Library Theory and Research Panel Data Curation Project

- Anna Maria Tammaro, LTR Chair ,University of Parma
- Krystyna K. Matusiak, University of Denver
- Frank Andreas Sposito, University of Denver
- Terry Weech, University of Illinois Urbana-Champaign

Outline

- Project overview
- Research design
- Phase I: Literature review and vocabulary
- Phase II: Content analysis of positon announcements
- Phase III: Findings from the interviews and document analysis
- LIS education
- Conclusion

Project Overview

- The main goal of the project was to identify the characteristics of roles and responsibilities of data curators in the international and interdisciplinary contexts

- The objectives were to prepare:
  - A vocabulary (a list of terms)
  - An ontology (formal representation of a set of concepts)

Research Design

- Mixed-method design
  - Quantitative content analysis of job announcements for data curators and RDM librarians
  - Semi-structured interviews with professionals working as data librarians, data curators, or research data managers
- Research questions:

  R1: How is data curation defined by practitioners / professional working in the field?

  R2: What terms are used to describe the roles for professionals in data curation area?

  R2: What are primary roles and responsibilities of data curators?

  R3: What are educational qualifications and competencies required of data curators?

Research Phases

Phase I – Literature review and vocabulary

Phase II – Content analysis of the position announcements with data curation responsibilities in libraries, archives, and research centers

- Select job postings from the following sites:
  - American Library Association (ALA) Job list: http://joblist.ala.org/
  - **Code4lib:** http://jobs.code4lib.org/jobs/data-curation/
  - **The IASSIST Jobs Repository:** http://www.iassistdata.org/resources/jobs/all

Phase III – Interviews with data curators + questionnaires and document analysis

PHASE I: VOCABULARY

Key phrases and corpora

Key phrases

Key phrases were identified only on frequencies and syntactic information

Keyphrase Digger* (http://dh.fbk.eu/technologies/kd) to extract relevant terms from the corpora

*Fondazione Bruno Kessler (FBK, see http://ict.fbk.eu/), University of Trento

Corpora

- Bibliography (2015 and 2017)
- Job descriptions
- Questionnaire
- Interviews
- Project Edison for Data scientist

It was apparent that there was a minimal overlap between any two pairs of corpora : different communities have different terminology!

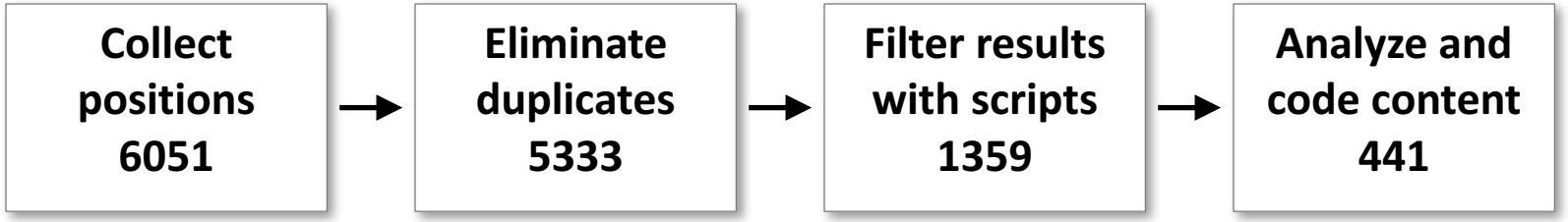
Vocabulary

Term	Definition	Related Term	WIKIDATA Code
Research Data Management (RDM)	Activities around the life cycle of research-related data	Research Data: collection of facts produced through systematic inquiry (Q15809982)	(Q30089794)
Data curation	Work performed to ensure meaningful and enduring access to data	Digital curation: selection, preservation, maintenance, collection and archiving of digital assets (Q5276060)	(Q15088675)
Data management	All disciplines related to managing data as a valuable resource	Data Management Plan (Q17085509)	(Q1149776)
Digital Preservation	Formal endeavor to ensure that digital information of continuing value remains accessible and usable	Preservation: maintenance of objects as closely as possible to their original condition also called conservation (Q1479406)	(Q632897)
Data Science	Interdisciplinary field about processes and systems to extract knowledge or insights from data	Data Scientist: a person studying and working with data (Q29169143)	(Q2374463)

PHASE II: CONTENT ANALYSIS

Content Analysis: Process

Phase I



- Environmental scan
- >10 sources
- Grouped as
  - CODE4LIB
  - IASSIST
  - OTHER

- Org & title match
- Duplicates:
  - within sources
  - across sources
  - over time

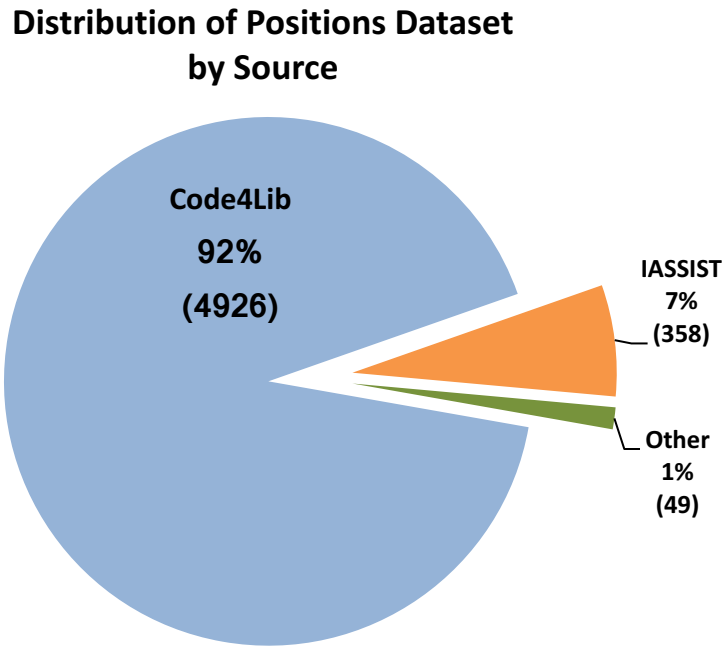
- 15 expressions
- Very wide net
- Data/digital
- Curator/curation
- Data management
- Data librarian
- Data science
- Data services

- Coded for data curator based on responsibilities
  - *Primary*
  - *Secondary*
- Also coded for other attributes

Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: Sources

- Code4Lib and IASSIST positions systematically scraped from websites: includes all records
- “Other” positions drawn from various sources
- Duplicate listings default to Code4Lib
- Duplicate listings within same site default to most recent date

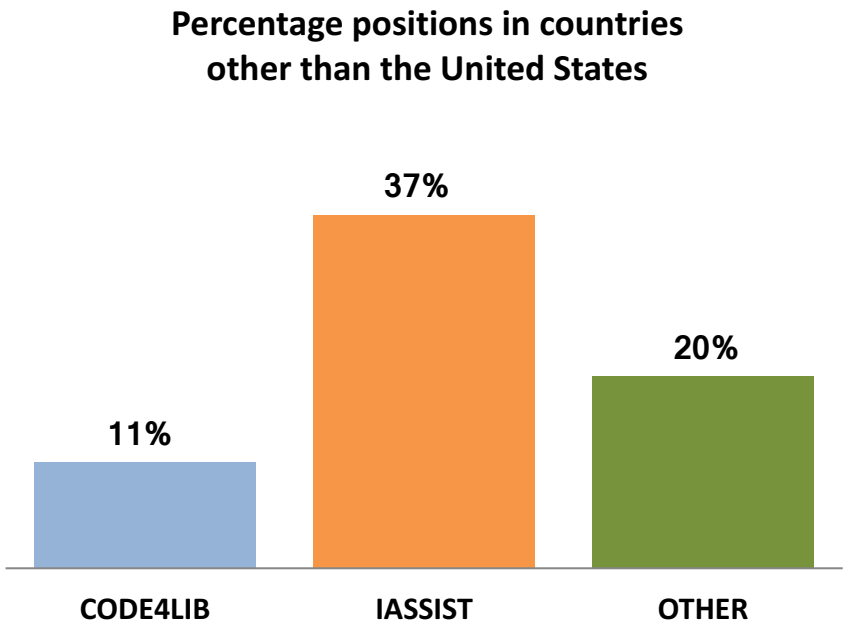


Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: Countries

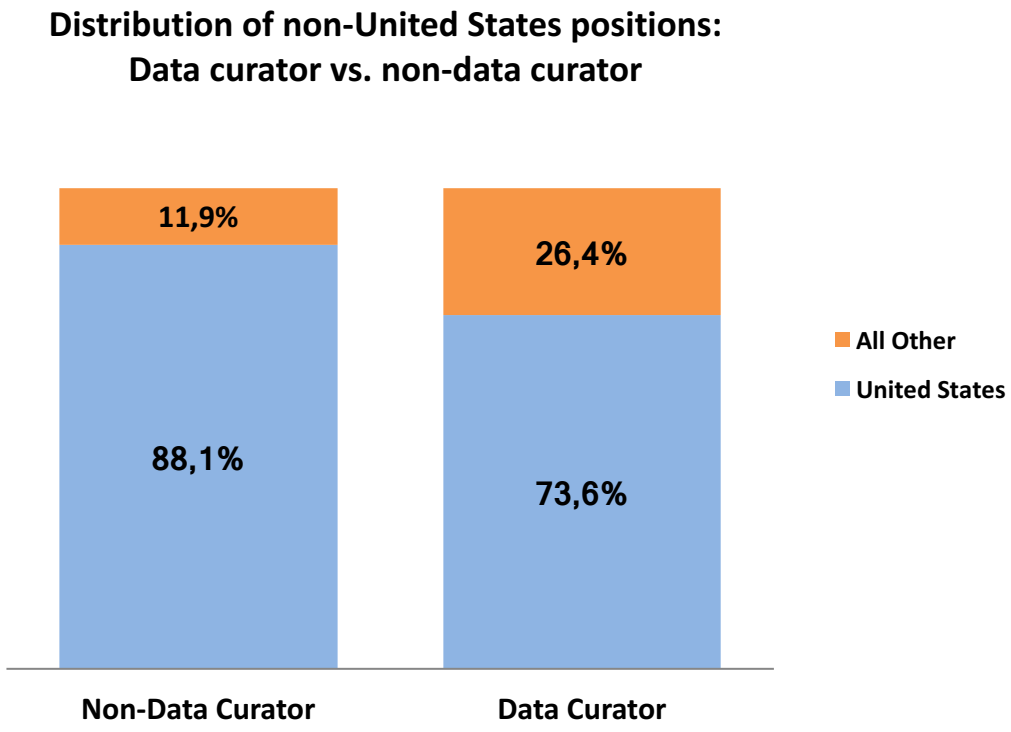
- Almost entirely English language
- But not uniquely North American
- 34 countries represented:

Afghanistan, Australia, Austria, Belgium, Canada, China, Czech Republic, Denmark, Egypt, Finland, France, Germany, Hungary, Ireland, Italy, Japan, Kuwait, Mexico, Netherlands, New Zealand, Norway, Poland, Qatar, Saudi Arabia, Singapore, South Africa, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, and United States



Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: Countries

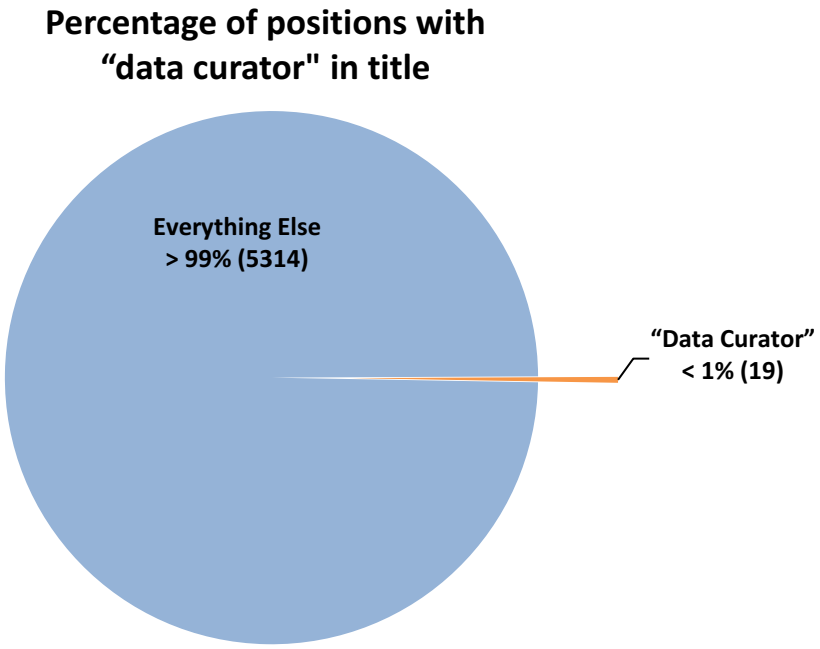


Note: Code4Lib and IASSIST positions only

Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: “Data Curator”

- How to know what a data curator is before looking at the positions?
- Analyze jobs with “data curator” in the title
- Bad news: there aren’t many



Title (normalized)	Count	Pct
Data Curator	12	63.2%
Scientific Data Curator	2	10.5%
Research Data Curator	2	10.5%
Assistant Data Curator	1	5.3%
Humanities Data Curator	1	5.3%
GeoSpatial Data Curator	1	5.3%

Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: "Data Curator"

Curating data or curating people (and processes)?

Sample 1: New York University - Data Curator

"The Data Curator will manage the **Data Lifecycle** from beginning to end. …[He/she] **prepares files** for deposit, including analysis of the structure of study data; **data normalization, cleaning,** [and] **authority management**; organization of digital and physical inventories; … **manage data ingest** and access workflows; catalog data using and **maintaining controlled vocabularies**…"

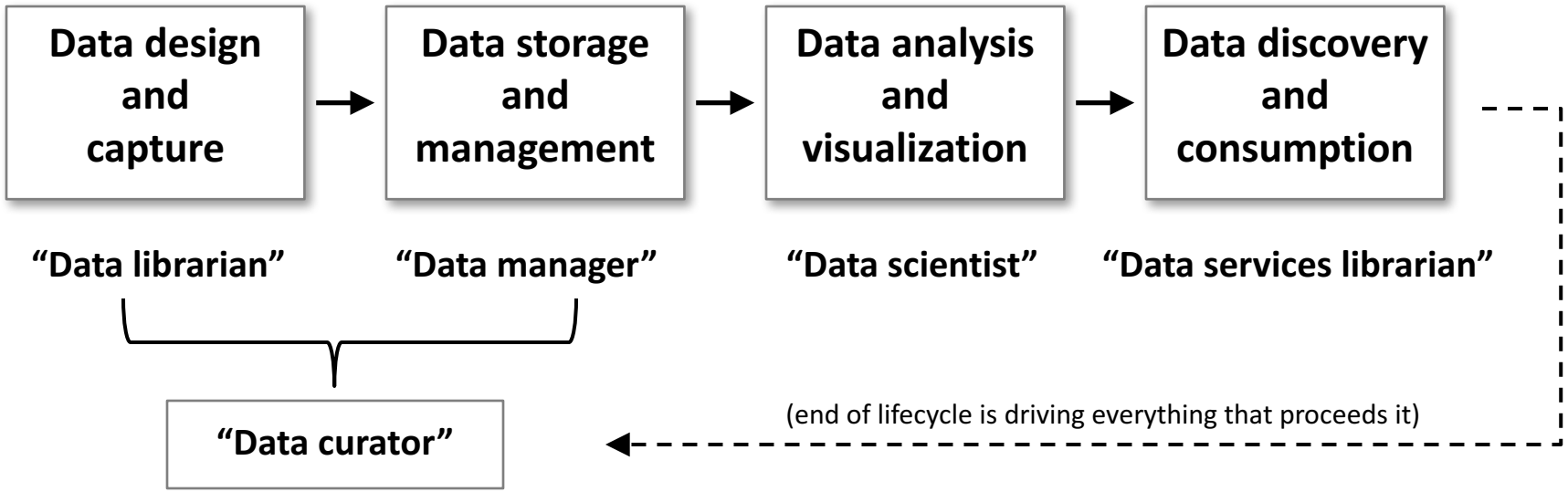
Sample 2: University of Melbourne - Research Data Curator

"Provision of **guidance and support** in identification and **establishment of methods** for the **long-term management** of research data throughout the research **lifecycle**, including **issues of digital preservation** and access, documentation, data repository management, intellectual property rights, and security of sensitive data."

Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: "Data Curator"

Sketch of the data lifecycle



- *Always* concerned about data quality
- *Always* working *directly* with
  - people
  - data/digital objects
  - both
- *Often* working with higher-level
  - policy/planning
  - workflow design

"Curator"

- Latin *curare:* "to take care of"
- Also "to cure"
- "Overseer, manager, guardian"
- Originally caring for minors or disabled persons

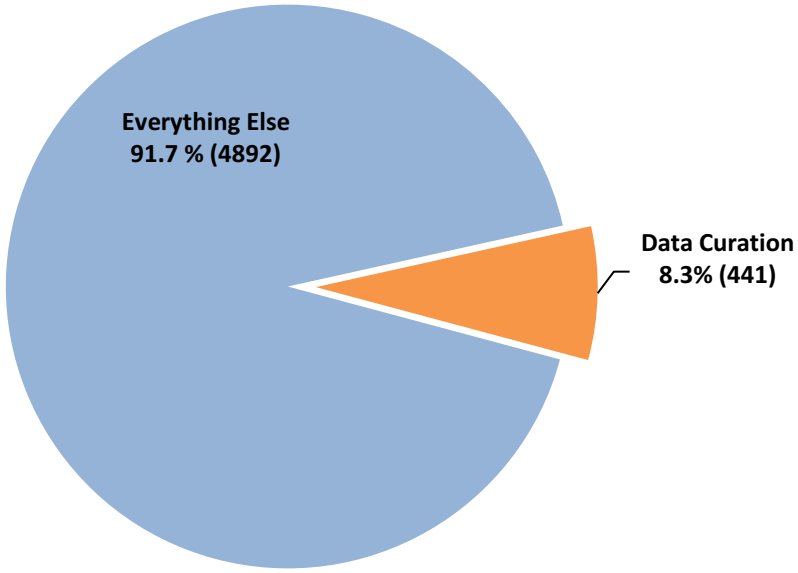
***Please note: all analyses and results are preliminary and may change as our research advances*.**

Lifecycle typology adapted from Matthew Mayernik, Research Data Services Specialist, NCAR.

Content Analysis: Data Curation

Title (normalized)	Count	Pct
Librarian - Data	15	3.4%
Librarian - Data Services	13	2.9%
Data Curator	12	2.7%
Librarian - Digital Scholarship	10	2.3%
Librarian - Digital Initiatives	10	2.3%
Librarian - Research Data	8	1.8%
Research Data Manager	8	1.8%
Librarian - Data Curation	6	1.4%
Librarian - Digital Preservation	6	1.4%
Digital Archivist	6	1.4%
Librarian - Scholarly Communications	6	1.4%
Research Data Specialist	5	1.1%
Librarian - Data Management	5	1.1%
Librarian - Geographic Information Systems	5	1.1%
Data Management Specialist	4	0.9%
Librarian - Social Sciences Data	4	0.9%
Librarian - Research Data Management	4	0.9%
Archivist	3	0.7%
Librarian - Digital Curation	3	0.7%
Data Manager	3	0.7%
Librarian - Digital Projects	3	0.7%
Data and Informatics Consultant	3	0.7%
Librarian - Research Data Services	3	0.7%
Data Archivist	3	0.7%
Research Information Scientist	3	0.7%
Other (255 unique titles)	290	65.8%

Percentage of Positions with Data Curation Responsibilities



- **Diverse array of titles**
- **54% “primary”**
- **46% “secondary”**

Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: Data Curation

Some descriptive characteristics

Countries	Count	Pct
United States	325	73.6%
Canada	42	9.5%
United Kingdom	41	9.3%
Australia	13	2.9%
Germany	7	1.6%
Ireland	3	0.7%
United Arab Emirates	2	0.5%
Hungary	1	0.2%
South Africa	1	0.2%
Singapore	1	0.2%
Sweden	1	0.2%
Egypt	1	0.2%
Netherlands	1	0.2%
Finland	1	0.2%
New Zealand	1	0.2%

Organization Types	Count	Pct
University	329	74.6%
University Library	43	9.8%
Research Center	27	6.1%
Government	21	4.8%
Other Organization	7	1.6%
Public Library	6	1.4%
Corporation	4	0.9%
Museum	2	0.5%
Not-for-profit	1	0.2%
Government Library	1	0.2%

Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: Data Curation

Incidence of certain characteristics and responsibilities

Characteristic	Count	Pct
Instruction	299	67.8%
Reference	290	65.8%
Outreach	270	61.2%
Access	258	58.5%
Preservation	254	57.6%
Data librarian	249	56.5%
Policy	238	54.0%
Data manager	208	47.2%
System design	187	42.4%
Research support	164	37.2%
Scholarly communication	162	36.7%
Best practices	137	31.1%
Lifecycle	115	26.1%
Domain expertise	115	26.1%
Open Access	111	25.2%
Statistics	102	23.1%
Data management plan	93	21.1%
Rights	81	18.4%
Data services librarian	70	15.9%
Steward	44	10.0%
Data science	43	9.8%

Note: Not natural language; count denotes incidence of term and/or broader concept.

Please note: all analyses and results are preliminary and may change as our research advances

Content Analysis: Data Curation

Incidence of degree qualifications

Among all data curation positions

Degree	Count	Pct
MLIS	119	27.0%
PHD	34	7.7%
BA/BS	13	2.9%

	Data librarian		*Data manager*		*Data scientist*		*Data service librarian*	
Degree	**Count**	**Pct**	**Count**	**Pct**	**Count**	**Pct**	**Count**	**Pct**
MLIS	86	34.5%	60	28.8%	11	25.6%	30	42.9%
PHD	16	6.4%	16	7.7%	4	9.3%	6	8.6%
BA/BS	5	2.0%	9	4.3%	5	11.6%	1	1.4%

Please note: all analyses and results are preliminary and may change as our research advances

PHASE III: INTERVIEWS

Interviews

- Participant recruitment
  - Post to the IFLA listserve
  - Convenience sampling
  - Snowball sampling
- 24 Interviews
  - Conducted over Skype, Zoom, or phone
  - Lasted approx. 30 - 60 min.
  - Recorded
  - Transcribed
- Additional data collection techniques
  - Questionnaires
  - Documents
- Data analysis
  - Debriefing
  - Memos
  - Coding

Codebook – 19 Groups

	Group Code	Group Label
2	COMMUNITYBUILDING	Community Building Activities
3	DATACURATIONSERVICE	Data Curation Services
4	DATATECHSERVICE	Data Services
5	DATATYPE	Data Types
6	ACADEMICDOMAIN	Domain
7	EDUCATIONLEVEL	Education Level
8	EXPERIENCE	Experience
9	INFRASTRUCTURE	Infrastructure
10	LIBPUBSERVICE	Outreach and Training Services
11	NATIONALCHARACTERISTICS	National Characteristics
12	ORGANIZATIONCHARACTERISTICS	Organizational Characteristics
13	MOTIVATION	RDM Motivations
14	RESEARCHSUPPORTSERVICE	Research Support Activities
15	CLIENTCHARACTERISTICS	Researcher Characteristics
16	CHALLENGES	Role Challenges
17	ROLECHARACTERISTICS	Role Characteristics
18	COMPETENCE	Role Competencies
19	UNITCHARACTERISTICS	Unit Characteristis

Codebook – 178 Codes

Code ID	Code	Index	Group Label	Code Label
102170	SOCIALSCIENCE	1	Domain	Social Science
102453	PHYSICALSCIENCE	2	Domain	Physical Sciences
102429	HEALTHSCIENCE	3	Domain	Health Science
102426	GEOSCIENCE	4	Domain	GIS/Geo Sciences
102171	HUMANITIES	5	Domain	Humanities
102417	ENGINEERING	6	Domain	Engineering
102416	EMERGENT	0	Role Challenges	Emerging and Evolving Trends
100134	ORGANIZATION	1	Role Challenges	Organization Policy and Support
102422	EXPERIENCE	2	Role Challenges	Lack of Experience
102421	EXPECTATIONS	3	Role Challenges	High Expectations
102189	COMPLIANCE	4	Role Challenges	Researcher Compliance
102469	SHARING	5	Role Challenges	Community Resistance
100011	TIME	6	Role Challenges	Time Constraints
102397	AWARENESS	7	Role Challenges	Community Awareness
102455	POWER	8	Role Challenges	Organization Power Dynamics
102448	NATIONAL	9	Role Challenges	National Policy and Support
102447	MOTIVATION	10	Role Challenges	Unmotivated Researchers
102449	NEED	0	Researcher Characteristics	Perceived Need
102434	JUNIOR	1	Researcher Characteristics	Junior Faculty
102223	FACULTY	2	Researcher Characteristics	Senior Faculty

Interview Participants

- Participants' background
  - 24 Interviews with 26 professionals (two interviews included a team of two people)
  - 26 participants - 11 female and 15 male
  - Most participants had several years of professional library or research experience
  - All have Masters degrees; 15 Masters in Library and Information Science (MLIS)
  - Ten PhDs in a variety of disciplines, including biology, environmental science, history, information science, medical informatics, or philosophy

Institution Type	No. of Sites
University library	17
Campus-wide research data service center	3
University department - embedded service	2
Data archive	1
Research center	1
Total	**24**

Countries

Country	No. of Sites
Australia	3
Austria	1
Canada	3
Germany	2
Netherlands	2
Sweden	2
Switzerland	1
United Kingdom	5
United States	5
Total	24

Position Titles

Position Title	No.
Research Services / Research Data Management / Digital Curation / Data Curation / Research Data **Coordinator**	6
Data Curation / Research Data / Research Data Management / Data **Librarian**	5
Project Manager Research Data Management / Research Data **Manager**	4
Head of Department / Head, Research Data Services/ Deputy Head	3
Project Scientist / Project Coordinator	2
Data Curation Scientist	1
E-research Project Officer	1
Humanities Data Curator	1
Principal Librarian	1
Research Data Officer	1
Business Developer	1
Total	**26**

Emergent Field

- Newly created positions (most cases in our sample)
  - Offer some flexibility
  - Require to develop a program from the ground up
- Primary responsibilities in the early phase
  - Needs assessment
    - Conduct studies to learn about researchers' practices and needs for data management
  - Policy development for open access and institutional data management
  - Infrastructure development
    - Participate in selecting or developing a data repository
  - Outreach to researchers
    - Faculty and graduate students
    - Faculty committees

"In the beginning, lots of time on outreach. Heaps of time." Participant D

Outreach and Advocacy

- Primary responsibility for all participants
  - Between 40% to 90% of time devoted to outreach and training
  - More time in the early phase but outreach remains a key responsibility for data curators with more than 2 years of experience
- Outreach efforts focused on:
  - Raising research data awareness
  - Informing about research data services
  - Educating about good data management practices
  - Promoting open access and data sharing
  - Building community around research data management (RDM)

"My role is very much … it's almost like a salesman role. I'm pretty much pitching the idea to researchers, first of all, the concept of open data and then secondly, the idea of opening it up through Figshare" Participant B

Roles

"The role of the data librarian is more about helping them [researchers] to describe data, getting it online, helping them to share it, that sort of thing."
Participant B

"My role is to provide organizational **leadership** on data issues, **collaborate** within and external to the organization, conduct independent research, **engagement** with the [center's] staff to promote and enable effective **data and metadata management**."
Participant G

Consultation Services

- All participants were engaged in consultative / informational services
  - Early in the data life cycle
- Forms of services
  - One-on-one consultations
  - Workshops and seminars for faculty and graduate students
  - Online tutorials and guidelines
  - Embedded in the projects or research labs (rare)
- Guidance and support in:
  - Compliance with funders' requirements
  - **Data management plans (DMP)**
- Other areas include:
  - Data management best practices
    - Data formats
    - File naming
  - Metadata
  - Data release, sharing, and re-use
    - Data citation standards
  - Digital preservation and archiving

Training Workshops

"The content of the workshops is usually on data management best practices and I'm using DataONE educational modules which I've adopted and customized for my university, so it deals with everything along the lines of the data life cycle model that DataONE has developed **from planning to metadata, description, quality control, preservation, access, and re-use**. So I'm talking about things like file name conventions, file versioning, and metadata, how to use tools for data entry, and how to minimize errors and increase data quality, and I talk about different file preservation formats. It's about an hour and a half workshop where I pretty much cover the entire data life cycle." Participant H

“Curating People”

“When you hear the phrase data curation, you think about … the act of sitting down at your computer and opening a data file, and maybe writing some documentation, or running the script to make sure that everything is doing what is supposed to do, documenting what you did, and all this stuff, this kind of work as a data curator, maybe takes place in a lab or maybe in a research center even, […]

but when it’s at a campus level like this, data curation is more about **providing information about good data curation practices to the people who need to curate their data or could be curating data**.”

Participant E

Technical Services

- Technical infrastructure and the level of support depend on institutional settings
  - Limited number of participants were involved in offering technical services
  - Usually at the end of data life cycle
- Support in
  - **Data management**
    - Data formats and file naming conventions
    - Data cleaning and verification
    - Data conversion
  - **Data description and documentation**
    - Metadata creation using standardized schemas
    - Data linking
  - **Data deposit / publishing**
    - Ingest into repository systems
    - Assigning identifiers
    - Data anonymization
    - Data security
  - **Archiving and preservation**

Changing Research Culture

"When we're trying to raise awareness, we really try to stress the benefits of data management as well. And it's that balance, really. Trying to get the right balance between making it known that funders require this, certain journal publishers require data preservation, open access and so on. So we want to get that message across but at the same time we also want to try to change the research culture. […] That's really what we want to be leading to, it's not just about compliance but actually trying to **change research culture and get people to think it's good research practice**." Participant V

Evolving Profession

- Encompasses technical and public services skills
  - Data librarians
  - Data managers
  - RDM coordinators
- Requires expertise in multiple areas of data creation, management, publishing, and preservation
- Plays a role of an intermediary / “translator”
- Involves collaboration and interdisciplinary work
  - Building bridges between a library, IT unit, campus departments, or specialized centers

“It's almost like multiple jobs in one really. I'm a technical services librarian, as well as an outreach librarian.” Participant C

International Context

- Terminology
  - Different understanding of terms: data curation, data stewardship, and research data management
- Different policy and funding contexts across countries
  - Institutional, national, or regional policy development
  - Funding and requirements for depositing data
  - Open access advocacy
- Infrastructure
  - Data repository systems
  - The significance of a well-developed digital library / repository infrastructure

"Australia has a very strong commitment to research data management and, things have been moving swiftly over the past couple of years, partially because of a big amount of funding put in by the federal government." Participant C

International Context

- Different models of services and levels of collaboration
  - University library
  - A university-wide network of research data experts
  - Research Data Service (RDS) centers
  - Embedded in the faculty-led research projects, research labs, or departments throughout the university

"We want to put data stewards, […] in each of the faculties so that we have experts in research management who are not sitting in the library but are actually sitting in faculties and understand the subject specific issues of data in their fields" Participant P

Conclusion

- *R1: How is data curation defined by practitioners / professional working in the field?* The concept is evolving
- *R2: What terms are used to describe the roles for professionals in data curation area?* RDM is the preferred term
- *R3: What are primary roles and responsibilities of data curators? Main* roles and responsibilities of data curators are identified
- *R4: What are educational qualifications and competencies required of data curators?* Differences in concepts and terminology should be further investigated

THANK YOU!

Anna Maria Tammaro, University of Parma
annamaria.tammaro@unipr.it

Krystyna K. Matusiak, University of Denver
krystyna.matusiak@du.edu

Frank Andreas Sposito, University of Denver
frank.sposito@du.edu

Terry Weech, University of Illinois Urbana-Champaign
weech@illinois.edu