1-1-2011

Using Data Curation Profiles (DCPs) as a means of raising data management awareness

Jeremy R. Garritano
Purdue University, jgarrita@purdue.edu

Using Data Curation Profiles (DCPs) as a means of raising data management awareness

Jeremy R. Garritano

Chemical Information Specialist & Assoc Prof of Library Science

Purdue University Libraries

jgarrita@purdue.edu

241st ACS National Meeting, Anaheim, CA

March 28, 2011

Outline

- Background and Development
- DCP Toolkit
- DCP Sections
- Tips/Lessons Learned
- Training of Purdue Librarians
- www.datacurationprofiles.org

IMLS Grant

- National Leadership Grant LG-06-07-0032-07
- Purdue University Libraries and University of Illinois GSLIS
- Develop a model for the description of datasets so they may be curated and shared

Uses of the DCPs

At individual level:

- Provides a way for information professional to interact with researchers/groups
- Provides a thorough way for researchers/group to consider data management needs

At institutional level:

- Serves as documentation for the management of a particular data set
- Can be shared to make sure everyone is on the same page
- Helps institutions identify the types of tools, infrastructure and responsibilities for data services staff and data management program

At an even broader level:

- May be used by others as a guide in developing data services at their own institutions
- May be used as objects of research to better understand data management needs and similarities/differences across disciplines

DCP Prototypes and Creation

- Initially looked at the fields of astronomy, ecology, and crystallography
- Looked for commonalities and developed interview questions
- 19 researchers were interviewed
- Draft DCPs created
- Focus groups of librarians were conducted

The DCP Toolkit

- DCP user guide (librarian)
- Interviewer’s manual (librarian)
- Interview worksheet (researcher)
- DCP template (blank and a sample)

DCP user guide

- “The User Guide provides information about the Data Curation Profiles, including background information, the purpose and use of Data Curation Profiles, and directions on how to construct a Data Curation Profile.”
- Includes checklist

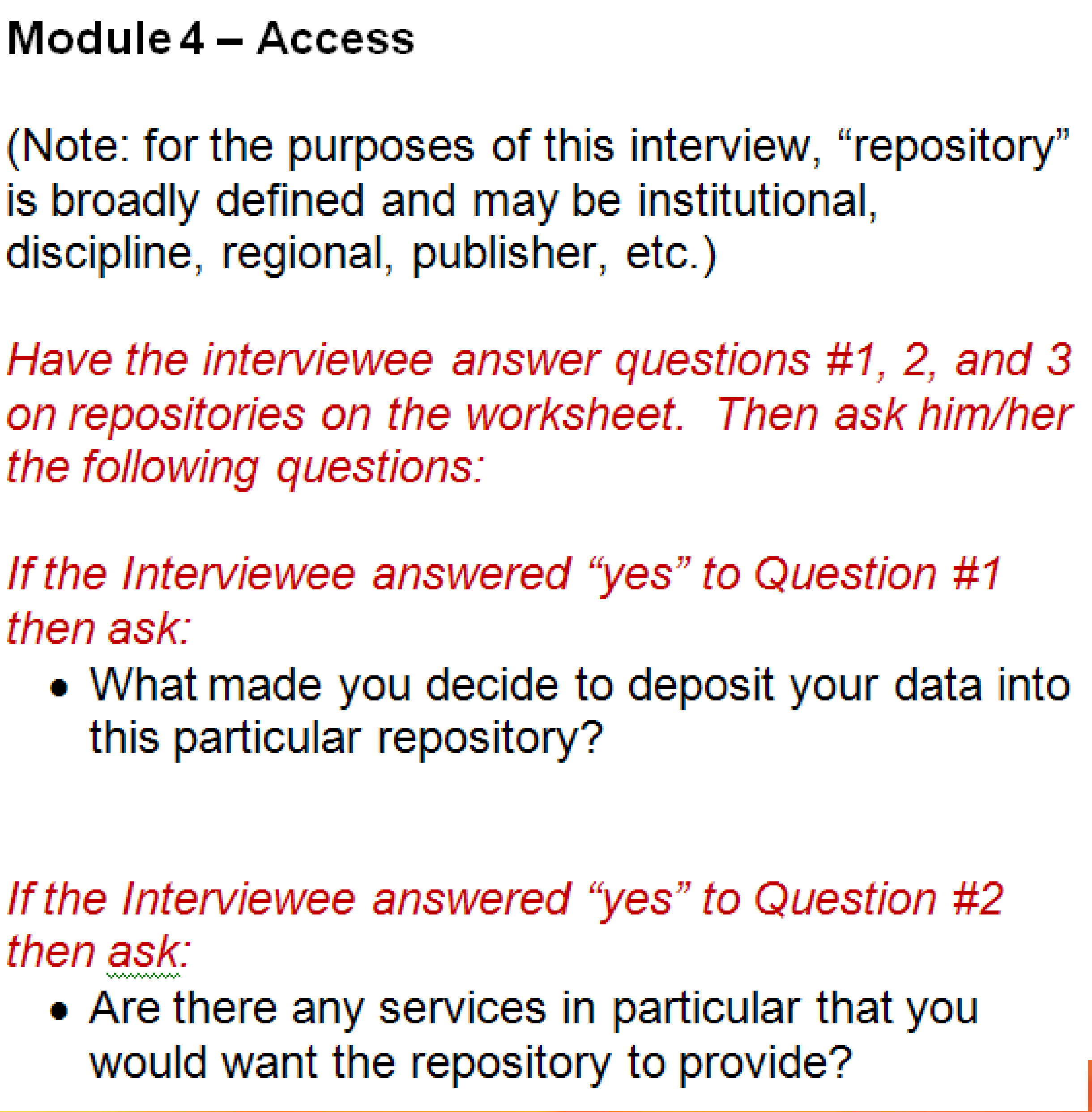
Methodology

- Preparation (1-2 hours)
    - IRB approval(?)
- Interviews (1-3 hours)
- Construction of the profile (5-10 hours)

Interviewer’s manual

- “The Interviewer’s Manual provides the framework for the interview. It contains text and questions to be read to the participating researcher over the course of the interview. Some of the questions to be asked will be in response to the answers given by the researcher in the Interview Worksheet.”

Sample from interviewer’s manual



Module 4 – Access

(Note: for the purposes of this interview, “repository” is broadly defined and may be institutional, discipline, regional, publisher, etc.)

Have the interviewee answer questions #1, 2, and 3 on repositories on the worksheet. Then ask him/her the following questions:

If the Interviewee answered “yes” to Question #1 then ask:

- What made you decide to deposit your data into this particular repository?

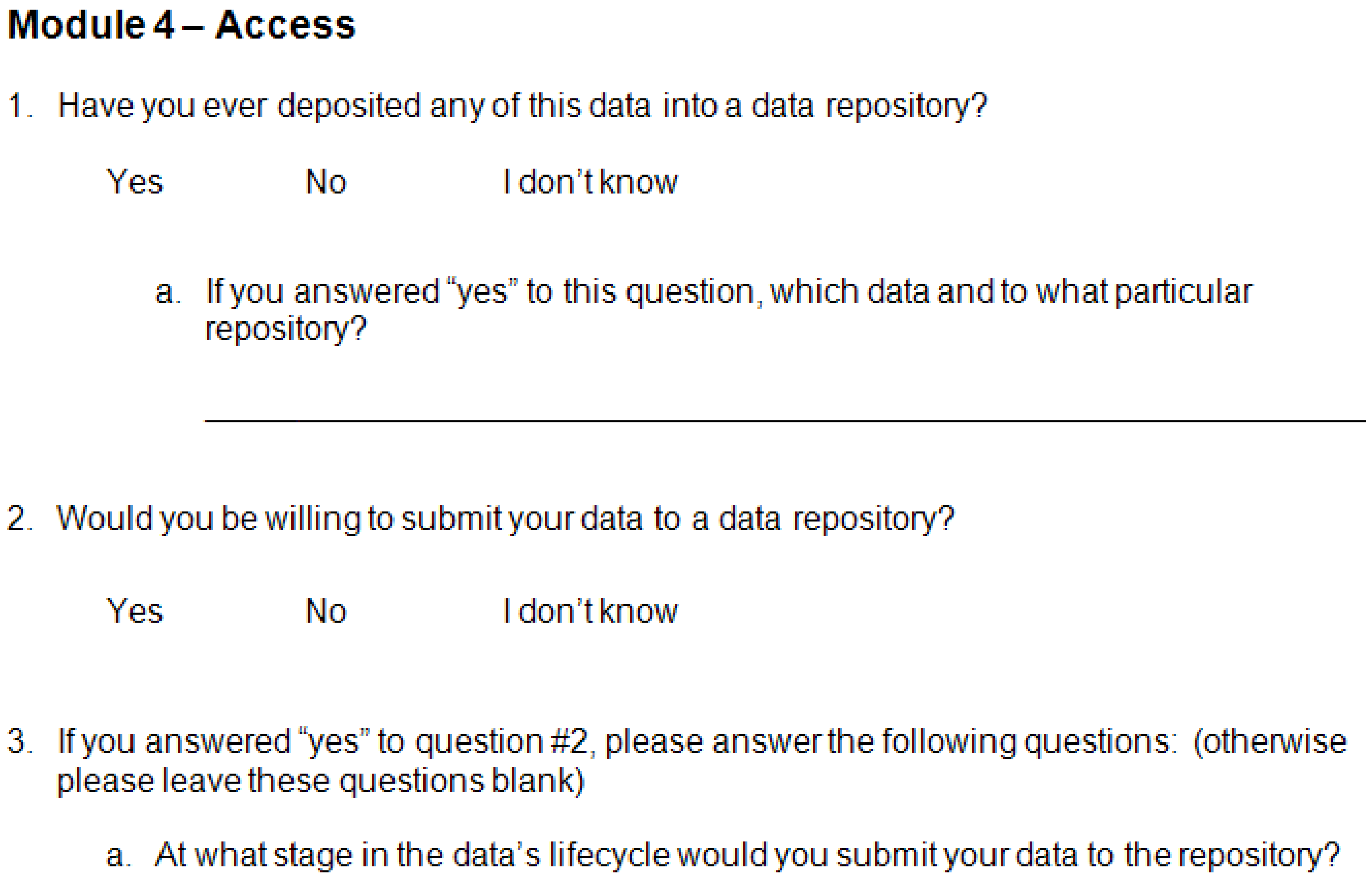
If the Interviewee answered “yes” to Question #2 then ask:

- Are there any services in particular that you would want the repository to provide?

Interview worksheet

- “The Interview Worksheet is to be given to the researcher by the interviewer at the start of the interview. It is the worksheet that the participating researcher will fill out over the course of the interview. In addition to capturing important information, the responses provided by the researcher will serve as the basis for further discussion during the interview.”

Sample from interview worksheet



Module 4 – Access

1. Have you ever deposited any of this data into a data repository?

   Yes   No   I don't know

   a. If you answered "yes" to this question, which data and to what particular repository?

      ____

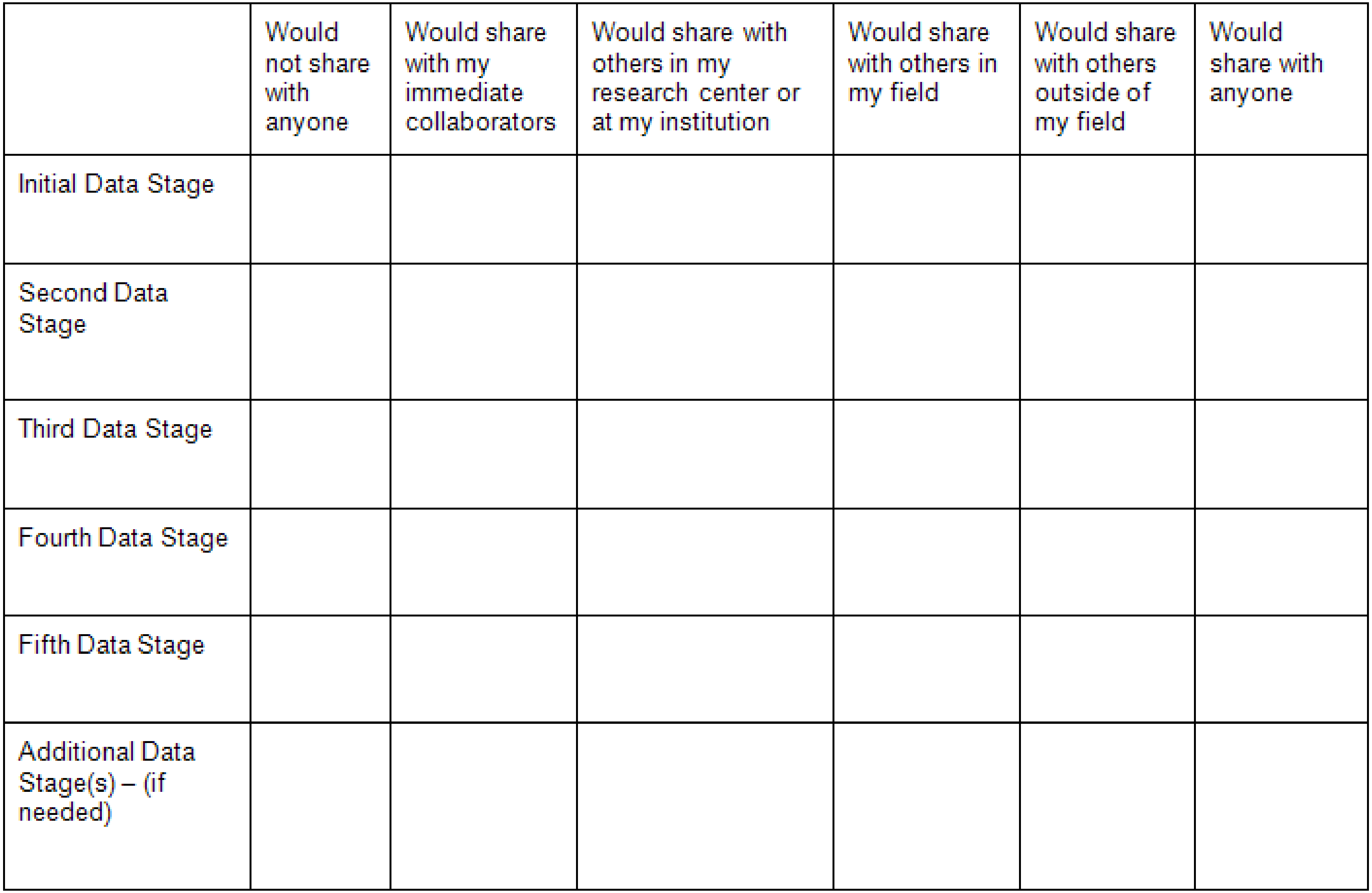
2. Would you be willing to submit your data to a data repository?

   Yes   No   I don't know

3. If you answered "yes" to question #2, please answer the following questions: (otherwise please leave these questions blank)

   a. At what stage in the data's lifecycle would you submit your data to the repository?

Sample from interview worksheet



	Would not share with anyone	Would share with my immediate collaborators	Would share with others in my research center or at my institution	Would share with others in my field	Would share with others outside of my field	Would share with anyone
Initial Data Stage						
Second Data Stage						
Third Data Stage						
Fourth Data Stage						
Fifth Data Stage						
Additional Data Stage(s) – (if needed)						

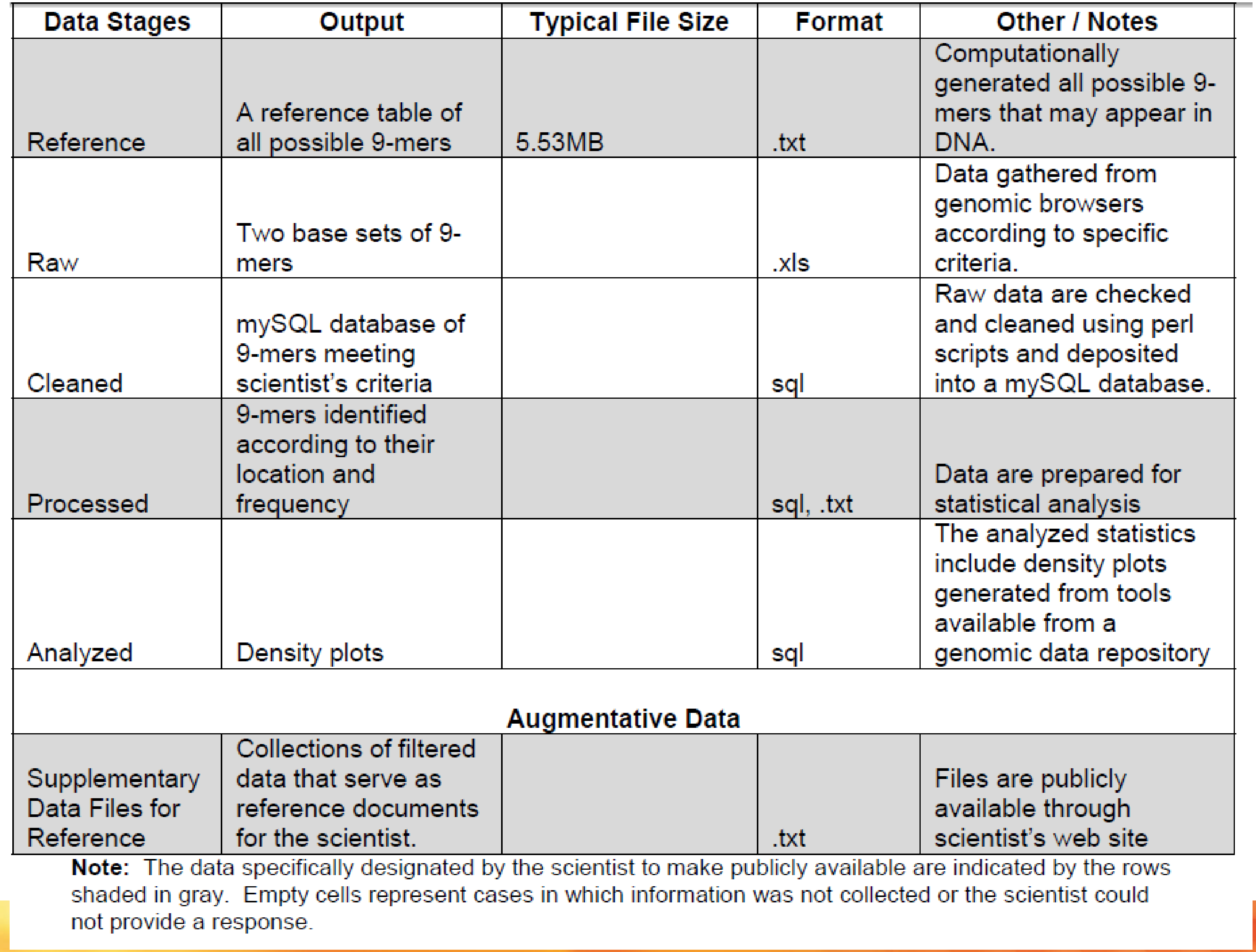
DCP template

- “The Data Curation Profile Template describes the structure of the Data Curation Profile. Each section or sub-section within the Data Curation Profile template contains a brief definition of the information that is needed to populate an individual Data Curation Profile for the participating researcher.”

Sections of the DCP

- Section 1 - Brief summary of data curation needs
- **Section 2 - Overview of the research**
- **Section 3 - Data kinds and stages**
- Section 4 - Intellectual property context and information

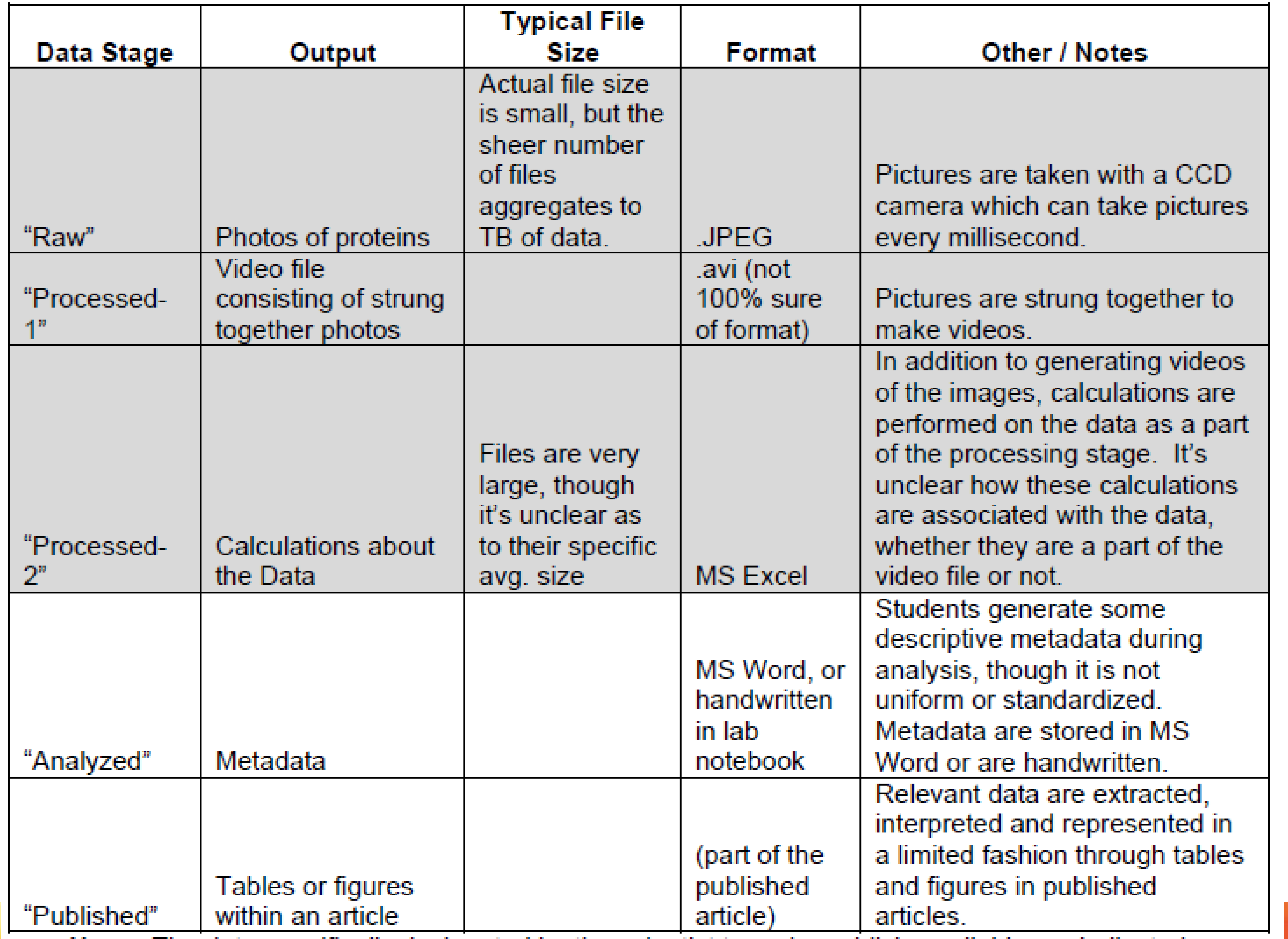
Example of data table



Data Stages	Output	Typical File Size	Format	Other / Notes
Reference	A reference table of all possible 9-mers	5.53MB	.txt	Computationally generated all possible 9-mers that may appear in DNA.
Raw	Two base sets of 9-mers		.xls	Data gathered from genomic browsers according to specific criteria.
Cleaned	mySQL database of 9-mers meeting scientist's criteria		sql	Raw data are checked and cleaned using perl scripts and deposited into a mySQL database.
Processed	9-mers identified according to their location and frequency		sql, .txt	Data are prepared for statistical analysis
Analyzed	Density plots		sql	The analyzed statistics include density plots generated from tools available from a genomic data repository
		Augmentative Data		
Supplementary Data Files for Reference	Collections of filtered data that serve as reference documents for the scientist.		.txt	Files are publicly available through scientist's web site

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

Example of data table



Data Stage	Output	Typical File Size	Format	Other / Notes
"Raw"	Photos of proteins	Actual file size is small, but the sheer number of files aggregates to TB of data.	.JPEG	Pictures are taken with a CCD camera which can take pictures every millisecond.
"Processed-1"	Video file consisting of strung together photos		.avi (not 100% sure of format)	Pictures are strung together to make videos.
"Processed-2"	Calculations about the Data	Files are very large, though it's unclear as to their specific avg. size	MS Excel	In addition to generating videos of the images, calculations are performed on the data as a part of the processing stage. It's unclear how these calculations are associated with the data, whether they are a part of the video file or not.
"Analyzed"	Metadata		MS Word, or handwritten in lab notebook	Students generate some descriptive metadata during analysis, though it is not uniform or standardized. Metadata are stored in MS Word or are handwritten.
"Published"	Tables or figures within an article		(part of the published article)	Relevant data are extracted, interpreted and represented in a limited fashion through tables and figures in published articles.

Sections of the DCP

- **Section 5 - Organization and description of data (incl. metadata)**
- Section 6 - Ingest / Transfer
- **Section 7 – Sharing & Access**
- Section 8 - Discovery

Sections of the DCP

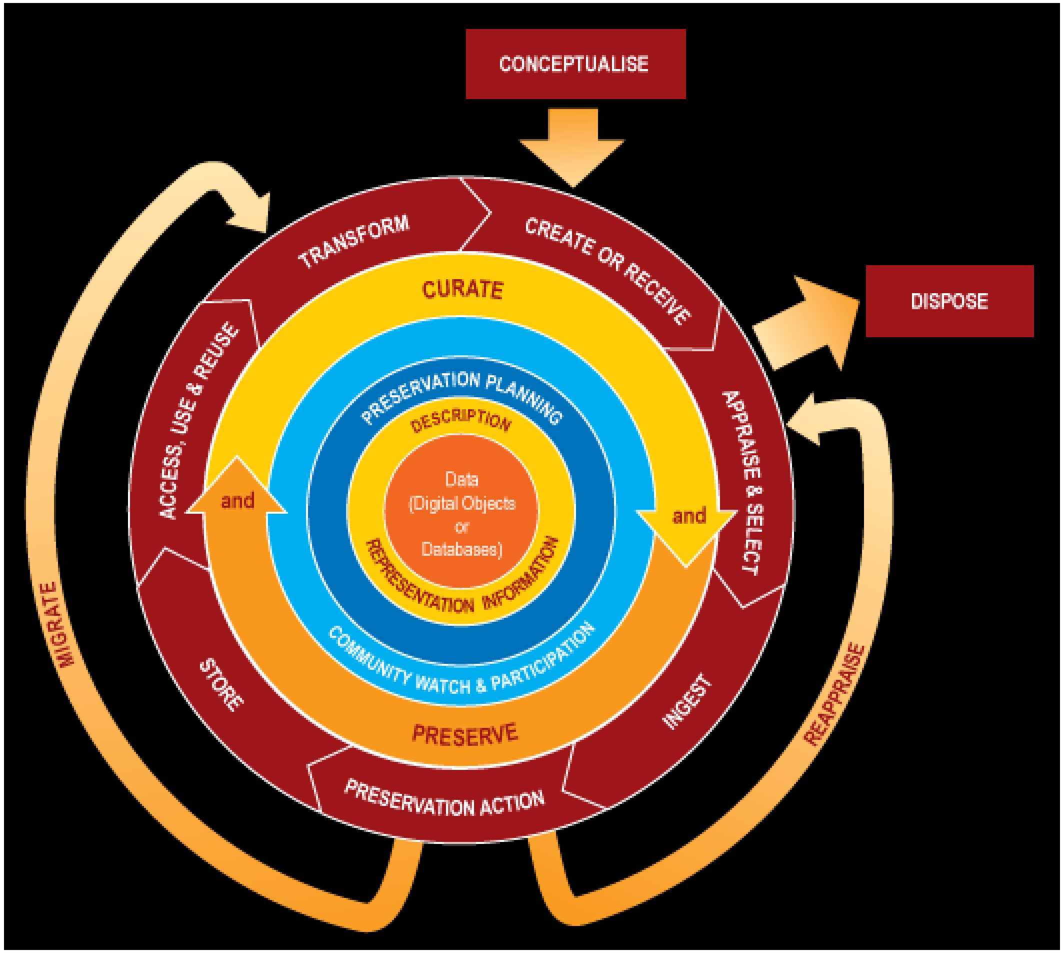
- Section 9 - Tools
- Section 10 – Linking / Interoperability
- Section 11 - Measuring Impact
- Section 12 – Data Management
- Section 13 - Preservation

Tips / Lessons Learned

- Ask for an authored article for preparation
- Graduate student(s) are often present (necessary)
- Record the conversation (transcription/index)
- A follow-up interview is almost always needed
- Takes a lot of time (2+ hours)
- Show/describe DCC Curation Lifecycle Model

DCC Curation Lifecycle Model

http://www.dcc.ac.uk/resources/curation-lifecycle-model



Training of Purdue Librarians

- 3 separate training sessions
- 1. Intro to the Data Curation Profile
  - – Introduces concepts and terminology related to data management and curation
  - – Provides background into how the Profile was developed
  - – Outcome: librarians should be able to discuss researcher projects to assess needs for reference and build deeper liaison relationships

Training of Purdue Librarians

- 2. Data Curation Profile Toolkit
  - Introduces DCP Toolkit
  - Walks through the DCP sections
  - Uses librarian’s research as simple example
  - Outcome: librarians should be able to engage in deeper conversations with researchers, including the disposition and dissemination of their research outputs and data

Training of Purdue Librarians

- 3. Using the Data Curation Profile
- Work with specific examples using the Profile Toolkit
- Hands-on interpretation of researcher examples/interactions
- Outcome: librarians should be able to work with faculty to elicit a DCP

Modifications

- For NSF Data Management Plan Requirements

www.datacurationprofiles.org

- History/background
- Download the toolkit
- Submit a profile
- Sample completed profiles
- Workshops
- Online forum

Future workshops

Monday, April 4, 2011
GWLA / University of Washington - Seattle, WA

Friday, April 29, 2011
CARL / Colorado State University - Fort Collins, CO

Friday, July 8, 2011
SWFLN / Florida Southern College - Lakeland, FL

Thursday, October 20, 2011
Internet Librarian - Monterey, CA

Acknowledgements

- Purdue University personnel
  - Scott Brandt
  - Jake Carlson
  - Meagan Sapp Nelson
  - Michael Witt

- University of Illinois personnel
  - Melissa Cragin