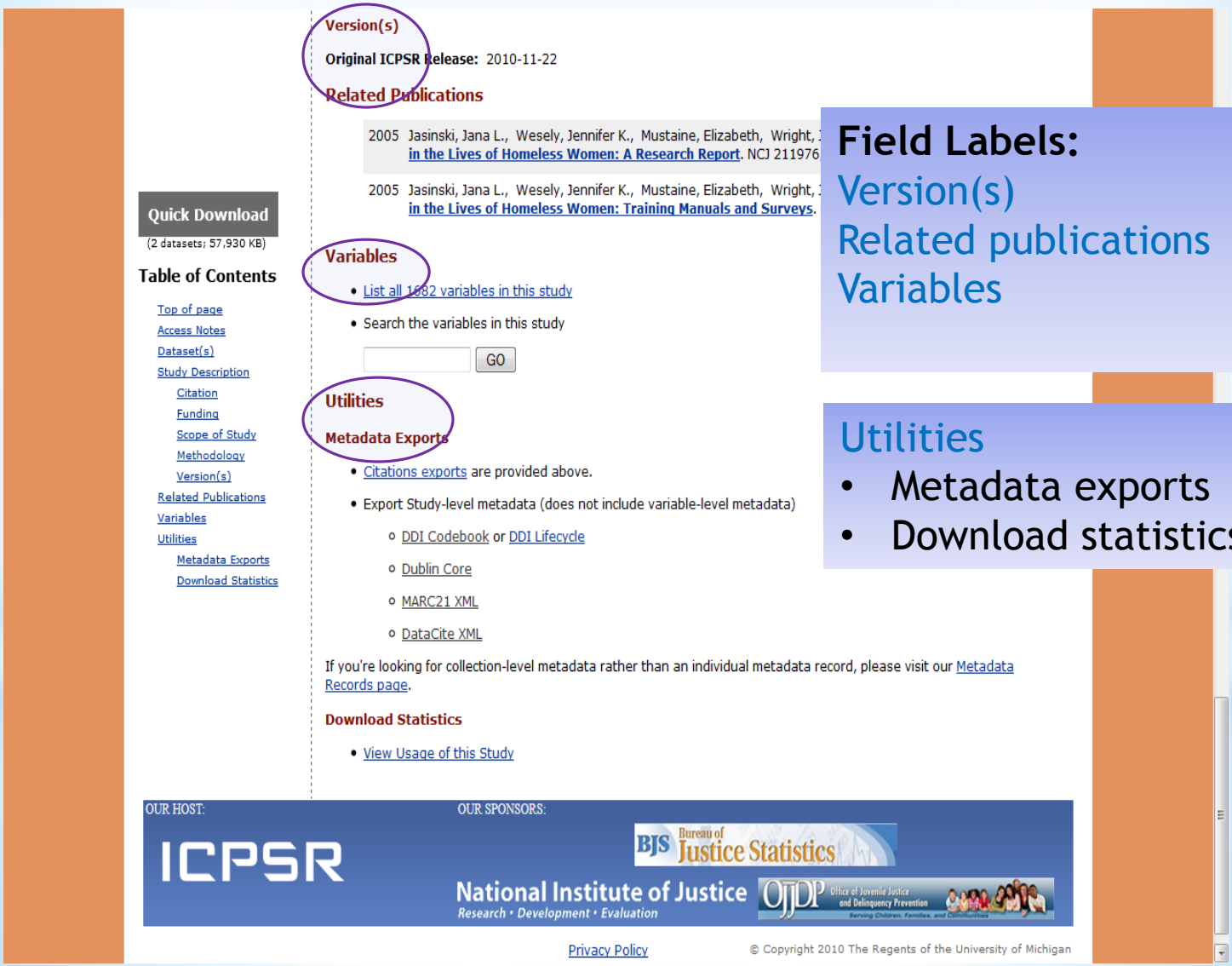
Social Science Dataset Display (Cont.)



Field Labels:

- Sample
- Mode of data collection
- Description of variables
- Response rates
- Presence of common scales
- Extent of processing

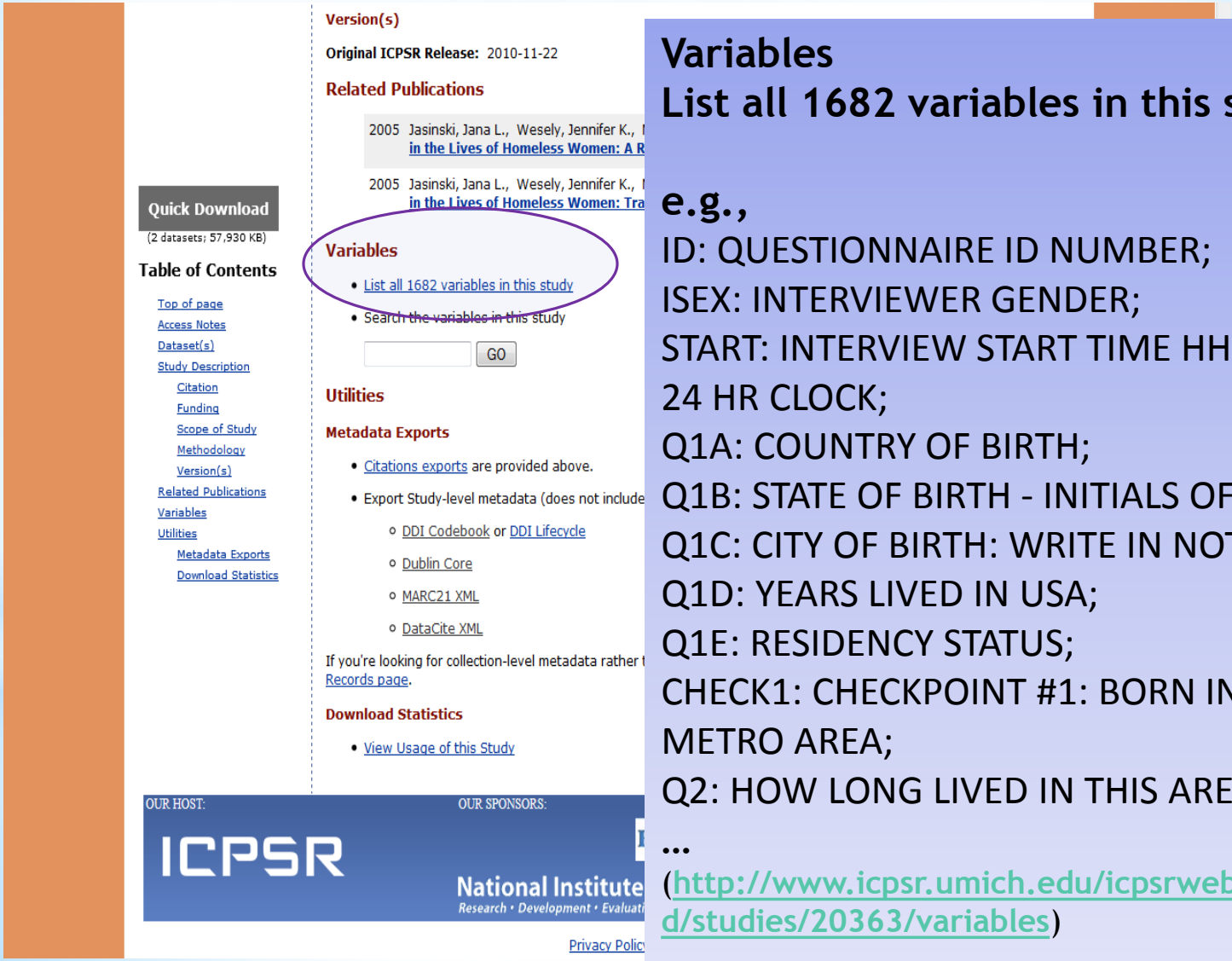
Social Science Dataset Display (Cont.)



Field Labels:
Version(s)
Related publications
Variables

Utilities
- Metadata exports
- Download statistics

Social Science Dataset Display (Cont.)



Variables
List all 1682 variables in this study

e.g.,
ID: QUESTIONNAIRE ID NUMBER;
ISEX: INTERVIEWER GENDER;
START: INTERVIEW START TIME HH:MM USE 24 HR CLOCK;
Q1A: COUNTRY OF BIRTH;
Q1B: STATE OF BIRTH - INITIALS OF STATE;
Q1C: CITY OF BIRTH: WRITE IN NOT APP;
Q1D: YEARS LIVED IN USA;
Q1E: RESIDENCY STATUS;
CHECK1: CHECKPOINT #1: BORN IN SAME METRO AREA;
Q2: HOW LONG LIVED IN THIS AREA?
…
(http://www.icpsr.umich.edu/icpsrweb/NACJD/ssvd/studies/20363/variables)

Social Science Dataset in DDI Format

docDscr

The **Document Description** consists of **bibliographic information** describing the DDI-compliant document itself as a whole.

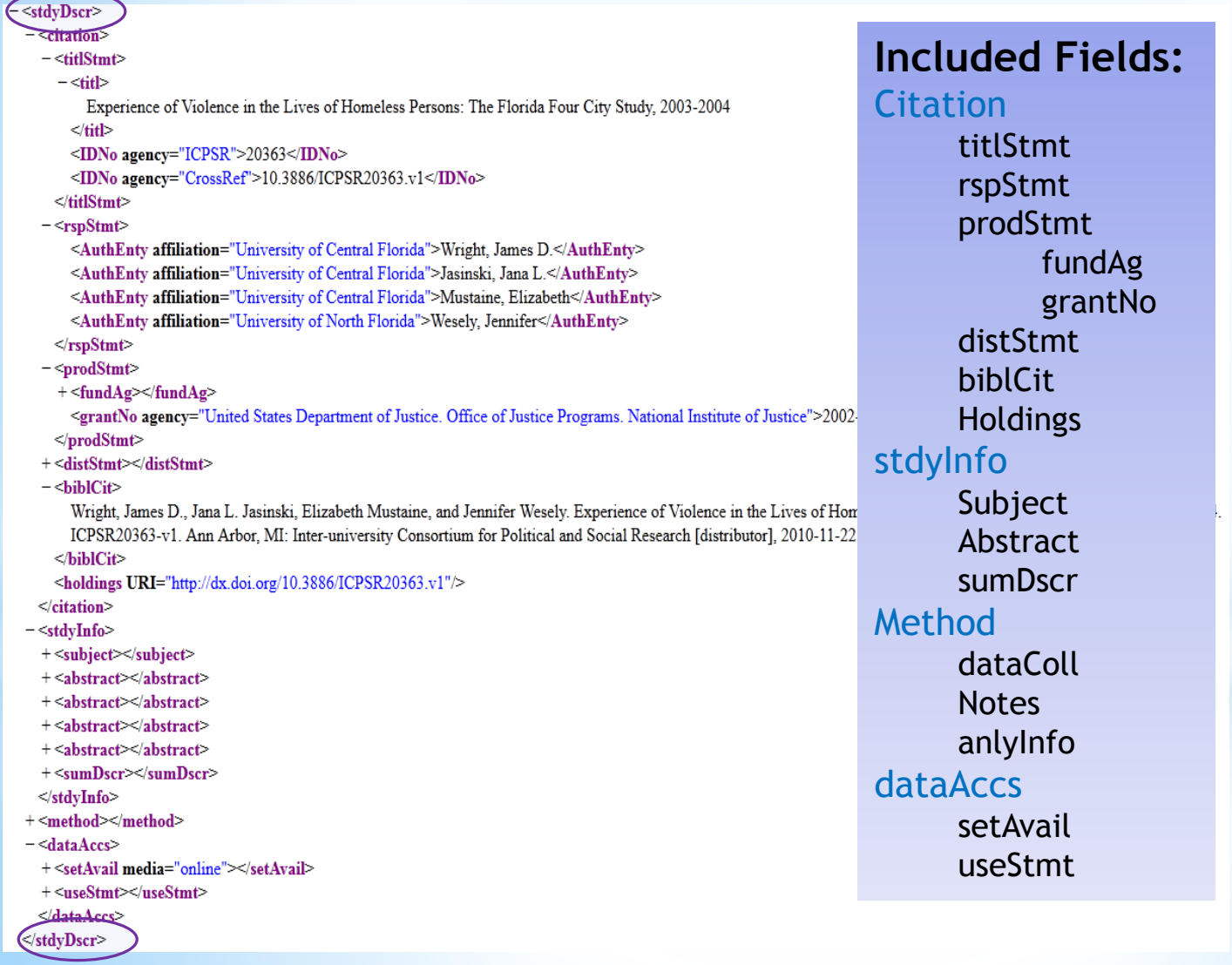
Included Fields:

citation

- titleStmt
- prodStmt
- verStmt
- holdings

http://www.icpsr.umich.edu/icpsrweb/ICPSR/ddi2/studies/20363

Social Science Dataset in DDI Format (Cont.)



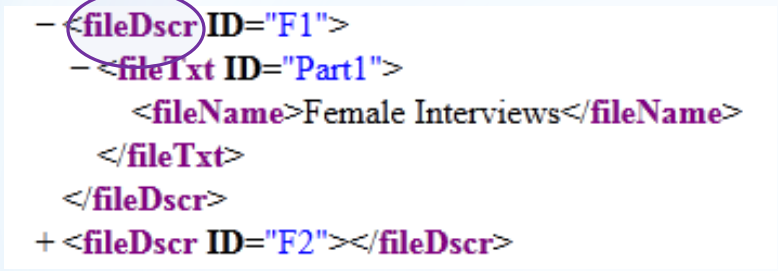
```
- <stdyDscr>
  - <citation>
    - <titlStmt>
      - <titl>
          Experience of Violence in the Lives of Homeless Persons: The Florida Four City Study, 2003-2004
        </titl>
        <IDNo agency="ICPSR">20363</IDNo>
        <IDNo agency="CrossRef">10.3886/ICPSR20363.v1</IDNo>
      </titlStmt>
    - <rspStmt>
        <AuthEnty affiliation="University of Central Florida">Wright, James D.</AuthEnty>
        <AuthEnty affiliation="University of Central Florida">Jasinski, Jana L.</AuthEnty>
        <AuthEnty affiliation="University of Central Florida">Mustaine, Elizabeth</AuthEnty>
        <AuthEnty affiliation="University of North Florida">Wesely, Jennifer</AuthEnty>
      </rspStmt>
    - <prodStmt>
      + <fundAg></fundAg>
        <grantNo agency="United States Department of Justice. Office of Justice Programs. National Institute of Justice">2002-
      </prodStmt>
    + <distStmt></distStmt>
    - <biblCit>
        Wright, James D., Jana L. Jasinski, Elizabeth Mustaine, and Jennifer Wesely. Experience of Violence in the Lives of Hom
        ICPSR20363-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-11-22
      </biblCit>
      <holdings URI="http://dx.doi.org/10.3886/ICPSR20363.v1"/>
    </citation>
  - <stdyInfo>
    + <subject></subject>
    + <abstract></abstract>
    + <abstract></abstract>
    + <abstract></abstract>
    + <abstract></abstract>
    + <sumDscr></sumDscr>
    </stdyInfo>
  + <method></method>
  - <dataAccs>
    + <setAvail media="online"></setAvail>
    + <useStmt></useStmt>
    </dataAccs>
</stdyDscr>
```

Included Fields:

- Citation
  - titlStmt
  - rspStmt
  - prodStmt
    - fundAg
    - grantNo
  - distStmt
  - biblCit
  - Holdings
- stdyInfo
  - Subject
  - Abstract
  - sumDscr
- Method
  - dataColl
  - Notes
  - anlyInfo
- dataAccs
  - setAvail
  - useStmt

stdyDscr 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.

Data Documentation: Qualitative Data

- **Context and participant details of interviews can be:**
  - **A descriptive header or summary page** in transcripts or field notes.
  - A structured **data list**
  - **XML mark-up of data,** for example,
    - **Text Encoding Initiative (TEI)** to mark up interview transcript
    - **Qualitative Data Exchange Format (QuDEx)** for researcher annotations and data linking

Data Documentation: Qualitative Data (Cont.)

- **Anonymisation of textual data** (e.g., replacing real names of people, organizations and locations with pseudonyms)

- **File naming**
  - Meaningful, short names; identify file types (e.g. interviews, focus groups, field notes, audio recordings); avoid space, special characters; avoid long names

- **Organizing files in folders:** Create uniform and structured folder names based on cases, studies, locations, data types etc., or the original, anonymized, coded or annotated versions of data

- **Version control:** Version numbering in file names

- **Documentation:** Methodology description, project plan, interview guidelines, consent form templates; data analyses and manipulation

Data Documentation: Qualitative Data Example

Morbidity and mortality - Population, vital statistics and censuses
Child development and child rearing - Social stratification and groupings
Elderly - Social stratification and groupings
Family life and marriage - Social stratification and groupings
Gender roles - Social stratification and groupings
Use and provision of specific social services - Social welfare policy and systems

Depositor(s):
Blaxter, M., University of East Anglia

Principal Investigator(s):
Blaxter, M., University of East Anglia

Data Collector(s):
Blaxter, M., University of East Anglia

Sponsor(s):
Economic and Social Research Council

Abstract:
This is an enhanced qualitative study.

The research looked at beliefs and attitudes to **health** and medical care, inter-genera and social history of members of a grandmother generation. The original study includ daughters as well; this collection contains only the grandmother interviews.

Grandmothers are asked extensive questions about their own **health** and the **heal** members. Details are provided on episodes of illness and remedies used, both home a Specific topics of accidents, nutrition, dental care, and immunisation are covered.

More generally, grandmothers are asked about their views of their personal docto **health** services. They give opinions on the quality of **health** care before and after the National **Health** Service.

Grandmother-daughter relationships are explored, especially around the subject of off medical advice concerning care for the grandchildren.

The collection has been enhanced by: conversion from paper to searchable RTF fo extensive editing and formatting of all interview transcripts. This collection will also through ESDS Qualidata Online.

Morbidity and mortality - Population, vital statistics and censuses
Child development and child rearing - Social stratification and groupings
Elderly - Social stratification and groupings
Family life and marriage - Social stratification and groupings
Gender roles - Social stratification and groupings
Use and provision of specific social services - Social welfare policy and systems

Depositor(s):
Blaxter, M., University of East Anglia

Principal Investigator(s):
Blaxter, M., University of East Anglia

Data Collector(s):
Blaxter, M., University of East Anglia

Sponsor(s):
Economic and Social Research Council

Abstract:
This is an enhanced qualitative study.

The research looked at beliefs and attitudes to **health** and medical care, inter-generational relationships, and social history of members of a grandmother generation. The original study included interviews with daughters as well; this collection contains only the grandmother interviews.

Grandmothers are asked extensive questions about their own **health** and the **health** of other family members. Details are provided on episodes of illness and remedies used, both home and **health** services. Specific topics of accidents, nutrition, dental care, and immunisation are covered.

More generally, grandmothers are asked about their views of their personal doctors and institutional **health** services. They give opinions on the quality of **health** care before and after the introduction of the National **Health** Service.

Grandmother-daughter relationships are explored, especially around the subject of offering and taking of medical advice concerning care for the grandchildren.

The collection has been enhanced by: conversion from paper to searchable RTF format by OCR and extensive editing and formatting of all interview transcripts. This collection will also be made available through ESDS Qualidata Online.

- Example is from: A NESSTAR FOR QUALITATIVE DATA: BUILDING BLOCKS FOR DIGITAL FUTURES. By Corti Louise et al. available at: http://data-archive.ac.uk/media/376907/digitalfutures_dashish_21nov2012.pdf

Data Documentation: Qualitative Data Example

- Data List

Study Number 6124
Being a Doctor: a Sociological Analysis, 2005-2006
Nettleton, S

Interview ID	Gender	Description	Date of Interview	No of Pages	Text File Name
x001	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	07/09/2005	36	6124int001
x002	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	09/09/2005	41	6124int002
x003	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	09/09/2005	39	6124int003
x004	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	13/09/2005	36	6124int004
x005	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	13/09/2005	34	6124int005
x006	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	14/09/2005	50	6124int006
x007	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	21/09/2005	31	6124int007
x008	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	21/09/2005	35	6
x009	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	22/09/2005	33	6
x010	Female	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	22/09/2005	23	6
x011	Male	Interview with a Hospital Doctor in a Multi-Ethnic Northern City	22/09/2005	36	6
		w with a Hospital Doctor in a Multi-Ethnic Northern City	27/09/2005	41	6
		w with a Hospital Doctor in a Multi-Ethnic Northern City	27/09/2005	21	6
		w with a Hospital Doctor in a Multi-Ethnic Northern City	30/09/2005	20	6
		w with a Hospital Doctor in a Multi-Ethnic Northern City	05/10/2005	19	6
		w with a Hospital Doctor in a Multi-Ethnic Northern City	05/10/2005	27	6
		w with a Hospital Doctor in a Multi-Ethnic Northern City	07/10/2005	27	6
		w with a Hospital Doctor in a Multi-Ethnic Northern City	17/10/2005	11	6
		w with a Hospital Doctor in a Multi-Ethnic Northern City	19/10/2005	33	6124int019
		w with a Hospital Doctor in a Multi-Ethnic Northern City	07/11/2005	21	6124int020
		w with Hospital Doctor in Northern Tourist City	19/07/2005	50	6124int021
		w with Hospital Doctor in Northern Tourist City	10/08/2005	45	6124int022
z003	Male	Interview with Hospital Doctor in Northern Tourist City	17/08/2005	56	6124int023
z004	Male	Interview with Hospital Doctor in Northern Tourist City	07/11/2005	21	6124int024
z005	Female	Interview with Hospital Doctor in Northern Tourist City	14/11/2005	32	6124int025
z006	Male	Interview with Hospital Doctor in Northern Tourist City	15/11/2005	23	6124int026
z007	Male	Interview with Hospital Doctor in Northern Tourist City	16/11/2005	23	6124int027
z008	Female	Interview with Hospital Doctor in Northern Tourist City	17/11/2005	18	6124int028
z009	Male	Interview with Hospital Doctor in Northern Tourist City	18/11/2005	20	6124int029

Interview ID:
x001
x002
…

Text File Name:
6124int001
6124int002
…

Part III:
Dataset Metadata

Dataset Metadata

- Create and generate metadata for your research data and datasets in your research lifecycle to preserve the data in the long run.
  - Consider what information is needed for the data to be read and interpreted in the future.
  - Understand your funder requirements for data documentation and metadata. Funder requirements for NSF, GBMF, IMLS, NEH, NIH and NOAA can be found at https://dmptool.org/guidance.
  - Consult available metadata standards in your field. You may refer to Common Metadata Standards and Domain Specific Metadata Standards for details.

Dataset Metadata (cont.)

- Describe data and datasets created in your research lifecycle, and use software programs and tools to assist in data documentation. Assign or capture administrative, descriptive, technical, structural and preservation metadata for the data. Some potential information to document:

- **Descriptive metadata**
  - Name of creator of data set
  - Name of author of document
  - Title of document
  - File name
  - Location of file
  - Size of file

- **Structural metadata**
  - File relationships (e.g. child, parent)

Dataset Metadata (cont.)

- **Technical metadata**
  - Format (e.g. text, SPSS, Stata, Excel, tiff, mpeg, 3D, Java, FITS, CIF)
  - Compression or encoding algorithms
  - Encryption and decryption keys
  - Software (including release number) used to create or update the data
  - Hardware on which the data were created
  - Operating systems in which the data were created
  - Application software in which the data were created

- **Administrative metadata**
  - Information about data creation (e.g. date)
  - Information about subsequent updates, transformation, versioning, summarization
  - Descriptions of migration and replication
  - Information about other events that have affected the files

Dataset Metadata (cont.)

- **Preservation metadata**
  - **File format (e.g. .txt, .pdf, .doc, .rtf, .xls, .xml, .spv, .jpg, .fits)**
  - **Significant properties**
  - **Technical environment**
  - **Fixity information**

- Adopt a **thesauri** in your field if applicable or compile a data dictionary for your dataset.

- Obtain **persistent identifiers** (e.g. doi, purl) for datasets if possible to ensure data can be found in the future.

- For your full **data management plan**, visit UCF Libraries Data Management Guide. Also, refer to Digital Curation Centre's Checklist for a Data Management Plan (http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf).

Dataset Record Examples, Their Associated Standards, and Data Repositories

- Common Metadata Standards
- Disciplinary Metadata Standards

- **Activity**: Choose a dataset or a standard in your field to examine and critique
  - Social Science Dataset
  - Humanities Dataset
  - Biological Sciences Dataset
  - Biotechnology Dataset
  - Geospatial Dataset
  - Earth Science Dataset
  - Physical Science Dataset
  - Other…

Common Metadata Standards

- **Dublin Core (DC):** A general metadata standard for describing a wide range of digital resources.
  - Dublin Core Metadata Element Set, Version 1.1 (http://dublincore.org/documents/dces/)
  - 15 Elements: Title, Creator, Subject or keyword, Description, Publisher, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights
  - DCMI Metadata Terms (http://dublincore.org/documents/dcmi-terms/)
  - DC Qualifiers (http://dublincore.org/documents/usageguide/qualifiers.shtml)
- Encoded Archival Description (EAD)
  - A standard for encoding archival finding aids with XML.
- Government Information Locator Service (GILS)
  - The Global Information Locator Service defines a core element set for government information so that it can be more searchable and discoverable by the general public.
- ONIX for Books (ONline Information eXchange)
  - An international standard for representing and communicating book industry product information in XML format.

Common Metadata Standards (Cont.): Image and Multi-media Metadata Standards

Categories for the Description of Works of Art (CDWA)

A conceptual framework and guidelines for the description of **art objects and images**.

Visual Resources Association Core Categories (VRA Core)

A data standard for the description of **works of visual culture** as well as the **images** that document them.

PBCore

The metadata standard for **audiovisual media** developed by the public broadcasting community.

NISO Metadata for Digital Images

This technical metadata standard defines a set of metadata elements for **raster digital images** to enable users to develop, exchange, and interpret digital image files. The dictionary has been designed to facilitate interoperability between systems, services, and software as well as to support the long-term management of and continuing access to digital image collections.

Technical Metadata for Multimedia: MPEG-7

The Multimedia Content Description Interface MPEG-7 is an ISO/IEC standard and specifies a set of descriptors to describe various types of **multimedia** information and is developed by the Moving Picture Experts Group.

Disciplinary Metadata Standards: Social Sciences & Humanities

- DDI - Data Documentation Initiative
  - A metadata specification for the social and behavioral sciences. Expressed in XML, the DDI metadata specification supports the entire research data life cycle.

- Text Encoding Initiative (TEI): A standard for the representation of texts in digital form, chiefly in the humanities, social sciences and linguistics.
  - Humanities repositories and Projects:
    - Projects Using the TEI (from the official TEI website)
    - See Appendix 1 for a TEI project example

Disciplinary Metadata Standards: Biological Sciences

ABCD - Access to Biological Collection Data

A standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data).

Darwin Core

A metadata specification for information about the geographic occurrence of species and the existence of specimens in collections.

EML: Ecological Metadata Language

A metadata specification developed by the ecology discipline and for the ecology discipline. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data.

Disciplinary Metadata Standards: Health Sciences

Health Level 7 Standards

HL7 and its members provide a framework (and related standards) for the exchange, integration, sharing, and retrieval of electronic health information. HL7 standards support clinical practice and the management, delivery, and evaluation of health services.

The Cross-Enterprise Document Sharing (XDS) Metadata

The Healthcare Enterprise (IHE) XDS profile is a protocol for sharing clinical documents in health information exchanges. IHE IT Infrastructure Technical Framework volumes can be accessed at: http://ihe.net/Resources/Technical_Frameworks/

National Institute of Health (NIH) Common Data Elements (CDEs)

CDE is a data element that is common to multiple data sets across different studies. NIH encourages the use of CDEs in clinical research, patient registries, and other human subject research in order to improve data quality and opportunities for comparison and combination of data from multiple studies and with electronic health records.

ClinicalTrials.gov Protocol Data Element Definitions

It describes the registration data items (required and optional) that are entered via the Protocol Registration and Results System (PRS).

Disciplinary Metadata:
Biological Sciences and Health Sciences - Repositories

Dryad (https://datadryad.org/)
A digital repository for data underlying the international scientific publications, with an initial focus on evolutionary biology and related fields.

NIH Data Sharing Repositories page lists NIH-supported data repositories that make data accessible for reuse. Most accept submissions of appropriate data from NIH-funded investigators (and others).

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.

GBIF - Global Biodiversity Information Facility
GBIF is a free and open access global web portal promoting and facilitating the mobilization, access, discovery and use of biodiversity data.

ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world.

Examples:
Biological Science Dataset: See Appendix 2
Biotechnology Dataset @ GenBank:
http://www.ncbi.nlm.nih.gov/nucleotide?cmd=Retrieve&dopt=GenBank&list_uids=1293613
Biotechnology Dataset @ PubChem: http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=5760
Clinical Study Dataset @ ClinicalTrials: https://clinicaltrials.gov/show/NCT01196442

Disciplinary Metadata Standards: Earth Science

AgMES

Agricultural Metadata Element Set

AgMES is designed to include agriculture specific extensions for terms and refinements from established metadata standard such as Dublin Core and AGLS to facilitate resource discovery, interoperability and data exchange in the **agriculture** domain.

DIF

Directory Interchange Format

An early metadata initiative from the **Earth sciences** community, intended for the description of **scientific data sets**. It includes elements focusing on instruments that capture data, temporal and spatial characteristics of the data, and projects with which the dataset is associated.

ISO 19115:2003

An internationally-adopted schema for describing **geographic information and services**. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.

FGDC/CSDGM

Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata

Content standard for **digital geospatial metadata** maintained by the Federal Geographic Data Committee (FGDC). Often referred to as the "FGDC Metadata Standard."

(Climate and Forecast) Metadata Conventions

A standard for **climate and forecast** "use metadata" that aims both to distinguish quantities (such as physical description, units, or prior processing) and to locate the data in space-time.

Disciplinary Metadata:
Earth Science- Repositories and Data Centers

AGRIS

AGRIS - International System for Agricultural Science and Technology

A global public domain database using the AgMES standard to describe structured bibliographical records on agricultural science and technology.

NCDC - National Climatic Data Center

The world's largest climate data archive, providing climatological services and data worldwide. It currently promotes the FGDC/CSDGM metadata standard for its datasets.

CEOS International Directory Network

An international effort to assist users in locating Earth science data sets, data services, and visualizations using DIF metadata. It provides free, online access to metadata on scientific data in the Earth sciences: geoscience, hydrospheric, biospheric, satellite remote sensing, and atmospheric sciences.

See a Geospatial Dataset (appendix 3) and an Earth Science Dataset (appendix 4).

Disciplinary Metadata: Physical Science Metadata Standards and Repositories

- CIF - Crystallographic Information Framework
  - An extensible standard file format and set of protocols for the exchange of crystallographic and related structured data.

American Mineralogist Crystal Structure Database

American Mineralogist Crystal Structure Database

A CIF crystal structure database that includes every structure published in the American Mineralogist, The Canadian Mineralogist, European Journal of Mineralogy and Physics and Chemistry of Minerals, as well as selected datasets from other journals.

Crystallography Open Database

An open-access collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, many of which are in CIF form.

Physical Science Dataset Example: http://rruff.geo.arizona.edu/AMS/minerals/Abernathyite

Dublin Core Metadata	Standard DIF
Title	Entry_Title
Creator	Data_Set_Citation: Dataset_Creator
	Personnel: Role: Investigator: Last_Name
	Personnel: Role: Investigator: First_Name
	Personnel: Role: Investigator: Middle_Name
Subject and Keywords	Keyword
	Parameters: Category
	Parameters: Topic
	Parameters: Term
	Parameters: Variable
	Parameters: Detailed_Variable
	Source_Name
	Sensor_Name
	Project
	Location
Description	Summary
Publisher	Data_Set_Citation: Dataset_Publisher
	Data_Center: Data_Center_Name
	Data_Center: Data_Center_URL
	Data_Center: Data Center Contact, Last_Name
	Data_Center: Data Center Contact, First_Name
	Data_Center: Data Center Contact, Middle_Name
Contributor	Personnel: Role:
	Personnel: Last_Name
	Personnel: First_Name
	Personnel: Middle_Name
Date	Data_Set_Citation: Dataset_Release_Date
Resource Type	Data_Set_Citation: Data_Presentation_Form

Generic Dataset and Disciplinary Dataset: Metadata Mapping

- The purpose of metadata mapping is for greater data sharing, exchange, interoperability, usability and re-usability.
- Example: Mapping of Dublin Core to DIF

Format	Group: Distribution
	Distribution_Media
	Distribution_Size
	Distribution_Format
	Fees
Resource Identifier	Data Center: Data_Set_ID
	Data_Set_Citation: Online_Resource
	Related_URL: URL_Content_Type
	Related_URL: URL
Source	Related_URL: URL_Content_Type
	Related_URL: URL
	Source_Name
Language	Data_Set_Language
Relation	Parent_DIF
	Data_Set_Citation: Online_Resource
	Related_URL: URL_Content_Type
	Related_URL: URL
	Reference
Coverage	Location
	Spatial_Coverage: Southernmost_Latitude
	Spatial_Coverage: Northernmost_Latitude
	Spatial_Coverage: Easternmost_Longitude
	Spatial_Coverage: Westernmost_Longitude
	Temporal_Coverage: Start_Date
	Temporal_Coverage: Stop_Date
	Paleo_Temporal_Coverage: Paleo_Start_Date
	Paleo_Temporal_Coverage: Paleo_Stop_Date
	Paleo_Temporal_Coverage: Chronostratigraphic_Unit
Rights Management	Use_Constraints
	Access_Constraints

Generic Dataset and Disciplinary Dataset: Metadata Mapping

- Compared to generic Dublin Core, discipline metadata standards provide more granularity.

Metadata Standards and Datasets

- Common Metadata Standards (http://guides.ucf.edu/metadata/genMetaStandards)
- Disciplinary Metadata Standards (http://guides.ucf.edu/metadata/domMetaStandards)
- Questions on metadata standards:
  - Do they make sense to you?
  - Are the standards adequate in your field? Can data be well documented?
  - Have you used any standard, or, will you consider it in your future study and research?

Disciplinary Metadata and Data Repositories

For more information on disciplinary metadata standards, tools, and use cases, please refer to UK Digital Curation Centre (DCC)'s Disciplinary Metadata page.

For more information on data repositories and digital repositories, please refer to Databib, OpenDOAR and OAD.

DataBib: Databib is a community-driven, annotated bibliography of research data repositories. Databib is now merged with re3data.org (http://www.re3data.org/).

OpenDOAR

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

OAD

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

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:

Dataset: Experience of Violence in the Lives of Homeless Persons: The Florida Four City Study, 2003-2004 (ICPSR 20363)

http://dx.doi.org/10.3886/ICPSR20363.v1

http://dx.doi.org/	10.3886/	ICPSR20363 .v1
resolver service	prefix (assigning body)	suffix (resource)

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.

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:**
    - Wright, James D., Jana L. Jasinski, Elizabeth Mustaine, and Jennifer Wesely. Experience of Violence in the Lives of Homeless Persons: The Florida Four City Study, 2003-2004. ICPSR20363-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-11-22. doi:10.3886/ICPSR20363.v1
    - **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 3.1 (released 2014-10)
  (http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf)

Table 1: DataCite Mandatory Properties

ID	Property	Obligation
1	Identifier (with type sub-property)	M
2	Creator (with name identifier sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M

Table 2: DataCite Recommended and Optional Properties

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

DataCite XML File

```
− <resource xsi:schemaLocation="http://datacite.org/schema/kernel-2.2 http://schema.datacite.org/meta/kernel-2.2/metadata.xsd">
    <identifier identifierType="DOI">10.3886/ICPSR20363.v1</identifier>
  + <creators></creators>
  − <titles>
    − <title>
        Experience of Violence in the Lives of Homeless Persons: The Florida Four City Study,
      </title>
    </titles>
  − <publisher>
      Inter-university Consortium for Political and Social Research
    </publisher>
    <publicationYear>2010</publicationYear>
  + <subjects></subjects>
  + <dates></dates>
    <resourceType resourceTypeGeneral="Dataset"> survey data </resourceType>
  − <alternateIdentifiers>
      <alternateIdentifier alternateIdentifierType="ICPSR Study Number">20363</alternat
    </alternateIdentifiers>
    <version>1</version>
  + <descriptions></descriptions>
  </resource>
```

FIELDS:
resource
creator
title
publisher
publicationYear
subject
date
resourceType
alternativeIdentifier
version
description
…

http://www.icpsr.umich.edu/icpsrweb/ICPSR/datacite/studies/20363

Controlled Vocabulary

- **Controlled vocabulary** is a standardized set of terms used to organize knowledge for subsequent retrieval. It can facilitate search and browsing. It can be universally agreed on or locally created.

- **What to consider in applying or designing a thesauri for your project:**
  - **Scope of the material (core and surrounding topics, your purpose, existing thesauri and your resource);**
  - **Your project needs and intended audience;**
  - **Funder requirements and institutional expectation;**
  - **What types of controlled vocabularies you may need: subject, genre, physical format, personal names, organization names, events…**
  - **When choosing particular terms over others, consider three warrants: literary warrant (discipline and field literature), user warrant and organizational warrant.** (Gazan, CONTROLLED VOCABULARY & THESAURUS DESIGN, http://www.loc.gov/catworkshop/courses/thesaurus/pdf/cont-vocab-thes-trnee-manual.pdf)

Controlled Vocabulary (Cont.)

- **For traditional library catalog:**
  - MARC Code List for Countries http://www.loc.gov/marc/countries/
  - MARC Code List for Languages http://www.loc.gov/marc/languages/
  - MARC Source Codes for Vocabularies, Rules, and Schemes http://www.loc.gov/marc/sourcecode/form/formsource.html
- **For digital and online resources:**
  - Internet Media Types www.iana.org/assignments/media-types/index.html
  - MODS Note Types http://www.loc.gov/standards/mods/mods-notes.html
  - DCMI Type Vocabulary http://dublincore.org/documents/dcmi-terms/index.shtml#H7

Controlled Vocabulary: Subject Thesauri

- **Subject Thesauri and Ontologies**
  - AGROVOC (Agricultural Organization of the United Nations Vocabulary)
  - Astronomy Thesaurus
  - CAB Thesaurus (for life sciences, technology and social sciences)
  - CIF dictionaries (for Physics)
  - Eurovoc (European Union Thesaurus)
  - Ethnographic Thesaurus
  - Gene Ontology
  - GeoNames
  - Getty Institute Art and Architecture Thesaurus Online
  - Getty Institute Thesaurus of Geographic Names
  - ICD (International Classification of Diseases)
  - Library of Congress Authorities for subject headings
  - Library of Congress Thesaurus for Graphic Materials
  - Logical Observation Identifiers Names, and Codes (LOINC)
  - MESH (Medical Subject Headings)
  - Public Health Language
  - Rare Books and Manuscripts Section (RBMS) Controlled Vocabularies
  - RxNorm (for drugs)
  - SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms)
  - STW Thesaurus for Economics
  - UNBIS Thesaurus
  - UNESCO Thesaurus
  - USDA National Agricultural Library Agriculture Thesaurus

Question: Have you ever used thesauri in your study and research?

Controlled Vocabulary: Name Authorities

Library of Congress Name Authority File (LC/NAF)

The LC/NAF provides authoritative data for names of persons, organizations, events, places, and titles.

Virtual International Authority File (VIAF)

The VIAF™ (Virtual International Authority File) **combines multiple name authority files** into a single OCLC-hosted name authority service. The goal of the service is to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web.

Getty Union List of Artist Names (ULAN)

The ULAN includes proper names and associated information about **artists**. *Artists* may be either individuals (persons) or groups of individuals working together (corporate bodies). Artists in the ULAN generally represent creators involved in the conception or production of **visual arts and architecture.**

Linked Data: Metadata Structure Standards

Resource Description Framework (RDF)

RDF is a standard model for data interchange on the Web. RDF extends the linking structure of the Web to **use URIs to name the relationship between things as well as the two ends of the link** (this is usually referred to as a **"triple"**). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

SKOS: Simple Knowledge Organization for the Web

SKOS is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary.

Linked data examples:

- FAST: Faceted Application of Subject Terminology;
- Dewey Decimal Classification;
- Open Metadata Registry (RDA vocabularies)
- Library of Congress Linked Data Service

…

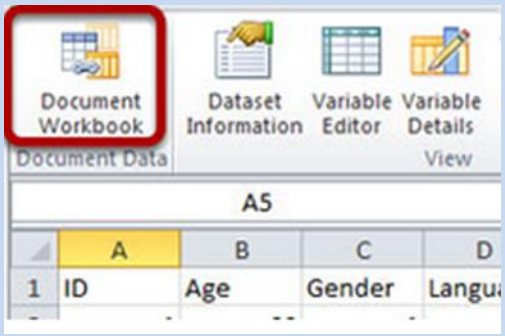
MADS/RDF

The Metadata Authority Description Schema (MADS) is an XML schema for an element set that may be used to provide metadata about **authorized forms of agents** (people, organizations), **events, and terms** (topics, geographics, genres, etc.). MADS/RDF builds on MADS/XML as a **knowledge organization system.**

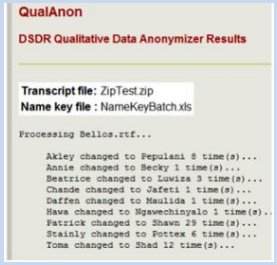
Web Ontology Language (OWL)

The OWL 2 Web Ontology Language, is an **ontology language for the Semantic Web with formally defined meaning**. OWL 2 ontologies provide classes, properties, individuals, and data values and are stored as Semantic Web documents. OWL 2 ontologies can be used along with information written in **RDF**, and OWL 2 ontologies themselves are primarily exchanged as RDF documents.

Curation Tools: Data Preparation

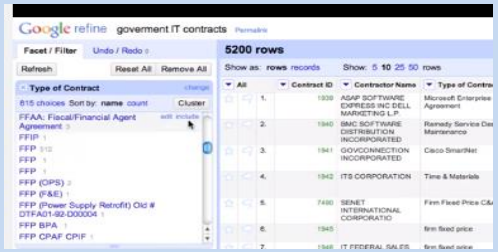


Colectica for Microsoft Excel
A free tool to document your spreadsheet data using the Data Documentation Initiative (DDI) metadata format, the open standard for data documentation.
http://www.colectica.com/software/colecticaforexcel



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
unlocking | data | creating | knowledge

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: Data Analysis and Visualization

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/

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

Altova XMLSpy is an advanced XML editor for modeling, editing, transforming, and debugging XML-related technologies.
http://www.altova.com/xmlspy.html

Curation Tools: Lab Management

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/

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 theof services and scripts that scientists with large-scale data use to execute research.
https://kepler-project.org/

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

Part IV:
Thoughts and Services

A Researcher's View vs. A Curator or a Librarian's Perspective on Data Documentation

- Section II addresses data documentation more from the researcher's view?
- Section III interprets data documentation more from a curator or librarian's perspective?
- What do researchers really care about?
- Will each party see the other side's points and emphases?

A life cycle approach

CDL Curation and Publishing Services

http://www.cdlib.org

Create, edit, share, and save data management plans

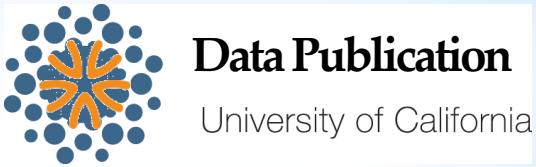
Open source add-in for Microsoft Excel as a data collection tool

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

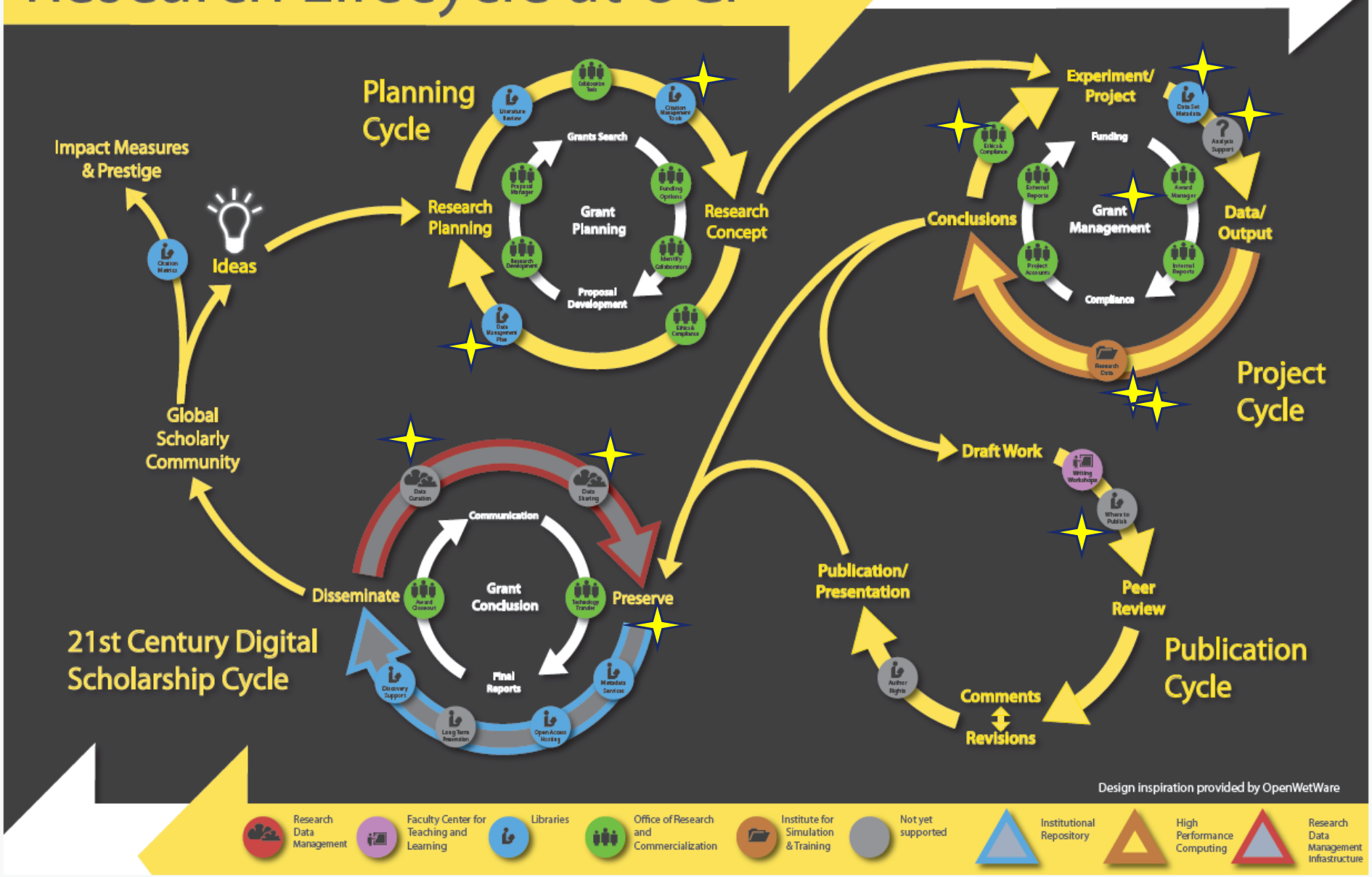


* *This slide is by Joan Starr, California Digital Library.*

http://www.slideshare.net/joanstarr/dataset-metadata-tools-approaches-for-access-preservation?from_search=1

Research Lifecycle at UCF

University of Central Florida Libraries



Data Set Related Services

http://library.ucf.edu/ScholarlyCommunication/UCFResearchLifecycle.pdf

Dataset Metadata Service at UCF

- **"Data Set (also called 'Dataset') Metadata" provides researchers consultation on:**
  - Project and dataset documentation;
  - 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.

Related Library Resources

- Scholarly Communication: (http://library.ucf.edu/ScholarlyCommunication/)
  - SC Contact Information (http://library.ucf.edu/ScholarlyCommunication/Contact.php)
- UCF Library Research Guides (http://guides.ucf.edu)
  - Metadata Guide (http://guides.ucf.edu/metadata)
  - Data Management Guide (http://guides.ucf.edu/data)
- Research and Information Services (http://library.ucf.edu/Reference/)
  - Subject Librarians (http://library.ucf.edu/SubjectLibrarians/)

Appendix 1: Humanities Data in TEI Format

Overall structure of an ENRICH-conformant XML document. ENRICH is “European Networking Resources and Information concerning Cultural Heritage.” Examples from “The ENRICH Schema — A Reference Guide.” The guide is a conformant subset of Release 1.4 of TEI P5.

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

http://projects.oucs.ox.ac.uk/ENRICH/Deliverables/referenceManual_en.html

The minimal required structure for teiHeader:

```
<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>
 <revisionDesc>
  <change when="2008-01-0
<!-- [revision information] -
  </change>
 </revisionDesc>
</teiHeader>
```

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

Appendix 1: Humanities Data in TEI Format (Cont.)

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

Fields:

- msDesc
  - msIdentifier
    - Settlement
    - repository
    - Idno
    - altIdentifier
  - msContents
    - P
      - quote
      - title
  - physDesc
    - p
      - material
  - History
    - p
      - origPlace
      - origDate
      - quote

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

```
<msDesc xml:id="ex1" xml:lang="en">
 <msIdentifier>
  <settlement>Oxford</settlement>
  <repository>Bodleian Library</repository>
  <idno>MS. Add. A. 61</idno>
  <altIdentifier type="former">
   <idno>28843</idno>
  </altIdentifier>
 </msIdentifier>
 <msContents>
  <p>
   <quote xml:lang="lat">Hic incipit Bruitus Anglie,</quote> the
   <title xml:lang="lat">De origine et gestis Regum Angliae</title>
    of Geoffrey of Monmouth (Galfridus Monumetensis):
    beg. <quote xml:lang="lat">Cum mecum multa &amp;amp; de multis.</quote>
    In Latin.</p>
 </msContents>
 <physDesc>
  <p>
   <material>Parchment</material>: written in
    more than one hand: 7¼ x 5⅜ in., i + 55 leaves, in double
    columns: with a few coloured capitals.</p>
 </physDesc>
 <history>
  <p>Written in
  <origPlace>England</origPlace> in the
  <origDate>13th cent.</origDate> On fol. 54v very faint is
  <quote xml:lang="lat">Iste liber est fratris guillelmi de buria de ... Roberti
    ordinis fratrum Pred[icatorum],</quote> 14th cent. (?):
  <quote>hanauilla</quote> is written at the foot of the page
    (15th cent.). Bought from the rev. W. D. Macray on March 17, 1863, for
    £1 10s.</p>
 </history>
</msDesc>
```

The official TEI P5 guideline is at: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

More TEI projects and examples are available at the TEI website: http://www.tei-c.org/Activities/Projects/

Appendix 2: Biological Science Dataset

dc.contributor.author	Crawford, Nicholas G.
dc.contributor.author	Faircloth, Brant C.
dc.contributor.author	McCormack, John E.
dc.contributor.author	Brumfield, Robb T.
dc.contributor.author	Winker, Kevin
dc.contributor.author	Glenn, Travis C.
dc.date.accessioned	2012-05-18T15:48:08Z
dc.date.available	2012-05-18T15:48:08Z
dc.date.issued	2012-05-16
dc.identifier	doi:10.5061/dryad.75nv22qj
dc.identifier.citation	Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biology Letters 8(5): 783-786.
dc.identifier.uri	http://hdl.handle.net/10255/dryad.38214
dc.description	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…
relation.haspart	doi:10.5061/dryad.75nv22qj/1
relation.haspart	doi:10.5061/dryad.75nv22qj/2
relation.haspart	…

Dryad
(https://datadryad.org/)

This is an example of full metadata view.

http://www.datadryad.org/handle/10255/dryad.38214?show=full

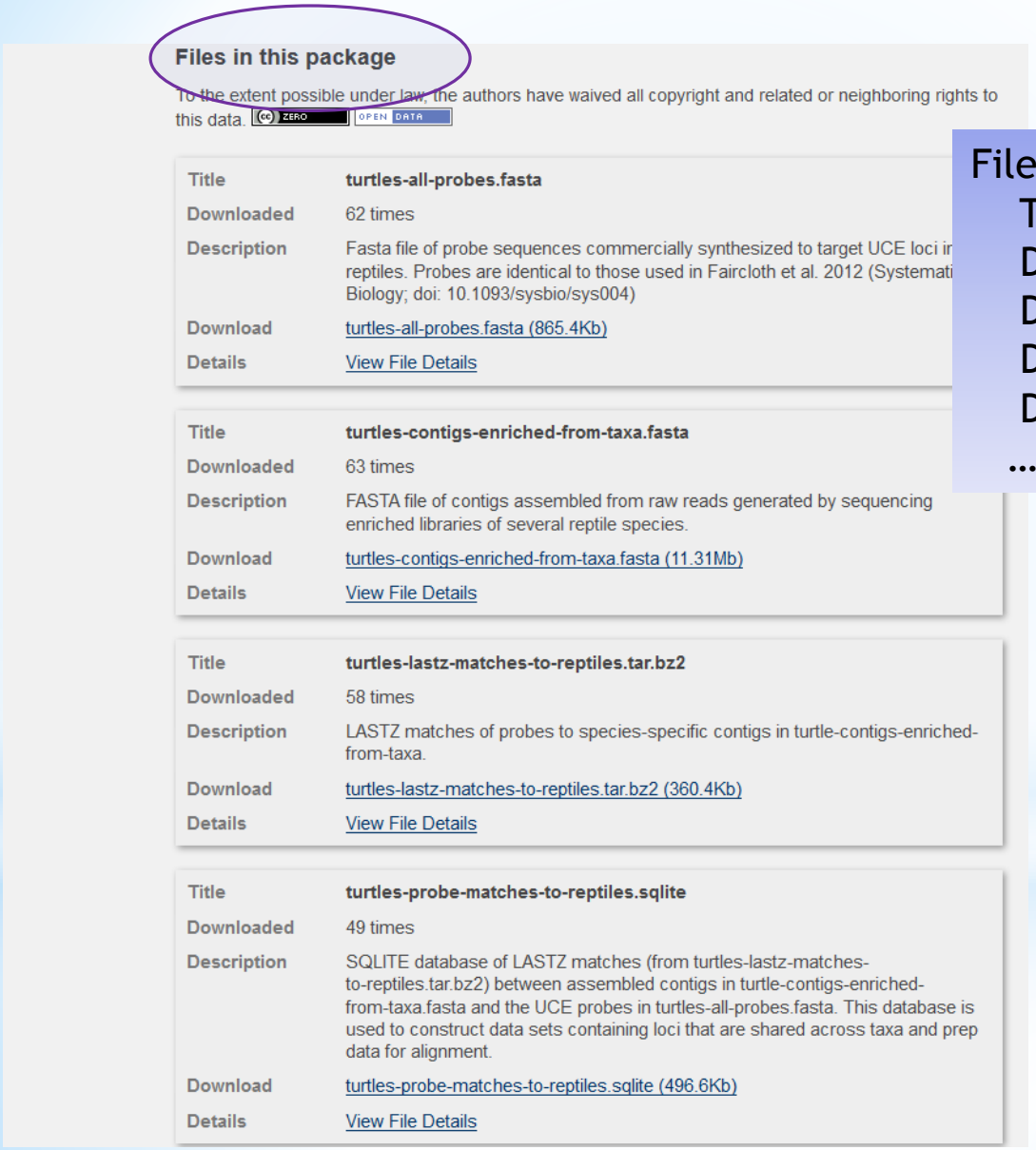
Appendix 2: Biological Science Dataset (Cont.)

dc.relation.isreferencedby	doi:10.1098/rsbl.2012.0331
dc.relation.isreferencedby	PMID:22593086
dc.subject	ultraconserved elements
dc.subject	phylogenomic
dc.subject	phylogenetics
dc.subject	reptiles
dc.subject	turtles
dc.subject	evolution
dc.subject	archosaurs
dc.title	Data from: More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs
dc.type	Article
dwc.ScientificName	Pantherophis guttata
dwc.ScientificName	Pelomedusa subrufa
dwc.ScientificName	Chrysemys picta
dwc.ScientificName	Alligator mississippiensis
dwc.ScientificName	Crocodylus porosus
dwc.ScientificName	Sphenodon tuatara
dwc.ScientificName	Gallus gallus
dwc.ScientificName	Taeniopygia guttata
dwc.ScientificName	Anolis carolinensis
dwc.ScientificName	Homo sapiens
dc.contributor.corresponding Author	Faircloth, Brant C.
prism.publicationName	Biology Letters

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.

Appendix 2: Biological Science Dataset (Cont.)



Files in this package

To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data. CC ZERO | OPEN DATA

Title	**turtles-all-probes.fasta**
Downloaded	62 times
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 (Systemati Biology; doi: 10.1093/sysbio/sys004)
Download	turtles-all-probes.fasta (865.4Kb)
Details	View File Details

Title	**turtles-contigs-enriched-from-taxa.fasta**
Downloaded	63 times
Description	FASTA file of contigs assembled from raw reads generated by sequencing enriched libraries of several reptile species.
Download	turtles-contigs-enriched-from-taxa.fasta (11.31Mb)
Details	View File Details

Title	**turtles-lastz-matches-to-reptiles.tar.bz2**
Downloaded	58 times
Description	LASTZ matches of probes to species-specific contigs in turtle-contigs-enriched-from-taxa.
Download	turtles-lastz-matches-to-reptiles.tar.bz2 (360.4Kb)
Details	View File Details

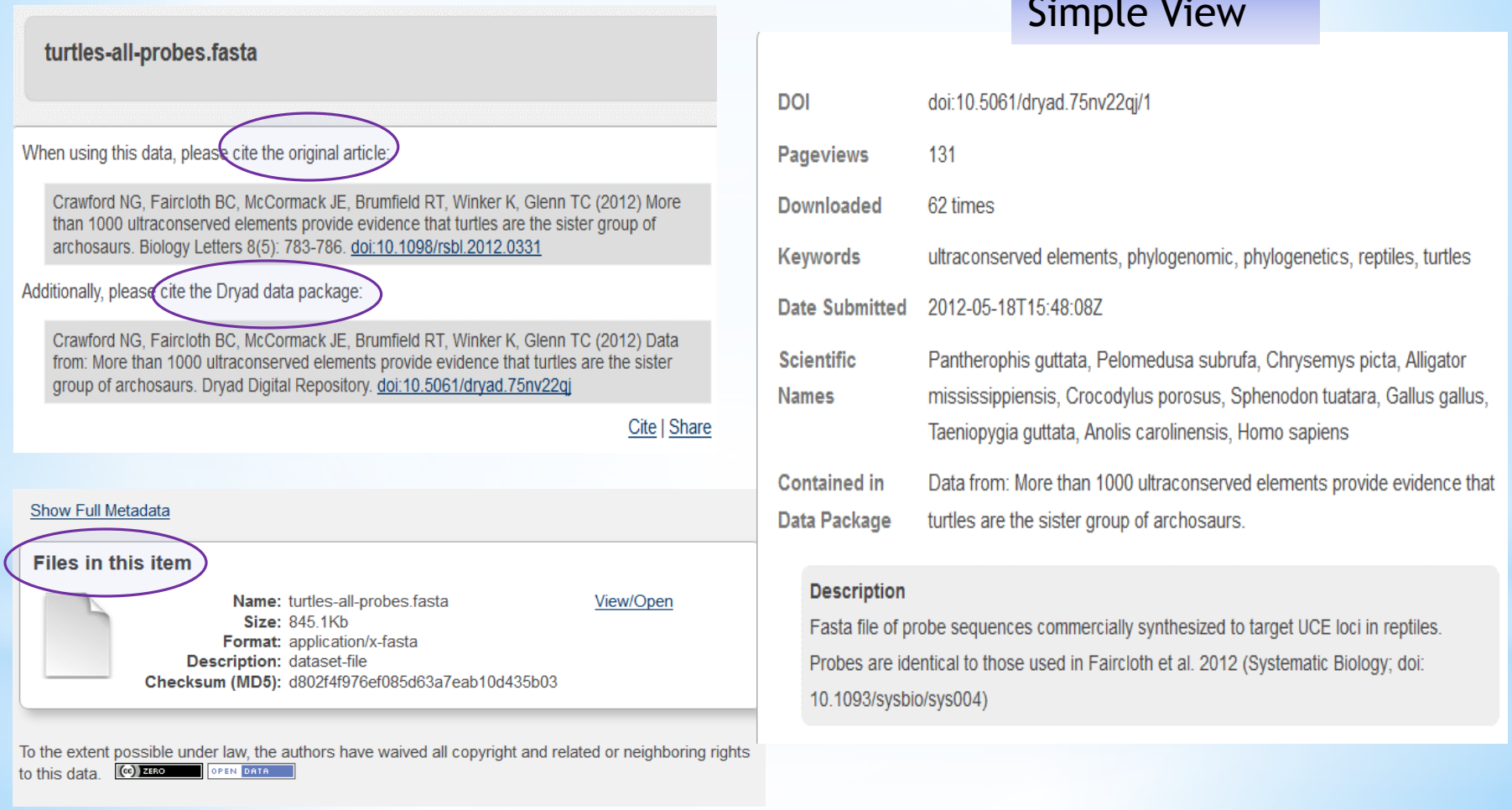
Title	**turtles-probe-matches-to-reptiles.sqlite**
Downloaded	49 times
Description	SQLITE database of LASTZ matches (from turtles-lastz-matches-to-reptiles.tar.bz2) between assembled contigs in turtle-contigs-enriched-from-taxa.fasta and the UCE probes in turtles-all-probes.fasta. This database is used to construct data sets containing loci that are shared across taxa and prep data for alignment.
Download	turtles-probe-matches-to-reptiles.sqlite (496.6Kb)
Details	View File Details

Files in this package
- Title
- Downloaded
- Description
- Download
- Details

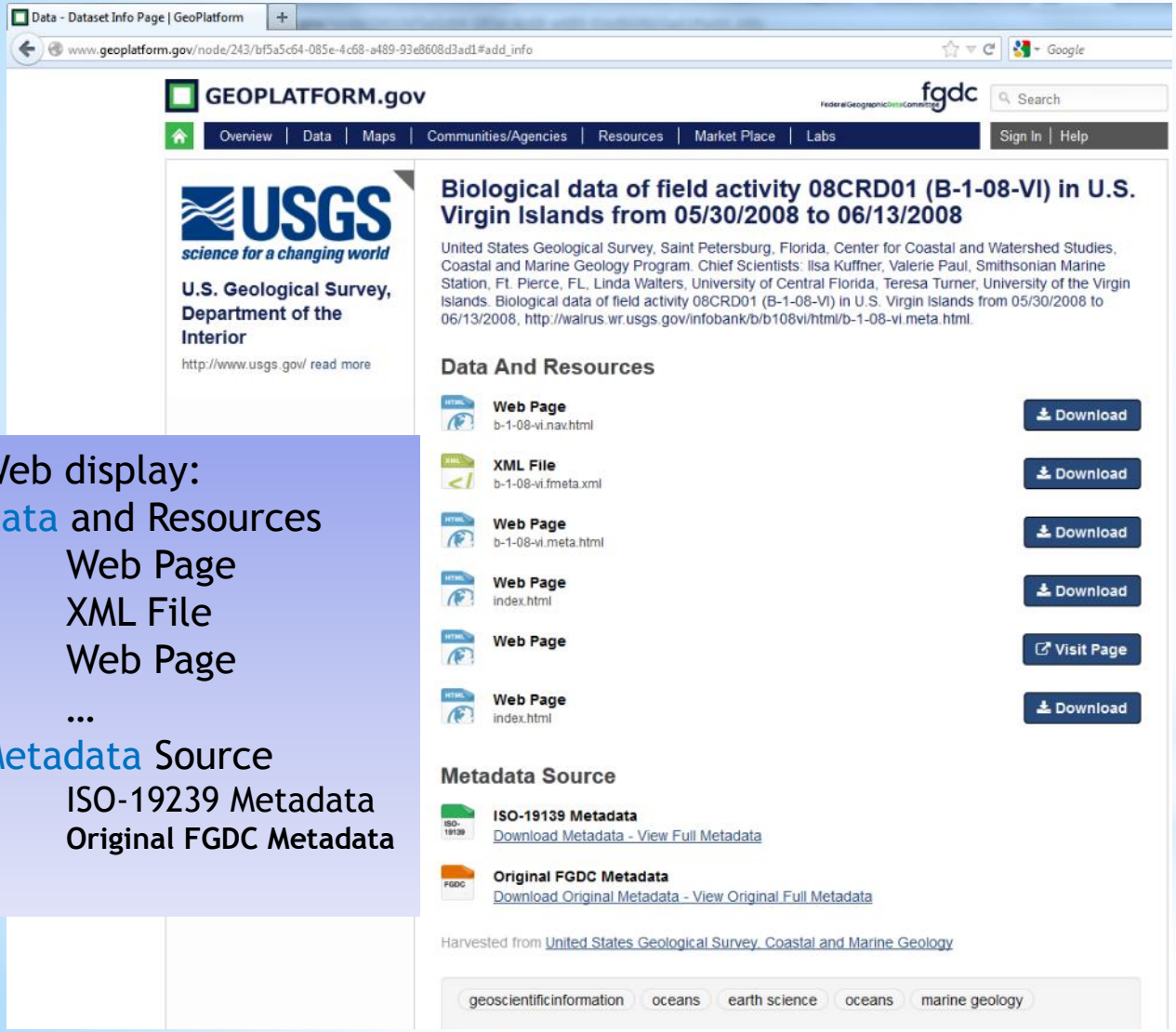
...

Appendix 2: Biological Science Dataset (Cont.)

- If clicking View File Details, it displays:



Appendix 3: Geospatial Dataset



Web display:

Data and Resources
- Web Page
- XML File
- Web Page
- …

Metadata Source
- ISO-19239 Metadata
- **Original FGDC Metadata**

http://www.geoplatform.gov/node/243/bf5a5c64-085e-4c68-a489-93e8608d3ad1

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.

Appendix 3: Geospatial Dataset (Cont.)

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

Metadata

FGDC/CSDGM 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>

Organisation Name: U.S. Geological Survey (USGS) <http://www.usgs.gov>, Coastal and Marine Geology (CMG) <http://walrus.wr.usgs.gov>

Position Name: **InfoBank Group Leader** <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

Metadata Standard Version: ISO 19115-2:2009(E)

http://walrus.wr.usgs.gov/infobank/b/b108vi/html/b-1-08-vi.fmeta.outline.html

Appendix 3: 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…

FGDC/CSDGM Metadata

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

Organisation Name: U.S. Geological Survey (USGS) <http://www.usgs.gov>, Coastal and Marine Geology (CMG) <http://walrus.wr.usgs.gov>

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

Appendix 3: 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:

URL: http://walrus.wr.usgs.gov/infobank/b/b108vi/html/b-1-08-vi.nav.html

Distributor:

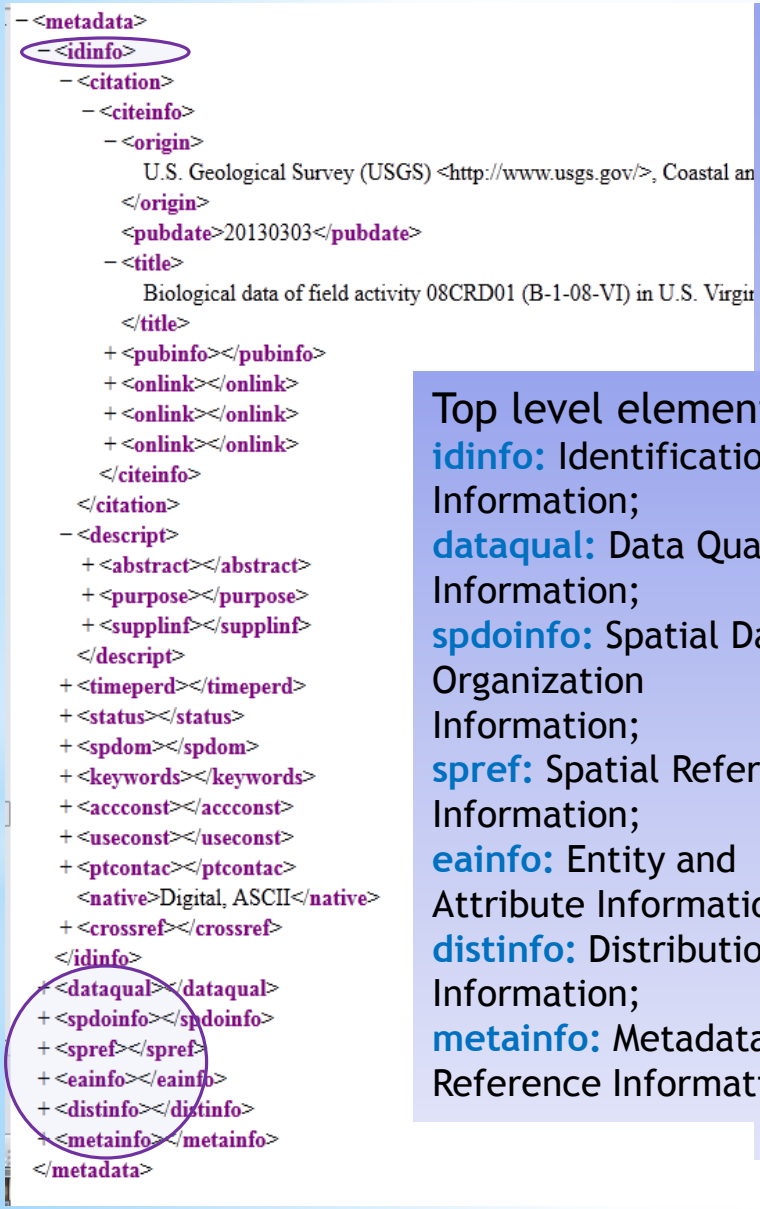
Distributor Contact: …

Quality

Scope: Dataset

FGDC/CSDGM Metadata

Appendix 3: Geospatial Dataset (in XML View)



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.

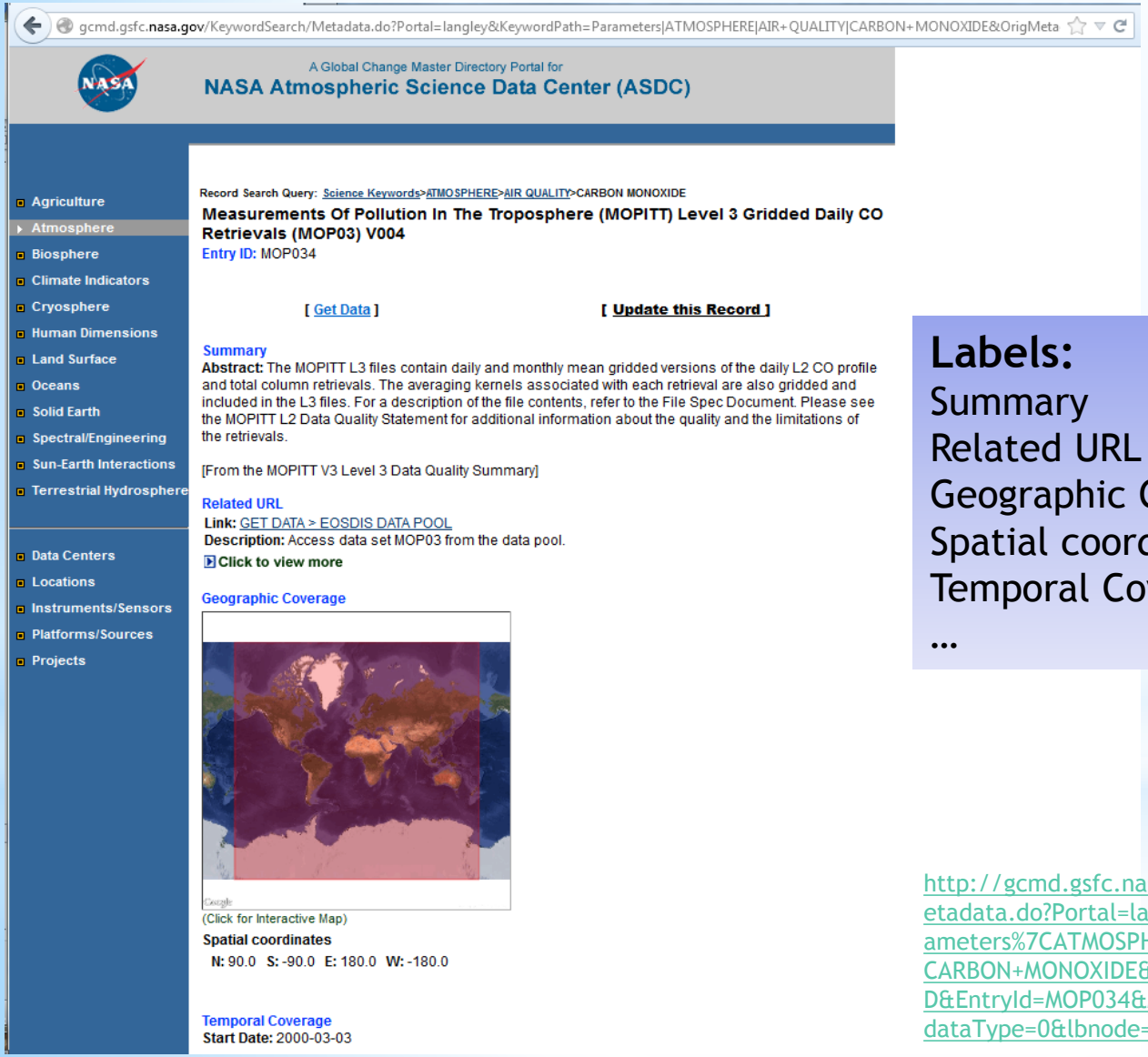
CSDGM Fields (under idinfo):

Idinfo

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

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

Appendix 4: Earth Science Dataset

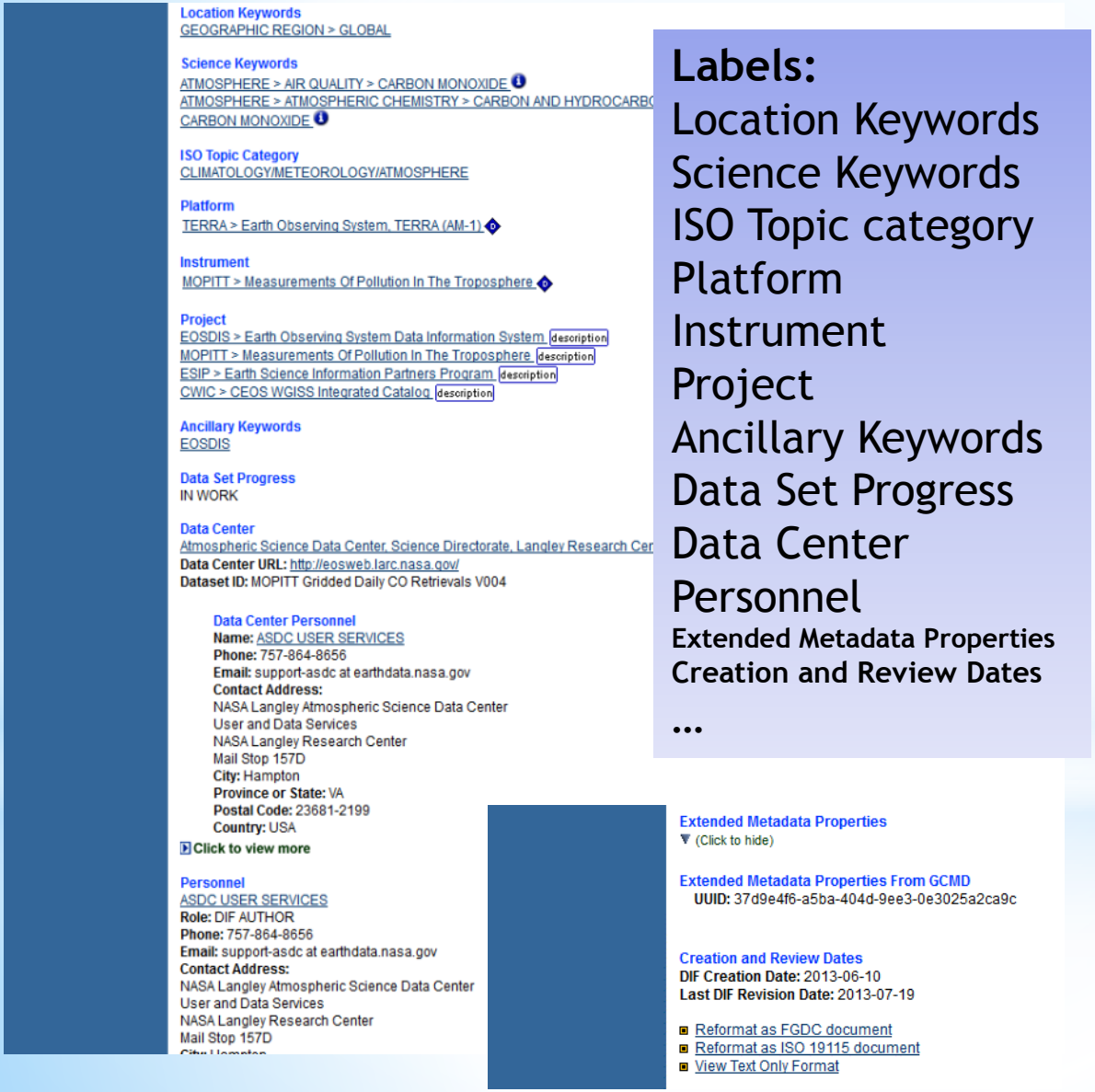


NASA Atmospheric Science Data Center (ASDC)

Labels:
Summary
Related URL
Geographic Coverage
Spatial coordinates
Temporal Coverage
…

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

Appendix 4: Earth Science Dataset (Cont.)



Labels:
Location Keywords
Science Keywords
ISO Topic category
Platform
Instrument
Project
Ancillary Keywords
Data Set Progress
Data Center
Personnel
Extended Metadata Properties
Creation and Review Dates
…

Directory Interchange Format (DIF):

a descriptive and standardized format for exchanging information about scientific data sets. The DIF Writer's Guide: http://gcmd.gsfc.nasa.gov/User/difguide/difman.html.

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)

Contact:
Sai Deng, Metadata Librarian and Associate Librarian
sai.deng@ucf.edu
407-823-4312 (Office)

Thank you!