Data Curation Education: Cross-Disciplinary Analysis of Master's Programs

Ayoung Yoon
Indiana University Indianapolis, Indianapolis, Indiana, United States

Angela P. Murillo
Indiana University Indianapolis, Indianapolis, Indiana, United States

Thomas Jettpace
Indiana University Indianapolis, Indianapolis, Indiana, United States

With growing emphasis on data curation practice in both science and industry, there has been a call for information professionals to take on a substantial role in data curation. Library and information science (LIS) education has been responding to this call by offering various training opportunities from Master's education to professional development. The most recent effort to systematically review a data curation curriculum offered by ALA-accredited LIS schools was in 2012, so it is time to revisit the progress and evolution of data curation education. The main goal of this study is to analyze the course content from the syllabi of various programs to understand what is being taught in LIS schools throughout graduate-level education. Further, because the need for data curation is apparent across different disciplines, and thus not only LIS but also other disciplines have been offering data curation courses, this study also analyzed syllabi from other disciplines. A total of 80 syllabi were analyzed in this study: 15 syllabi from 9 ALA-accredited institutions and 65 syllabi from 53 institutions of Carnegie Classification (CC). Our findings suggest a notable growth in LIS education in data curation since 2012, but LIS education still provides less training in technical skills. There was also a distinctive difference in educational approach to teach data curation between LIS (user- and service-oriented) and other disciplines (technical skills—focused), which brought different strengths and weaknesses in curriculum.

Keywords: data curation education, Master's education, syllabi analysis

Commonly defined as the ongoing processing and maintenance of data throughout its lifecycle to ensure long-term accessibility, sharing, and preservation (NLM, n.d.), data curation has become more critical in recent years, with increased volume, popularity, and emphasis on public access to existing data (Palmer et al., 2013). Contributing to data management throughout its lifecycle, data curation makes data easily findable, retrievable for future research, and useful for end users while adding value through annotations, merging, and cleaning. Many have argued for the importance of well-curated data in science and research, noting their vital function within scientific practice across different analytic methods and techniques (Gold, 2010) and the part they play in enabling new discoveries and fostering interdisciplinary research (Witt et al., 2009). Data curation is important not only in many science disciplines with data-driven research, such as physics and medicine, but also in disciplines with a long history of data sharing (e.g., in the social sciences), as well

© Association for Library and Information Science Education, 2023
Journal of Education for Library and Information Science 2023
Vol. 64, No. 4 DOI: 10.3138/jelis-2022-0025

KEY POINTS:

- A significant distinction exists between the field of LIS (library and information science) and other disciplines in the realm of teaching data curation.
- LIS emphasizes a strongly service-oriented approach, whereas other disciplines tend to prioritize technical skills.
- A shared deficiency in both LIS and other domains pertains to the absence of content relevant to certain soft skills, such as project or workflow management, as well as interpersonal and professional communication.

as in some disciplines that have recently started responding to data curation initiatives (e.g., in the humanities). Recognizing that much existing data have not been properly curated (Heidorn, 2011), most federal funding agencies now require data management plans in grant proposals, which is the first step toward a full curation lifecycle.

There is also a growing awareness within the data industry of the increasing importance of data curation and the variety of its implications for business, as harnessing and leveraging data to understand and solve the most critical business problems is a top priority of many organizations today. In business contexts, data curation first acts as a bridge, by facilitating the process of collecting and managing the data so that the various stakeholders can make use of these data in their respective ways. Second, data curation systematically organizes the data that are created so that data analysts and scientists can make use of these data to derive insights that business can then leverage. Lastly, data curation ensures quality so that data analysts are able to trust the data provided to them (ProWebScraper, 2019). The term *data management* has been more commonly adopted and used within the business context, emphasizing the "multiple disparate functions and systems [that work] together to move, organize, and secure data such that it is accurate, precise, accessible and protected" (Uzialko, 2022). Uzialko (2022) argues for the importance of data management in leading companies to successfully implement business-enabling strategies by harnessing large amounts of enterprise data for decision making.

With this significance and emphasis on data curation practice in both science and industry, there has been a call for information professionals to take on a substantial role in data curation. As a response, in the past decades, a growing number of library and information science (LIS) graduate programs in the United States have launched new training opportunities to meet the science and industry needs for data curation professionals (Fulton et al., 2011; Gold, 2010; Heidorn, 2011; Ray, 2009; Walters, 2009; Yakel, 2007). Since this growth began, a review of curriculum offered by ALA-accredited LIS schools was performed in 2012 by Harris-Pierce and Liu (2012). While they reported that the increasing number of schools started offering courses in data curation as a special topic, it is critical to revisit the data curation curriculum to check the progress and evolution of the programs. Specifically, it would be worthwhile to check how some of the findings from Harris-Pierce and Liu have been changing since 2012: for example, how the special topics courses have been upgraded to regular courses and how those courses have been keeping pace with the burgeoning need for skilled professionals to manage the so-called "data deluge" (p. 611).

Given the significance of data curation education, the overarching goal of this study is to analyze the course content from the syllabi of various programs to understand what is being taught in library schools throughout graduate-level education. Further, because the need for data curation is apparent across different disciplines, and thus not only LIS but also other disciplines have been offering educational opportunities in data curation, this study also analyzed syllabi from non-LIS courses. While syllabus analysis is one way of examining educational content and approach, major topics, integrated technology, and aimed skills to achieve, the comparison between LIS education and other disciplinary approaches will highlight the uniqueness that LIS education can bring to data curation while revealing areas for future collaboration across disciplines.

Literature review

Types of educational efforts

Perhaps the most common educational avenue for preparation for data curation is a Master's degree. It is generally agreed that data curation falls within the confines of digital librarianship and thus within MLIS/LIS/IS programs in higher education institutions (Heidorn, 2011). Within the iSchools and ALA-accredited library science programs, there are a variety of standout programs in data curation, such as those at the University of Illinois at Urbana-Champaign (UIUC), the University of Arizona at Tucson, Clayton State University, Arizona State University at Phoenix, Syracuse University, the University of Michigan at Ann Arbor, the University of Tennessee at Knoxville, the University of North Texas, and the University of North Carolina at Chapel Hill (UNC-Chapel Hill) (Botticelli et al., 2011; Committee on Future Career Opportunities and Educational Requirements for Digital Curation, 2015).

However, this is not the only formal academic route to data curation, nor is it the only type of program that teaches data curation skills, which demonstrates the interdisciplinary nature of data curation and the broad spectrum of needs of data curators. Some examples of alternative programs are programs in general data science, chemistry, museum studies, and nursing (Reisner et al., 2014; Tibbo & Duff, 2008). For example, in examining the Doctor of Nursing Practice (DNP) program at Johns Hopkins School of Nursing, Sylvia and Terhaar (2014) found many elements of applied data curation within the curriculum. Similarly, Virkus and Garoufallou (2019, 2020) found that, while data science is most often not found within MLIS/MLS/IS programs, the skills taught in data science programs have some overlap with those required for data curation, such as metadata management, data storage and preservation, and data quality. Furthermore, Wang (2018) described how data science and information science taught in LIS programs have begun to overlap, such as the offering of data science specializations within MLIS programs (e.g., University of Illinois, Indiana University in Bloomington), the offering of specific courses such as Introduction to Data Science and Visualization (e.g., University of Washington), and graduate-level programs in data science (e.g., Syracuse University). Wang describes how the skills required for data science and information science overlap with and complement each other, such as the need for data librarians to have both data curation and data science skills.

There are also several professional development options for attaining the skills and proficiencies required of data curators. Some professional development programs are associated with government agencies, such as the Library of Congress Digital Preservation Outreach and Education Program or the Australian National Data Service (Gold, 2010). Others are associated with professional associations and non-profit organizations, such as the Science and Technology Section of the Association of College and Research Libraries (ACRL) Division of the American Library Association (Committee on Future Career Opportunities and Educational Requirements for Digital Curation, 2015). Other continuing education opportunities are also offered by higher education institutions, such as the University of North Carolina at Chapel Hill or Purdue University at West Lafayette (Gold, 2010; Keralis, 2012). Lastly, many professional organizations provide free data curation guidance through case studies, reports, and other online material that can be utilized by both data curators and data curation educators (e.g., Digital Curation Centre, Research Data Alliance, Data Curation Network, etc.).

Skills and competencies

Not all the literature precisely describes the necessary skills and competencies using concrete terms, but, in general, previous studies have listed some skills and competencies that are required for data curation in two types: (1) technical or "hard" skills and (2) "soft" skills. While those types are not mutually exclusive, as some skills are combination of both traits, it helps to understand the different nature of skills. We followed the definitions proposed by Heery and Noon (2017), who describe "hard skills" as competencies about specific job-related technical abilities, which are relatively easy to measure and are often validated with some form of qualification, whereas "soft skills" are competencies associated with activities such as service handling, communication, problem solving, and teamwork.

The hard skills cover a large variety of topics, ranging from performing quite specific (but situationally dependent) tasks, such as cleaning data or developing metadata schemes, to using and troubleshooting various technologies, such as databases or XML (Bishop et al., 2020; Carlson et al., 2011; Chen & Zhang, 2017; Johnston et al., 2018; Madrid, 2011). Relatedly, there are also many "soft" skills and proficiencies that aid in data curation, such as project management, professional communication, knowledge of research best practices, and administration (Bishop et al., 2020; Carlson et al., 2011; Chen & Zhang, 2017; Kennan, 2016; Kim et al., 2011; Lankes et al., 2008; Lee & Stvilia, 2017). Kim et al. (2011) found that there are six major types of duties for data curators in which these skills are required or performed: collecting primary data, collecting secondary data, storing data, managing data, analyzing data, and presenting data. Palmer et al. (2014) found that there are three types of major duties conducted by data curators: technical duties (i.e., preserving and analyzing data), service duties (i.e., providing training, assisting with data management plans), and administrative or managerial duties (i.e., overseeing collections, developing policies). Table 1 presents a comprehensive (though not exhaustive) list of the various hard and soft skills and competencies expected of data curators. As noted, several soft skills are associated with technical understanding of data (e.g., data citation, data discovery services,

Table 1: Summary of skills and competencies for data curators

Hard skills for data curators	Soft skills for data curators
APIs	Ability to adapt to changing situations
Basic database skills	Ability to learn
Architecture and operation of data networks	Ability to serve diverse people
Basic IT skills	Ability to understand user behaviors
Cataloging standards	Ability to critically evaluate and synthesize relevant data sources*
Chain of custody	Ability to understand many aspects of the data lifecycle*
Data analysis	Ability to understand the diversity, size, and complexity of data sets*
Data cleaning	Advocacy
Data conversion	Capability building
Data formats	Data citation*
Data gateways/portals	Discovery services*
Data visualization	Embargo
File audit	Ethics
File format transformations	Fundraising
Full text indexing	Grant writing
HTML authoring tools	Interpersonal communication skills
Integrated library systems	Knowledge of policy
Master files and versioning	Liaison and support
Metadata	Negotiation
Migration	Professional communication
Natural language processing	Program development
Ontologies	Project management
Operating systems	Publishing preferences
Persistent identifiers	Research and trends
Preservation	Technical writing*
Programming savvy	Training of others
Quality assurance	Workflow management
Rights management	
Risk management	
Secure storage	
Sensor networks	
Software installation	
Standardizing documentation process	
Statistics	
Terms of use	
User-centered design	
XML	

data evaluation). These skills are marked with an asterisk (*), which reveals a close link between the soft and hard skills required to perform the role of data curators.

Limitations in previous efforts

Despite the evident need for highly skilled and professionally competent data curators, various studies have argued that there are still gaps in the current educational options for preparing students for career outcomes. To start with, there are very few dedicated data curation programs in general (e.g., data curation certificate) from a few institutions such as the University of Arizona, the University of Illinois at Urbana-Champaign, the University of North Carolina at Chapel Hill, and San Jose State University (Harris-Pierce & Liu, 2012; Keralis, 2012; Ortiz-Repiso et al., 2018). This extends to the general MLIS curricula, where data curation courses and content are still relatively rare despite the recent growing interest. While Varvel et al. (2012) found that only 8 percent of MLIS courses are "data centric," with another 11 percent being "data inclusive" (p. 528), it is not known how those MLIS programs have changed since then.

Previous studies have also reported a limitation within programs that specialize in data curation. One prominent issue that appeared in previous studies was a lack of content focused on technical competencies. Palmer et al. (2011) and Botticelli et al. (2011) both underscored the capability of working with technical infrastructure as an essential competency, at least "[to] be able to have an intelligent conversation with a technologist . . . [and] accomplish basic curation tasks, including web markup, simple database design and queries" (Botticelli et al., p. 159). Still, educators face a tough dilemma when it comes to teaching technical skills, as specific technologies will come and go (Botticelli et al.). Regardless, however, graduates need to be prepared to perform the job when they graduate, and there is a large sentiment that the teaching of technical skills is not happening (Goodsett & Koziura, 2016; Tammaro & Casarosa, 2014; Thomas & Urban, 2018). Another frequent issue is a lack of professional and applied experiences within data curation programs (Goodsett & Koziura, 2016; Thomas & Urban, 2018; Thompson et al., 2013). Further, previous studies also pointed out that programs may not be preparing their curricula to address future changes and challenges in the field, such as those related to culture, policy, and technology (Thomas & Urban, 2018).

Methods

This study employed the content analysis of course syllabi as a primary method for analysis. Content analysis is known as an observational research method, which is used to systematically evaluate the content (Kolbe & Burnett, 1991, p. 243), and thus it is appropriate for our research purpose, which is to examine texts that are not subject to the influence of interests.

Sample and data collection

In order to identify the data curation course syllabi taught in LIS schools, the project team first created a list of ALA-accredited library programs. To review data curation courses from other disciplines, we chose universities listed in the "doctoral university—highest research activity" category on the Carnegie Classification (CC) website. The Carnegie Classification is a classification framework developed by the Carnegie Commission on Higher Education, first published in 1973. Institutions are classified as R1 if they provide research/scholarship doctoral degrees and have very high research activity (American Council of Education, n.d.).

A total of 61 universities from the ALA list and 115 universities from the CC list were initially selected between fall 2019 and spring 2020. We visited each school's website to search for relevant courses. We manually reviewed the course catalog and Office of the Registrar, and also conducted a keyword search using the terms *data*, *data curation*, *data management*, and *curation/management*. While data curation and data management are not exactly the same, our goal was to identify courses relevant to data curation, and thus our search was kept as broad as possible to capture courses that might teach data curation concepts and practices while not using the term in the course title. For the same reason, we included the terms *curation* or *management* in our search to capture courses that might include data curation components, such as digital curation, which have been taught in ALA schools. To collect comparable data, we excluded any undergraduate courses appearing in the CC list. When a course was cross-listed for both undergraduate and graduate students, we included it. When a university had more than one campus, we reviewed only the courses offered by the main campus.

From the initial search, we collected information about the names and locations of institutions, URLs of the websites, whether the school or department had a course on data curation, the course title, availability of syllabi, and the course description. Then we reviewed the course descriptions and removed irrelevant courses, including courses relevant to databases (e.g., database management, database systems), broad data science (e.g., introduction to data science), and data analytics (e.g., data visualization, advanced statistics). We also excluded digital curation courses (e.g., introduction to digital curation), as many courses focused on cultural records (e.g., manuscripts, archives) and only one or two modules addressed data-specific topics. The final list of selected courses comprised 27 courses from the ALA list and 195 courses from the CC list. We also compared any overlaps between the two lists, as ALA-accredited schools were often listed in the CC list. When one library program appeared in both lists, depending on their associated program, courses were placed on the ALA and CC list. For example, if a course was delivered within an ALA-accredited program, it was added to the ALA list; if not, then it was placed in the CC list if it fit the CC criteria so that we could compare the ALA and CC approaches without any overlaps. From the final list of relevant courses, we collected the available syllabi from the web and also contacted the instructors when syllabi were not available online. The final number of courses available and used for the study was 80. See Table 2 for our screening and selection process.

The courses met the criteria if the majority of the topics specifically focused on the curation of data. The nine institutions in the final ALA analysis list include some "stand-out" ALA programs that previous studies discussed, such as UIUC, Syracuse University, and UNC-Chapel Hill, but also include other programs such as UCLA, the University of Denver, the University of British Columbia, Dalhousie University, and Indiana University Purdue University Indianapolis. Data curation—related courses at the CC institutions were

Table 2: Number of institutions and courses selected for the study

	Reviewed		Identified	Removed	Remaining	Syllabi available
ALA	Institutions	61	25	7	18	9
	Courses		44	15	29	**15**
CC	Institutions	115	88	8	80	53
	Courses		329	160	169	**65**
Total number of courses			373	175	198	**80**

housed in various departments, including business, statistics, sociology, healthcare, nursing, computing, data science, and even specific domain sciences such as atmospheric science, earth science, and anthropology.

Protocol development

We adopted the protocol used in Yoon et al.'s (2021) study, which analyzed digital preservation course syllabi. The protocol worked well for the syllabi of both the ALA schools and the CC schools, and only slight modifications were made in this study context. The protocol consisted of two parts: course basics (e.g., course title, frequency, mode of offering, prerequisites, etc.) and course content (e.g., course objectives, course topics, technology and tools used, textbooks, assignments, collaborative activities, etc.)

Data analysis

The research team coded the identified syllabi, and the inter-coder reliability among the team members was 94.8 percent, which was within the accepted range. Three different analysis methods were used to analyze the syllabi fully. First, descriptive analysis was performed using SPSS in regard to the course basics. Some textual data were analyzed manually using Excel. Second, we employed a topic modelling technique to identify key themes from the syllabi, including themes from the course topics and course objectives.

Key themes were determined from analyzing course topics and course objectives using topic modeling. The course topics were gathered from the schedule section of the course syllabus through headings that indicated which topics would be covered that week. Course objectives were gathered in the course learning objectives portion of the syllabus. The text from course topics and course objectives was extracted and placed manually into a Microsoft Excel spreadsheet. Some preprocessing of the data was conducted in the form of spell and grammar checking and updating terminology for consistency (i.e., data sets vs. datasets).

Topic model analysis was then conducted on four corpora. The four corpora consisted of the text from the following:

1. ALA syllabi course topics
2. CC syllabi course topics

3. ALA syllabi course objectives
4. CC syllabi course objectives

Since we were interested in understanding the major themes in the topics covered and the course objectives, we utilized topic modelling to analyze the four corpora, as it is challenging to determine themes from large-scale textual data. Topic modelling has "proved useful for analyzing and summarizing large-scale textual data" (Song & Ding, 2014). We used Latent Dirichlet Allocation (LDA), which is a "generative probabilistic model for collections of discrete data such as text corpora" (Blei et al., 2003, p. 993) and produces proportions of topics within documents. The concept behind LDA is that "one document contains multiple topics, and each topic requires specific words to describe it" (Song & Ding, 2014, p. 235); therefore, the observed variables are words in the documents, and the hidden variables are topics.

The probability of generating a word w from a document d is

$$P(w|d,\theta,\phi) = \Sigma_{z\in T} P(\mathrm{w}|z,\phi_2)\, P(z|d,\theta_d)$$

The likelihood of a document d is defined as

$$\mathrm{P}(Z, W|\Theta, \Phi) = \Pi_{\mathrm{d}\in\mathrm{D}}\Pi_{z\in T}\theta^{ndz} \times \Pi_{\mathrm{z}\in\mathrm{T}}\Pi_{v\in\mathrm{V}}\phi^{nzv}$$

We employed the Stanford Topic Modelling Toolbox (TMT) (Stanford Natural Language Processing Group, n.d.) to conduct the topic modelling analysis on the text corpora extracted from the course topics and course objectives. The TMT software works well with .txt and .csv files through a simple Java interface. TMT provides a series of tokenizers to prepare the data for analysis and the ability to create custom stop-word lists for the most ideal text corpus for analysis.

Additionally, the software prepares perplexity calculations for corpora to determine the most optimal number of topics. We ran perplexity calculations for the four corpora. In the majority of the cases, five topics were the optimal number; however, in two instances, ten was the optimal number, with five closely following. Since the calculations were very similar between five topics and ten topics, we created five topic models for each corpus, which helps maintain consistency for analysis and understanding of the themes.

The software provides the option to explore several topic model methods, including a Collapsed Variation Bayes Approximation and the Collapsed Gibbs Sampler. Both methods were tested, and the Collapsed Gibbs Sampler produced clearer results, likely due to its ability to handle shorter text documents more accurately (Yin & Wang, 2014). The TMT output provided the top terms per topic and the probability of each topic and term. For the analysis and ease of interpretation, we changed the probabilities into percentages and listed each term in order of probability.

For topic modelling, human interpretation and judgment are used to determine the theme of each topic. Therefore, the researchers interpreted the major themes of each topic model through the examination of the terms for each topic, the likelihood of those terms, and the meaning of the combination of terms. They used an iterative approach to determine the themes and titles for each topic model, individually examining the terms and topics, determining topic themes, and sharing their interpretations with each other. Through this analysis and extensive discussions, the researchers determined the topic meaning and topic title.

Results

Among the 61 institutions that we reviewed from the ALA list, about one-third (18 universities, 29.5%) offered courses relevant to data curation and management. This trend was opposite in CC schools, as about 70% of the institutions (53 universities, 69.5%) that we reviewed offered courses on data curation or management. The institutions also differed in terms of the levels of courses and the depth of coverage, as several institutions offered more than one course.

Course basics

The majority (87%) of courses from ALA schools used the term "data curation" or "data management," or both, in the course title, and all used the term "data," with variations of "open data," "data administration," and "data practices." A few courses indicated that they had a specific focus area, such as "research data" or "policy-based data management." In CC schools, 97% of the courses had the term "data management" in course titles, but they were often combined with other terms, such as those that specify disciplinary or types of data (e.g., biological data, public health data, spatial data, energy data, business data, qualitative data; 20%), terms that specify the process (e.g., analysis, preparation, and processing; 14%), and terms specifying the statistical skills or languages (e.g., programming, SAS, STATA; 15%). As described previously, the CC courses were housed in various departments, and the course title diversity reflects these domains and approaches to data curation.

The majority of courses relevant to data curation were at the introductory level, as stated in either the syllabus or course description (73% for ALA schools and 86% for CC schools). Only a small number of courses required a prerequisite to take the course for ALA schools (26%), and the prerequisite requirement was only for advanced data curation courses in those schools. However, 40% of courses from CC schools required one or more prerequisite courses (none: 60%; one: 20%; two: 7%; more than two: 13%), which clearly reflects the different approaches and types of needed prerequisite knowledge for data curation/management education. While ALA schools' prerequisites were either relevant to the degree requirement (e.g., foundational courses) or the introductory data curation course, CC schools' requirements often involved not only introductory data curation courses but also statistics (e.g., introduction to statistics) or an introductory level of technical courses (e.g., introduction to programming, database management/systems), which clearly shows that technical data skills are more emphasized in the CC schools' approach and needed for student success in these courses.

For both types of institutions, a yearly offering of the courses was most common (67% in ALA schools; 62% in CC). However, the frequencies varied; 20% of ALA school courses were offered every semester, as opposed to 18% for the CC school courses, and 13% of

Table 3: Topic model percentages for the course topics corpora

ALA course topics	Percentage	CC course topics	Percentage
T1-ALA: Research Data Management	47.43	T1-CC: Database Design	32.96
T2-ALA: Metadata Management	18.24	T2-CC: Data Analysis and Statistics	21.84
T3-ALA: Data Services and Sharing	12.94	T3-CC: Data Quality Assurance and Quality Control	19.97
T4-ALA: Database Design	11.44	T4-CC: Programming	15.45
T5-ALA: Data Archives and Preservation	9.95	T5-CC: Data Processing	9.78

courses from ALA schools were offered every other year compared to 3% of courses from CC schools. The majority of courses were offered in a face-to-face mode (73% of ALA schools; 78% of CC schools). While further investigation is necessary, the nature of data education, which often involves a hands-on approach, may have influenced the mode of instruction.

More than half of the courses from ALA schools (53%) did not require a textbook, while more schools from the CC list required the use of textbooks (58%). Most textbooks used by ALA schools were about research data management, data sharing and reuse, and scholarship involving data, all in the context of libraries (e.g., Borgman, 2015; Corti et al., 2019; Krier & Strasser, 2014; Ray, 2014). The textbooks used by CC institutions were more technical (e.g., topics on data mining and database management), often statistical (e.g., Baum, 2016; Delwiche & Slaughter, 2003), or oriented toward data analytics in general (e.g., Kabacoff, 2015; Wickham & Grolemund, 2017), but sometimes also included research data sharing and management (e.g., Corti et al., 2019).

Course topics and objectives

Topic models provide a way to conduct large-scale thematic textual analysis by producing the probability of a word in documents. In order to examine the themes taught in the courses, we created five topic models for the following corpora:

1. ALA syllabi course topics
2. CC syllabi course topics
3. ALA syllabi course objectives
4. CC syllabi course objectives

As discussed in the methods, we created five topic models per corpus, based on perplexity calculations and for consistency in the reporting of the findings.

In the results below, Tables 3 and 6 provide a summary of the topics for each corpus, and Tables 4, 5, 7, and 8 provide the top 15 terms for each corpus. Additionally, Appendices A through F provide visualizations of all topic model findings, including topic model

Table 4: The top 15 terms in ALA course topics

T1-ALA: Research Data Management (47.43%)	T2-ALA: Metadata Management (18.24%)	T3-ALA: Data Services and Sharing (12.94%)	T4-ALA: Database Design (11.44%)	T5-ALA: Data Archives and Preservation (9.95%)
Data	Metadata	Data	Data	Data
Research	Management	Cleaning	Standards	Preservation
Management	Libraries	Curation	Access/ Microsoft	Access
Sharing	Formats	Lifecycle	Modelling	Selection
Curation	Software	Services	Databases	Types
Practices	Models	Visualization	Managing	Provenance
Libraries	Types	Analysis	Types	Appraisal
Researchers	Standards	Documentation	Systems	Ethics
Reuse	Design	Version	SQL	Archives
Collection	Requirements	Reproducibility	Ontologies	Properties
Sustainability	Data	Copyright	Model	Policy
University	Property	Licensing	Access	Approaches
Types	Digital	Publishing	Normalization	Repositories
Archives	Hardware	GitHub	Design	Managing
Repositories	Services	Open	Abstraction	Issues

distributions, term comparisons for ALA and CC terms, and the top 15 terms for each topic model (Figures A1—A4, B1—B4, C1—C5, D1—D5, E1—E5, and F1—F5).

Course topics

We produced five topic models for the two course topics corpora: ALA course topics and CC course topics. The probability percentages for each topic model of the two corpora are shown in Table 3. As can be seen in the table, for the ALA topics, research data management is nearly one-half of the focus of the course topics, while about one-third of the CC topics focus on database design. These findings demonstrate that while many LIS data curation courses focus specifically on research data, CC courses seem to focus on general data management.

When reviewing the ALA topics, four of the five topic models focused on typical data curation or data lifecycle topics, namely T1-ALA: Research Data Management (47.43%), T2-ALA: Metadata Management (18.24%), T3-ALA: Data Services and Sharing (12.94%), and T5-ALA: Data Archives and Preservation (9.95%). These data lifecycle topics comprise 88.6% of the ALA syllabi. The Database Design topic is the only technical topic in the ALA syllabi and makes up only 11.44% of the themes of the ALA's course topics.

Table 5: The top 15 terms in CC course topics

T1-CC: Database Design (32.96%)	T2-CC: Data Analysis and Statistics (21.84%)	T3-CC: Data Quality Assurance and Quality Control (19.97%)	T4-CC: Programming (15.45%)	T5-CC: Data Processing (9.78%)
Data	Data	Data	SAS	Data
Databases	R	Management	Data	Access/Accessing
SQL	Analysis	Records	R Studio	Reading
Relational	Stata	Systems	Proc	Processing
Modelling	Cleaning	Security	Creating	Creating
Query	Regression	Quality	Learning	Combining
Design	Statistics	Control	Macros	Mapping
Xml	Visualization	Plans	Output	Editing
Management	Sampling	Version	Statistical	Working
Apache	Models	Strategy	Ods	Conditional
Joining	Transforming	Governance	Merging	Functions
Normalization	Importing	Research	Summarizing	Displaying
Cassandra	Descriptive	Tools	Programming	Raw
Algebra	Text	Metadata	SQL	Excel
Model	Commands	Capture	Working	Sets

In order to examine the ALA course topics more closely, Table 4 provides the top 15 terms for each topic model in the ALA course syllabi corpus. As our analysis shows, ALA course topics are shaped around research data management and services to support data preparation for sharing and reuse. The terms *sharing* and *reusing* appeared under T1-ALA: Research Data Management (RDM). As metadata is a vital part of RDM, T2-ALA: Metadata

Table 6: Topic model percentages for the syllabi course objectives corpora

ALA course objectives	Percentage	CC course objectives	Percentage
O1-ALA: Research Data Management	27.26	O1-CC: Data Management Foundations	25.89
O2-ALA: Data Services	26.13	O2-CC: Database Design	22.81
O3-ALA: User and Community Focused	25.75	O3-CC: Data Analysis and Statistics	20.48
O4-ALA: Legal, Ethical, and Context Focused	14.85	O4-CC: Domain-Specific Data Management	17.09
O5-ALA: Technology Focused	6.02	O5-CC: Data Processing	13.74

Table 7: The top 15 terms in each topic for ALA course objectives

O1-ALA: Research Data Management (27.26%)	O2-ALA: Data Services (26.13%)	O3-ALA: User and Community Focused (25.75%)	O4-ALA: Legal, Ethical, and Context Focused (14.85%)	O5-ALA: Technology Focused (6.02%)
Data	Data	Data	Curation	Managing
Research	Research	Information	Issues	Use
Management	Collections	Social	Associated	Technical
Lifecycle	Services	Policy	Legal	Skills
Curation	Practices	Public	Evaluate	Practice
Process	Appraising	Communities	Ethical	Computer
Role	Scholarly	Management	Private	Techniques
Challenges	Knowledge	Strategies	Technologies	Technologies
Strategies	Professional	Use/using	Perspectives	Issues
Documentation	Policies	Context	Context	Products
Needs	Criteria	Physical	Roles	Context
Plans	Principles	Develop	Emerging	Strategies
Techniques	Assess	Needs	Activities	Services
Professions	Digital	Roles	Reuse	Settings
Practice	Forms	Trends	Types	Knowledge

Management had the second-highest probability and included terms such as *standards* and *formats*, indicating the focus on metadata. The T3-ALA: Data Services and Sharing topic also demonstrated the librarian's role in serving researchers' data needs and service-oriented approach when teaching data curation and management. The terms, including *services*, *licensing*, and *publishing*, are all relevant to data services and sharing of research data.

While T4-ALA: Database Design is the only specifically technical category in this analysis, metadata management also often relates to dealing with technical standards and software, as the terms *formats*, *software*, *models*, and *hardware* appeared in T2-ALA: Metadata Management. Lastly, preservation is an essential component of data curation, but surprisingly T5-ALA: Data Archives and Preservation had the lowest probability; however, this topic was clearly focused on long-term preservation by including terms such as *selection*, *appraisal*, *provenance*, and *repositories*.

When we compare ALA course topics with CC topics, clearly the dominant topics for the CC syllabi are the technical aspects of data work, such as T1-CC: Database Design (32.96%), T2-CC: Data Analysis and Statistics (21.84%), T3-CC: Data Quality Assurance and Quality Control (19.97%), T4-CC: Programming (15.45%), and T5-CC: Data Processing (9.78%). Notably, no topics in the CC syllabi specifically relate to research data management, data sharing, or data archives and preservation, which demonstrated a different approach to education data curation and management.

Table 8: The top 15 terms in each topic for CC course objectives

O1-CC: Data Management Foundations (25.89%)	O2-CC: Database Design (22.81%)	O3-CC: Data Analysis and Statistics (20.48%)	O4-CC: Domain-Specific Data Management (17.09%)	O5-CC: Data Processing (13.74%)
Data	Database	Data	Data	Data
Analysis	Data	SAS	Management	Perform
Management	Relational	Datasets	Health	Apply
Basic	SQL	Statistical/Statistics	Business	Programming
Tools	Design	Analysis	Research	Delivery
Query	Queries	Software	Social	Extract
Cleaning	Systems	Methods	Clinical	Techniques
Quality	Applications	Stata	Science	Define
Processing	Management	Access/Microsoft	Systems	Management
Manipulate	Modelling	Excel	Researcher	Utilize
Modelling	Structured	R	Information	Identify
Databases	Model	Variables	Evaluate	Interpret
Import	Build	Package	Activities	Code
Use	Technologies	Create	Healthcare	Describe
Hands-on	Unstructured	Results	Knowledge	Interpret

In order to examine the CC course topics more closely, Table 5 provides the top 15 terms for each topic model for the syllabi corpus of CC course topics. Clearly, the focus of CC courses is data manipulation and use (e.g., analysis) using tools such as programming languages. While the term *data management* does not appear as a main topic in our analysis, we found different elements relevant to data curation and management. T3-CC: Data Quality Assurance and Quality Control, is an important part of data curation, ensuring quality and trust in data for current and future users, as terms such as *security*, *control*, *version*, *governance*, and *metadata* are included. Data processing is also essential not only in supporting use/reuse, as the terms *access* and *accessing* can be found in T5-CC: Data Processing, but also in quality control, as relevant terms like *mapping* and *functions* are included.

Course objectives

While there may be overlaps between course topics and course objectives, we also examined course objectives, as they are often directly relevant to student learning outcomes while describing the goals and intentions of course instructors. We produced five topic models for the two corpora, ALA course objectives and CC course objectives. The probability percentages for each topic model of the two corpora are shown in Table 6. As can be seen, the topic models for each corpus were fairly evenly distributed. The ALA course objectives topics had the largest range in topic distribution (from 6.02% to 27.26%).

In order to examine the ALA course objectives more closely, Table 7 provides the top 15 terms for each topic model in the ALA course objectives syllabi corpus. For the ALA course objectives, O1-ALA: Research Data Management held the highest probability, which aligns well with the course topic analysis. The service-oriented approach to data curation and management education is well reflected in the analysis of course objectives, as O2-ALA: Data Services and O3-ALA: User and Community Focused together comprise most of the themes in the corpus, with a combined total of 51.88%.

It is interesting to note that one of the major topics for the course objectives was regarding legal and ethical issues (O4-ALA in Table 7), while in our ALA course topics analysis, legal and ethical issues were not the course topics (see Table 4). In fact, the terms *ethics* and *policy* appeared only once within T5-ALA: Data Archives and Preservation, but instructors included legal and ethical issues as a part of the course learning objectives. Educators' intention to address legal and ethical issues when teaching data curation is worthwhile, but the mismatch between course topics and course objectives suggests that it is still unclear how their intention has been incorporated into teaching practices, either through course topics or other learning activities. Finally, technology-oriented objectives (O5-ALA) showed the lowest probability, which also aligned well with the findings from the ALA course topic analysis, as the technical course topic had the fourth-highest probability.

In order to examine the CC course objectives more closely, Table 8 provides the top 15 terms for each topic model in the CC course objectives. Compared to the ALA course objectives, the CC course objectives do not have a topic specifically related to research data management. However, there is a topic related to O1-CC: Data Management Foundations (25.89%), which encompasses some of the research data management themes. Like the course topic analysis, it is also clear that CC course objectives emphasized data handling and manipulation and have similar topics as the course topics analyzed, such as O2-CC: Database Design (22.81%), O3-CC: Data Analysis and Statistics (20.48%), and O5-CC: Data Processing (13.74%). One new topic that emerged from the objectives analysis is O4-CC: Domain-Specific Data Management (17.09%), which did not appear in any of the course topics analysis, as a topic or as terms in each topic. This is not surprising, given that 20% of CC courses had domain-specific terms in their course title (e.g., *biological data*, *public health data*, *spatial data*, *energy data*, and *business data*), indicating that the course was developed within a specific context. Nonetheless, how these disciplinary contexts were embedded in teaching data management needs to be further investigated.

Assignments

We reviewed the assignments to get a better idea of the skills and knowledge the instructors expected students to learn. A total of 59 assignments from ALA schools and 126 assignments from CC schools were analyzed from the syllabi (see Figure 1). While this analysis is limited—as the depth of information about each assignment varied and the majority of the

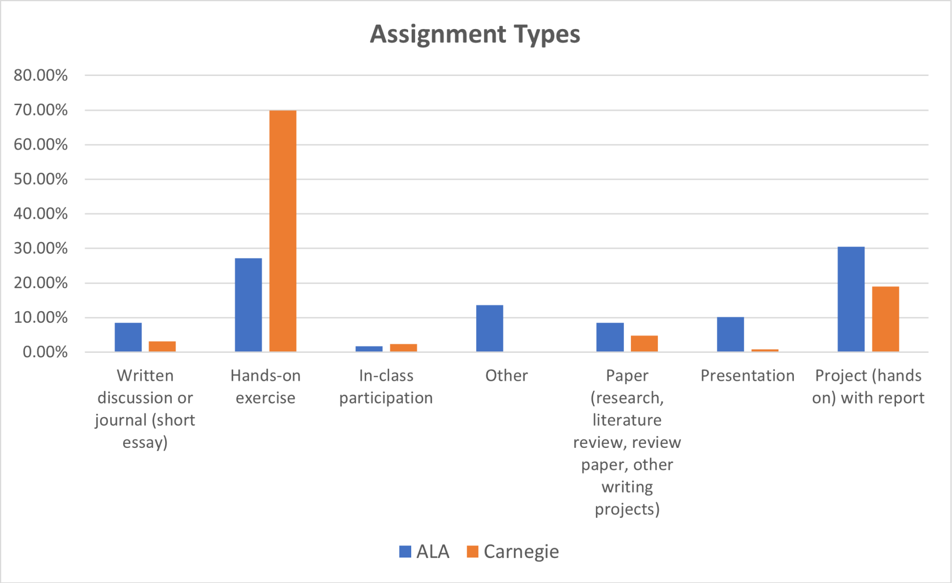

Figure 1: Types of assignments (ALA: $n = 59$, CC: $n = 126$)

syllabi did not provide details about the assignments—it still provides a snapshot of the nature of the assignments required for students.

Our analysis indicated that hands-on projects and exercises were the most-used types of assignments for both ALA and CC schools, although hands-on exercises were the dominant type of assignment for CC schools. It is not surprising to see that both types of schools placed a significant emphasis on hands-on work, as this type of assignment often provides an opportunity for students to work with tools and acquire hands-on experiences by adding a real-world context. Many projects in ALA schools concerned data management plans, which identify data management requirements and needs for stakeholders and propose a budget in a specific project context. Sometimes, the project required a meeting with a researcher to interview RDM requirements. Other exemplary projects from CC schools were more about practical data handling, such as analyzing data with real-world datasets, creating SQL queries, designing a data dictionary, validating data, performing data management, and conducting data analysis using various methods and techniques. Examples of hands-on exercises for ALA schools were often small components of data curation, such as creating metadata, building ontology, checking storage, and structuring and organizing data. However, hands-on exercises for CC schools required mostly data handling, such as importing and entering data, data merging and aggregating, checking data errors, building data models, and designing databases and queries. ALA schools tended to require more written assignments (either short essays or research/review papers) and presentations than CC schools, which may indicate that the course content in ALA schools may include more theoretical concepts than CC schools. The assignments in the "other" category for ALA schools were simply those without descriptions or titles.

Project-type assignments from CC schools tended to be more collaborative, as 47% of projects from CC schools (8 out of 17) were group projects, while only 27% of project-type assignments from ALA schools (3 out of 11) were team projects.

Technology

When we examined the tools and technologies integrated throughout the courses, mostly through the assignments but also through other types of learning activities, 45.7% (7 out of 15 courses) of ALA schools specified the tools and technologies that would be utilized in the courses, while 89.2% (58 out of 65 courses) of CC schools mentioned the tools in the syllabus. The nature of the tools used in the courses was distinctive, which explained the dominant number of courses requiring the use of tools in CC schools. Most tools utilized in the courses from CC schools were in five categories: general tools for data handling (e.g., Access, Open Refine), programs or languages relevant to statistical analysis (e.g., Stata, R, SAS), tools and languages relevant to databases (e.g., SQL, Access, Oracle, NoSQL, MySQL), tools that were big data—specific (e.g., Hadoop), and data visualization tools (MS Visio). SQL and Access were also used in the ALA schools' data curation course; however, data management tools developed in the field of library and information science were more commonly used, such as the DMP tool and the Data Curation Profile Toolkit. Tools for ontology creation (e.g., Protégé) and data repository or preservation infrastructures (e.g., Dataverse, iRODs, DSpace, GitGub) were other tools used in the courses from ALA schools.

Discussion

Our study examined data curation and management course syllabi from ALA and CC schools. Compared to the survey results of the ALA school's data curation curriculum conducted by Harris-Pierce and Liu (2012), the number of institutions that offered data curation courses seemed to increase. We used stricter criteria in our course sampling, focused solely on data, while Harris-Pierce and Liu's study included courses that covered a broader range of topics, such as digital collections and digital curation. Consequently, it is not possible to directly compare the number of courses offered by ALA schools. However, even with our strict course selections, we found more courses on data curation than in 2012, which indicates more institutions offered data-focused curation courses. This was confirmed when compared with Varvel et al.'s (2012) survey study on 55 LIS institutions' courses and programs for data professionals. While Varvel et al. identified 11 institutions that offered five programs specifically targeting data, our study identified 18 institutions that offered data curation courses, which is notable growth.

While data curation has become a topic of educational interest across different disciplines, our syllabus analysis suggests several interesting comparisons between ALA schools' and CC schools' educational approaches to teaching data curation and management courses. One notable difference is that ALA courses are designed in a highly service-oriented way, which means that the goal of education is to produce informational professionals whose role is to assist researchers with their data needs. RDM is particularly the focus of education, as the topic analysis indicated nearly 50% of the topics were relevant to RDM. It was also the course objective with the highest probability. Tools to support RDM were also utilized in ALA courses with a combination of database design tools. This is not surprising because many academic libraries have initiated RDM services in recent years and hired data librarians (Yoon & Schultz, 2017). Furthermore, this service-oriented approach is well reflected in the ALA course objectives analysis, as the results suggest a strong focus on user communities.

While ALA courses are more service-oriented and user-centered, CC courses focus more on the technical skills relevant to data manipulation and handling, seeing students as future data scientists or analysts who will perform hands-on data work while possessing data management skills with domain-specific expertise. As a result, most course topics were relevant to data analysis, programming, data processing, database design, and quality control and assurance, which is an important piece of curation. Still, quality control and assurance were not indicated in the course objectives, which may indicate that this is not the main goal of the course. Instead, they are taught as a subset of other main topics, such as data processing or complementary topics. Therefore, the terms *quality control*, *governance*, and *security* appeared sparsely in the course topic analysis.

Due to these two distinctive approaches, the skills covered by the two types of schools were also different. For instance, softer skills or other non-technical skills were covered throughout ALA courses (e.g., course topic terms such as *ethics*, *policy*, *copyright*, *licensing*, *publishing*, *sustainability*, and *lifecycle*; see Table 4), while only two terms relevant to soft skills, *governance* and *strategy*, appeared in CC course topics. There was also little or no mention of soft skill—relevant terms in the CC course objectives, except for *research* and *evaluation* (see Table 8). Perhaps, therefore, ALA courses have a variety of assignment types, from written and hands-on to projects, while CC courses heavily utilize hands-on exercises, which might be a better method for teaching hard skills. Technology and software utilization were also prevalent in CC courses, both reflected by the percentage of required tools mentioned in syllabi (89%) and the names of tools and software appearing in CC course objectives (see Table 8).

In addition, while ALA courses tried to address the full data lifecycle for future use, as reflected by the topical terms *sharing*, *publishing*, *reuse*, *preservation*, *archives*, and *repositories*, those terms did not appear in any CC course topics, as they seemed to focus more on current use. *Metadata* was the only term relevant to supporting the current and future use of data, which appeared under T3-CC: Data Quality Assurance and Quality Control. This could be a gap in CC schools' education approach, as researchers' understanding of the full data lifecycle often helps contribute to data sharing, which has become more common in many scientific disciplines.

This distinct approach is also well reflected, even when teaching technical skills, such as database design. Both the ALA and CC courses listed Database Design (T4-ALA, T1-CC) as among the major topics. However, CC courses focus more specifically on building and querying databases, the systems needed to manage the databases (Cassandra, Apache), and the mathematical theory that underpins database querying and joins (i.e., relational algebra). ALA courses teach higher-level concepts, such as standards and ontologies, while addressing SQL query design using MS Access.

Besides those comparisons, we found other interesting observations specific to each school category, as well as questions. While CC course topics and objectives were generally similar and well aligned, it was interesting that ALA course objectives varied from course topics but were still related. For example, the Technology-Focused course objective (O5-ALA) is relevant to course topics such as Metadata Management (T2-ALA) and Database Design (T4-ALA). The User- and Community-Focused course objective (O3-ALA) is relevant to course topics such as Data Services and Sharing (T3-ALA) and Data Archives and Preservation (T5-ALA). However, we would like to see how those objectives were implemented and taught in the actual course content. Moreover, one ALA course objective, Legal, Ethical, and Context Focused (O4-ALA), was not well reflected in the course topics, as it was one of the main course objectives in ALA courses (See Table 7). However, only a few terms reflect themes related to legal or ethical topics, two of which were included in T5-ALA: Data Archives and Preservation. We wondered what caused this mismatch. This may indicate that, although there was the intent to teach the legal and ethical issues of data curation, there might be only a small portion of class and teaching materials dedicated to that topic.

Our analysis also revealed gaps in both ALA and CC courses when teaching data curation and management. Courses from ALA schools generally address both soft and hard skills, but they focus less on some technical skills particularly relevant to data analysis, data conversion, data cleaning, statistics, data visualization, natural language processes, and programming. Those skills were suggested as the necessary hard skills from the previous literature (see Table 1), but they did not appear in our course topic or objective analysis. Even though it may not be necessary or desirable for LIS students to be experts in those skills, having at least basic knowledge in those areas will be helpful in order to have a broader view of the data lifecycle and to understand the nature of data and the use/reuse context, which is necessary to curate data for current and future use. Conversely, we observed that courses from CC schools generally lack sociotechnical perspectives, such as social, behavioral, and legal understanding. Even if the goal of education for CC courses is to produce data scientists who manage and analyze data, it is still critical to teach the legal and ethical compliance of data practices and the full data lifecycle, to contemplate the issue of data preservation, data sharing and reuse, and science reproducibility. Finally, we learned that courses from both schools did not have content relevant to other soft skills, such as project or workflow management and interpersonal and professional communication. As many data projects become more collaborative (Wilson, 2019), being able to work in a team environment in a collegial manner using professional communication is highly important.

Conclusion

Our study investigated data curation and management courses in ALA and CC schools by using the content analysis method on course syllabi. Our analysis demonstrated a growth in data curation education efforts since 2012, but it also presents areas for improvement, such as the need to reinforce more technical components in LIS data curation education. ALA courses have improved in providing technical skills, recalling the previous studies which argued that technical skills were not being taught (Goodsett & Koziura, 2016; Tammaro & Casarosa, 2014; Thomas & Urban, 2018). However, this can be ameliorated to reflect the skills expected for data curators. While librarians do not need to be programmers or data analysts, they should be able to talk to and communicate with domain experts, programmers, and technologists and be able to track and assess emerging technology. As ALA courses have educated students to understand how databases work, expanding the education to other technical and programming skills would complement existing educational efforts. In addition, while ethics and policy were clearly one of the main course objectives, further investigation is necessary to see how those objectives were met through actual course content to confirm that ALA courses sufficiently addressed those important topics.

While data curation and management are universal and essential skills in any data-centric field and are not monopolized areas of expertise specific to LIS, our study also revealed the unique aspect of LIS data curation education by offering a user-centered and service-oriented approach. This is well aligned with the provisioned role of librarians in the field of data curation, engaging with scientists during the research production cycle, supporting data handling and management, facilitating data deposits, and training and supporting data literacy by collaborating with various departments both within and outside campus (e.g., research office, domain-specific repositories, federal agencies) (Cragin, 2009). Because librarians play a role as a human interface between data and science, it is natural and important that LIS education brings several soft skills into data curation education. Knowledge about scholarly communication and the research process is thus expected for data curators and was integrated into ALA courses (see Table 7, O2-ALA: Data Services and O3-ALA: User and Community Focused). ALA education is also very strong in metadata and preservation, which is another unique aspect whereby both are core to supporting the long-term use and reuse of data, something that does not appear in CC courses. Typically, researchers or data scientists do not have knowledge of or access to tools for metadata generation, acquisition, data migration, or format validation (Heidorn, 2008) or have the skills necessary to prepare their data for sharing and reuse, which may lead to inefficient and costly practices (Cragin et al., 2010; Heidorn, 2008).

Our study also has a few limitations. First, because our unit of analysis is a syllabus, not a program, specialization, or degree, it is possible that our analysis does not reflect the whole picture of data curation education. For instance, although our analysis identified some gaps in education, such as project management and communication skills, it is possible that those skills are covered in other courses within the program, such as a library management course, which is typically a foundation course in many library science curricula. Also, since our analysis is based solely on the information provided on the syllabus, all of the topics, tools, concepts taught in the class may not be fully represented in our analysis. In addition, our study examined only a graduate-level curriculum, whereas postgraduate or professional development programs are also important to review to gain a holistic understanding of current educational efforts. For instance, there are a few free educational opportunities for professionals on RDM (e.g., Research Data Management Library Academy [RDMLA], Software Carpentry workshops, TrainRDM), as well as regional and ad-hoc workshops.

While topic modelling is an excellent tool for analyzing and finding patterns in unstructured text, there are several limitations to consider. The topics themselves can be difficult to interpret and require human judgement to determine the meaning of each topic. Additionally, the ALA corpora were quite small compared to the CC corpora, so the CC topic models were likely more robust and diverse in the themes they provided.

This research provides an understanding of the current landscape of data curation education, both at ALA-accredited MLIS programs and related CC institutions. Additionally, this research provides a comparison of educational trends for these programs and highlights areas of opportunities for strengthening data curation education. We are currently conducting ongoing research to deepen our understanding of data curation education, determine the data curation needs of organizations, and learn about the data curation competencies that professionals are using in the field. This ongoing research aims to create a Data Curation Competencies Framework through a thorough review of data curation literature and focus groups with data curation educators, data curation practitioners, and local and community organizations.

Ayoung Yoon is an associate professor at Luddy School of Informatics, Computing, and Engineering at Indiana University Indianapolis (IUI). She holds a PhD from the University of North Carolina at Chapel Hill and MSI from the University of Michigan. She specializes in research domains such as data curation, data sharing/reuse, open data, and community data. Her primary focus lies in fostering community capacity to create robust data sharing and reuse infrastructure and practice. Her research has been funded by the Institute of Museum and Library Services, Alfred P. Sloan Foundation, and Indiana University. Email: ayyoon@iu.edu

Angela M. Murillo is an assistant professor at the Luddy School of Informatics, Computing, and Engineering at IUI and program director for the Applied Data and Information Science Program. Her research focuses on data-related education, scientific data management, and cyberinfrastructure. Her research has been funded by the National Science Foundation (#2127548, #1842042), the Institute of Museum and Library Services (RE-252380-OLS-22), the IUPUI STEM Education Innovation and Research Institute, and the IUPUI Center for Teaching and Learning. Email: apmurill@iu.edu

Thomas (Wiley) Jettpace holds a BA from Indiana University – Bloomington in anthropology and psychology and an MS from IUI in human-computer interaction (HCI). He works as a product marketing manager in Chicago. His research interests are in data curation, social informatics, human data interaction, ubiquitous computing, and human-centered data science. Email: wileyjettpace@gmail.com

References

American Council of Education. (n.d.). *Carnegie classification of institutions of higher education*. https://carnegieclassifications.acenet.edu/classification_descriptions/basic.php

Baum, C. F. (2016). *An introduction to Stata programming* (2nd ed.). Stata Press.

Bishop, B., Gunderman, H., Davis, R., Lee, T., Howard, R., Samors, R., Murphy, F., & Ungvari, J. (2020). Data curation profiling to assess data management training needs and practices to inform a toolkit. *Data Science Journal*, *19*(1), 4. https://doi.org/10.5334/dsj-2020-004

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. The MIT Press.

Botticelli, P., Fulton, B., Pearce-Moses, R., Szuter, C., & Watters, P. (2011). Educating digital curators: Challenges and opportunities. *International Journal of Digital Curation*, *6*(2), 146–164. https://doi.org/10.2218/ijdc.v6i2.193

Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: A study of students and research faculty. *portal: Libraries and the Academy, 11*(2), 629–657. https://doi.org/10.1353/pla.2011.0022

Chen, H.-L., & Zhang, Y. (2017). Educating data management professionals: A content analysis of job descriptions. *Journal of Academic Librarianship*, *43*(1), 18–24. https://doi.org/10.1016/j.acalib.2016.11.002

Committee on Future Career Opportunities and Educational Requirements for Digital Curation. (2015). *Preparing the workforce for digital curation*. National Academies Press.

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: A guide to good practice* (2nd ed.). SAGE.

Cragin, M. H. (2009, July 13). *Data curation in LIS education and libraries* [Presentation]. ACRL-STS panel, Big Science, Little Science, E-Science: The Science Librarian's Role in the Conversation, ALA Annual Meeting, Chicago, IL. https://www.ideals.illinois.edu/handle/2142/13145

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. https://doi.org/10.1098/rsta.2010.0165

Delwiche, L. D., & Slaughter, S. J. (2003). *The little SAS book: A primer*. SAS Institute.

Fulton, B., Botticelli, P., & Bradley, J. (2011). DigIn: A hands-on approach to a digital curation curriculum for professional development. *Journal of Education for Library and Information Science*, *52*(2), 95–109.

Gold, A. (2010). *Data curation and libraries: Short-term developments, long-term prospects*. California Polytechnic State University.

Goodsett, M., & Koziura, A. (2016). Are library science programs preparing new librarians? Creating a sustainable and vibrant librarian community. *Journal of Library Administration*, *56*(6), 697–721. https://doi.org/10.1080/01930826.2015.1134246

Harris-Pierce, R. L., & Liu, Y. Q. (2012). Is data curation education at library and information science schools in North America adequate? *New Library World*, *113*(11/12), 598–613. https://doi.org/10.1108/03074801211282957

Heery E., & Noon M. (2017). *A dictionary of human resource management*. Oxford University Press.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280—299. https://doi.org/10.1353/lib.0.0036

Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. *Journal of Library Administration*, *51*(7–8), 662–672. https://doi.org/10.1080/01930826.2011.601269

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2018). How important is data curation? Gaps and opportunities for academic libraries. *Journal of Librarianship and Scholarly Communication*, *6*(1), eP2198. https://doi.org/10.7710/2162-3309.2198

Kabacoff, R. I. (2015). *R in Action: Data analysis and graphics with R* (2nd ed.). Manning Publications.

Kennan, M. A. (2016). Data management: Knowledge and skills required in research, scientific and technical organisations. *Proceedings of the World Library and Information Congress: IFLA General Conference and Assembly*. http://library.ifla.org/id/eprint/1466

Keralis, S. D. C. (2012). Data curation education: A snapshot. In L. Jahnke, A. Asher, & S. D. C. Keralis (Eds.), *The problem of data*. Council on Library and Information Resources. https://www.clir.org/pubs/reports/pub154/education/

Kim, Y., Addom, B. K., & Stanton, J. M. (2011). Education for eScience professionals: Integrating data curation and cyberinfrastructure. *International Journal of Digital Curation*, *6*(1), 125–138. https://doi.org/10.2218/ijdc.v6i1.177

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research*, *18*(2), 243–250. https://doi.org/10.1086/209256

Krier, L., & Strasser, C. A. (2014). *Data management for libraries: A LITA guide*. ALA Tech Source.

Lankes, R. D., Cogburn, D, Oakleaf, M., & Stanton. J. (2008). Cyberinfrastructure facilitators: New approaches to information professionals for E-Research. *Oxford e-Research Conference*. https://ora.ouls.ox.ac.uk/objects/uuid%3A64aa6f39-7e81-4d42-a008-ee2d7524bd67

Lee, D. J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE*, *12*(3), e0173987. https://doi.org/10.1371/journal.pone.0173987

Madrid, M. M. (2011). *A study of digital curator competences: A survey of experts* [Master's thesis, University of Parma]. https://hdl.handle.net/1889/1785

National Library of Medicine (NLM). (n.d.). *Data curation*. https://www.nnlm.gov/guides/data-glossary/data-curation

Ortiz-Repiso, V., Greenberg, J., & Calzada-Prado, J. (2018). A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the iSchools. *Journal of Information Science*, *44*(6), 768–784. https://doi.org/10.1177/0165551517748149

Palmer, C. L., Allard, S., & Marlino, M. (2011). Data curation education in research centers. *Proceedings of the 2011 iConference*, 738–740. https://doi.org/10.1145/1940761.1940891

Palmer, C. L., Thompson, C. A., Baker, K. S., & Senseney, M. (2014, March 1). Meeting data workforce needs: Indicators based on recent data curation placements. *IConference 2014 Proceedings*. https://doi.org/10.9776/14133

Palmer, C. L., Weber, N. M., Muñoz, T., & Renear, A. H. (2013). Foundations of data curation: The pedagogy and practice of "purposeful work" with research data. *Archive Journal*, 3. http://www.archivejournal.net/essays/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/

ProWebScraper. (2019, April 18). *What is data curation, and why is it important?* https://prowebscraper.com/blog/what-is-data-curation-and-why-is-it-important/

Ray, J. (2009). Sharks, digital curation, and the education of information professionals. *Museum Management and Curatorship*, *24*(4), 357–368. https://doi.org/10.1080/09647770903314720

Ray, J. M. (2014). *Research data management: Practical strategies for information professionals*. Purdue University Press.

Reisner, B. A., Vaughan, K. T. L., & Shorish, Y. L. (2014). Making data management accessible in the undergraduate chemistry curriculum. *Journal of Chemical Education*, *91*(11), 1943–1946. https://doi.org/10.1021/ed500099h

Song, M., & Ding, Y. (2014). Topic modeling: Measuring scholarly impact using a topical lens. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact* (pp. 235–257). Springer International Publishing. https://doi.org/10.1007/978-3-319-10377-8_11

Stanford Natural Language Processing Group. (n.d.). *Stanford Topic Modeling Toolbox*. https://nlp.stanford.edu/software/tmt/tmt-0.4/

Sylvia, M., & Terhaar, M. (2014). An approach to clinical data management for the doctor of nursing practice curriculum. *Journal of Professional Nursing*, *30*(1), 56–62. https://doi.org/10.1016/j.profnurs.2013.04.002

Tammaro, A. M., & Casarosa, V. (2014). Research data management in the curriculum: An interdisciplinary approach. *Procedia Computer Science*, *38*, 138–142. https://doi.org/10.1016/j.procs.2014.10.023

Thomas, C. V. L., & Urban, R. J. (2018). What do data librarians think of the MLIS? Professionals' perceptions of knowledge transfer, trends, and challenges. *College and Research Libraries*, *79*(3), 401–423. https://doi.org/10.5860/crl.79.3.401

Thompson, C. A., Senseney, M., Baker, K. S., Varvel, V. E., & Palmer, C. L. (2013). Specialization in data curation: Preliminary results from an alumni survey, 2008–2012. *Proceedings of the American Society for Information Science and Technology, 50*(1), 1–4. https://doi.org/10.1002/meet.14505001151

Tibbo, H. R., & Duff, W. (2008). Toward a digital curation curriculum for museum studies: A North American perspective. *Proceedings of 2008 annual conference of the International Documentation Committee of the International Council of Museums*. https://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2018/12/70_papers.pdf

Uzialko, A. (2022, February 16). *What is data management?* Business.com. https://www.business.com/articles/what-is-data-management/

Varvel, V. E., Bammerlin, E. J., & Palmer, C. L. (2012). Education for data professionals: A study of current courses and programs. *Proceedings of the 2012 iConference*, 527–529. https://doi.org/10.1145/2132176.2132275

Virkus, S., & Garoufallou, E. (2019). Data science from a library and information science perspective. *Data Technologies and Applications*, *53*(4), 422–441. https://doi.org/10.1108/DTA-05-2019-0076

Virkus, S., & Garoufallou, E. (2020). Data science and its relationship to library and information science: A content analysis. *Data Technologies and Applications*, *54*(5), 643–663. https://doi.org/10.1108/DTA-07-2020-0167

Walters, T. O. (2009). Data curation program development in U.S. universities: The Georgia Institute of Technology example. *International Journal of Digital Curation*, *4*(3), 83–92. https://doi.org/10.2218/ijdc.v4i3.116

Wang, L. (2018). Twinning data science with information science in schools of library and information science. *Journal of Documentation*, *74*(6), 1243–1257. https://doi.org/10.1108/JD-02-2018-0036

Wickham, H., & Grolemund, G. (2017). R for data science. O'Reilly.

Wilson, R. (2019). Collaborative data projects: What we learned by contributing to them. https://theodi.org/article/collaborative-data-projects-what-we-learned-by-contributing-to-them/

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, *4*(3), 93–103. https://doi.org/10.2218/ijdc.v4i3.117

Yakel, E. (2007). Digital curation. *OCLC Systems & Services: International Digital Library Perspectives*, *23*(4), 335–340. https://doi.org/10.1108/10650750710831466

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 233–242). https://doi.org/10.1145/2623330.2623715

Yoon, A., Murillo, A. P., & McNally, P. A. (2021). Digital preservation in LIS education: A content analysis of course syllabi. *Journal of Education for Library and Information Science*, *61*(1), 61–86. https://doi.org/10.3138/jelis.62.1-2018-0053

Yoon, A., & Schultz, T. (2017). Research data management services in academic libraries in the US: A content analysis of libraries' websites. *College & Research Libraries*, *78*(7), 920–933. https://doi.org/10.5860/crl.78.7.920

Appendix A: Topic model distributions

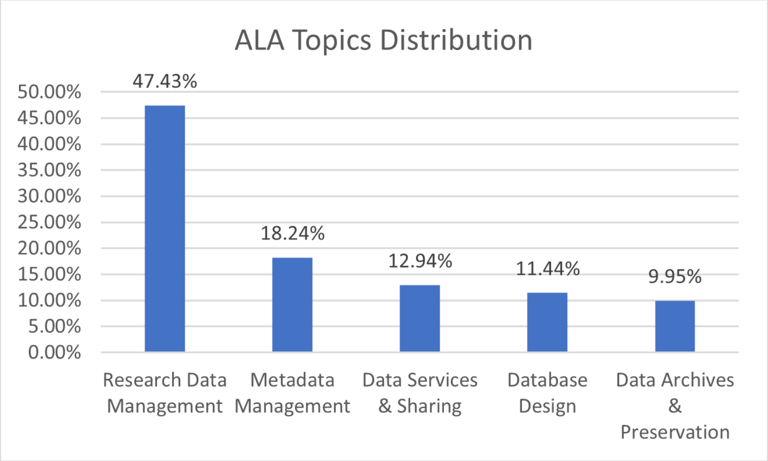

Figure A1: ALA topic model distribution (course topics)

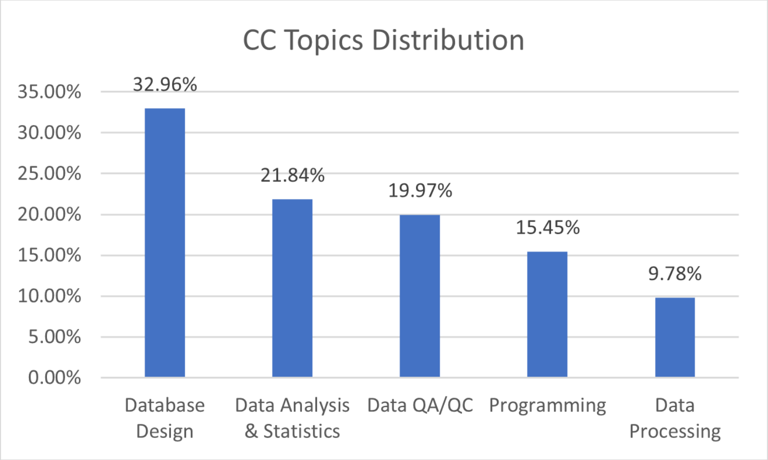

Figure A2: CC topic model distribution (course topics)

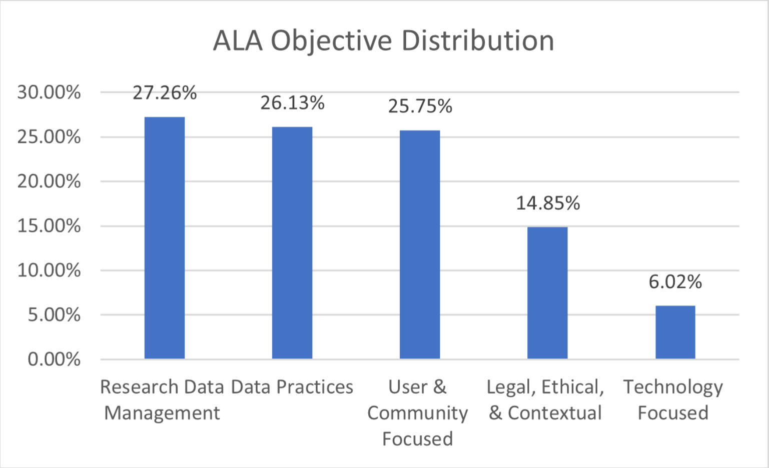

Figure A3: ALA topic model distribution (course objectives)

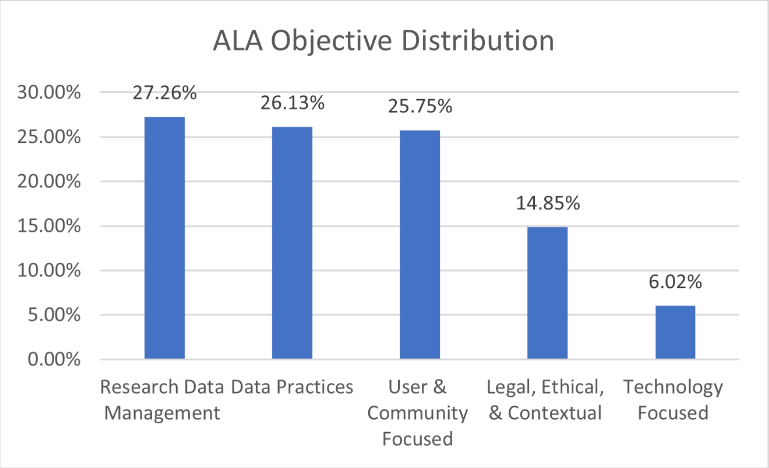

Figure A4: CC topic model distribution (course objectives)

Appendix B: ALA and CC term distribution comparison

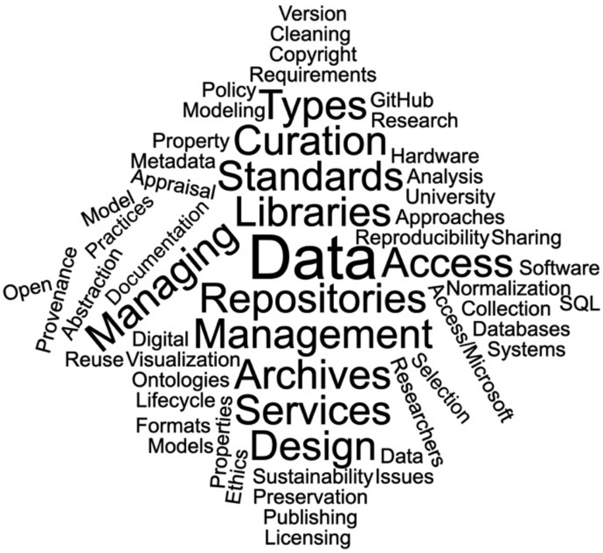

Figure B1: ALA term distribution (course topics)

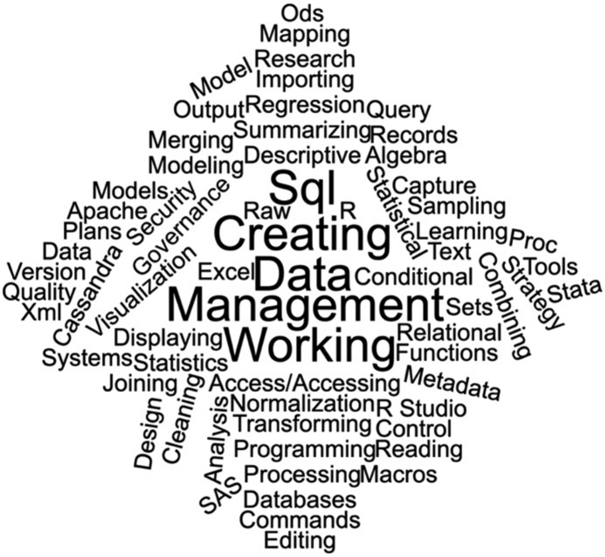

Figure B2: CC term distribution (course topics)

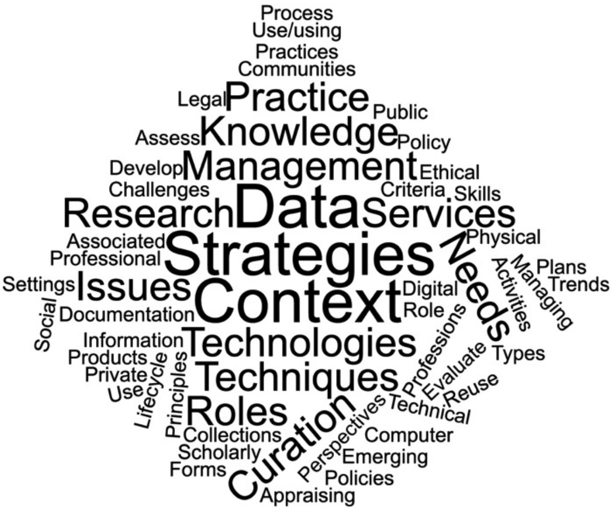

Figure B3: ALA term distribution (course objectives)

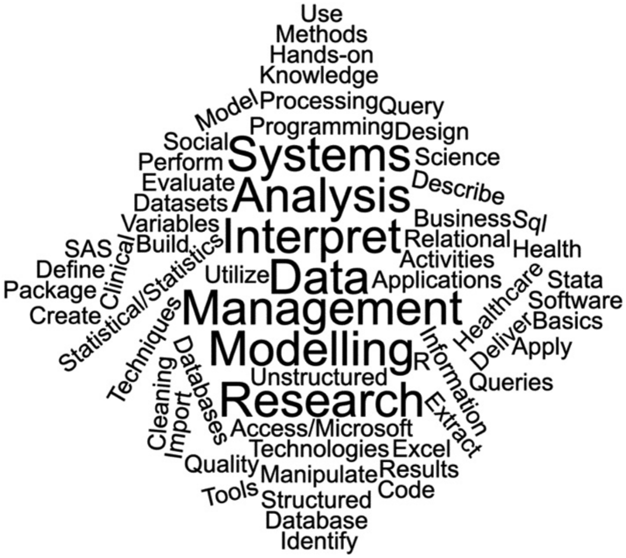

Figure B4: CC term distribution (course objectives)

Appendix C: ALA top 15 terms (course topics)

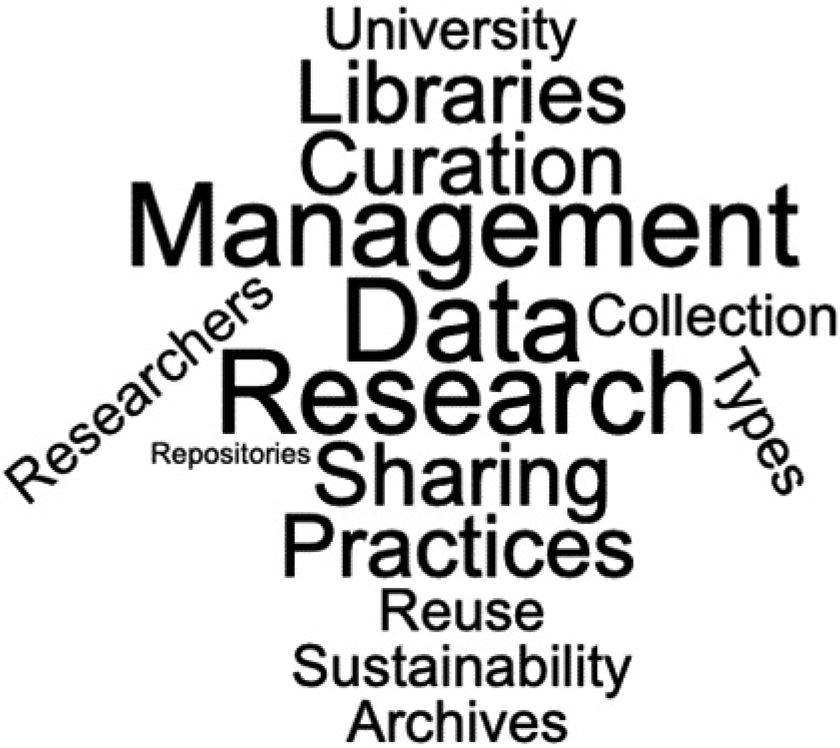

Figure C1: T1-ALA: Research Data Management (47.43%)

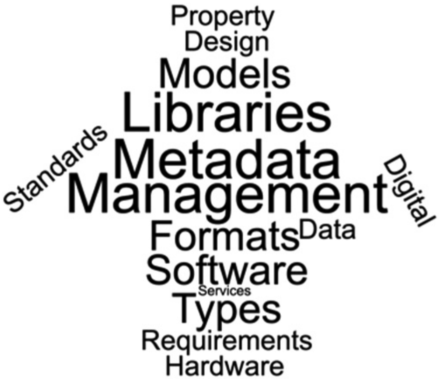

Figure C2: T2-ALA: Metadata Management (18.24%)

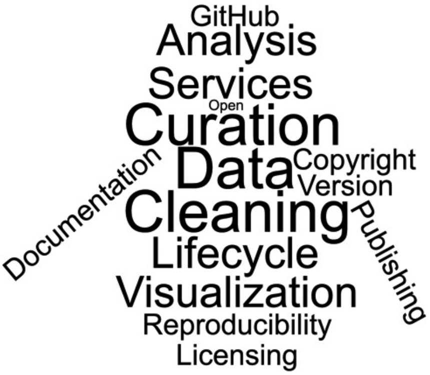

Figure C3: T3-ALA: Data Services and Sharing (12.94%)

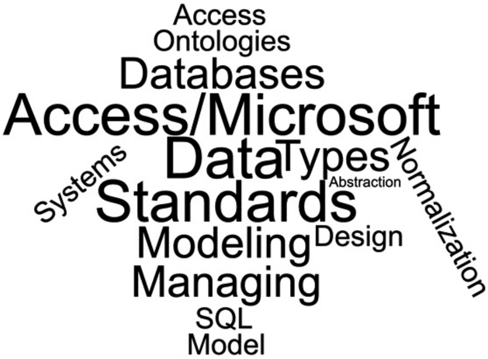

Figure C4: T4-ALA: Database Design (11.44%)

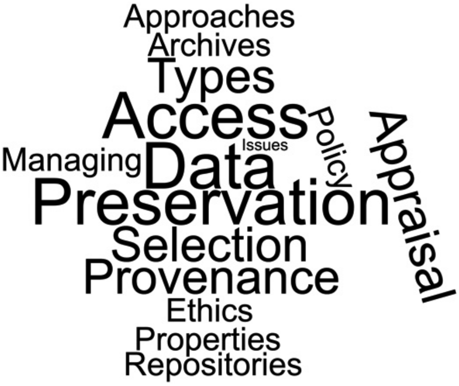

Figure C5: T5-ALA: Data Archives and Preservation (9.95%)

Appendix D: CC top 15 terms (course topics)

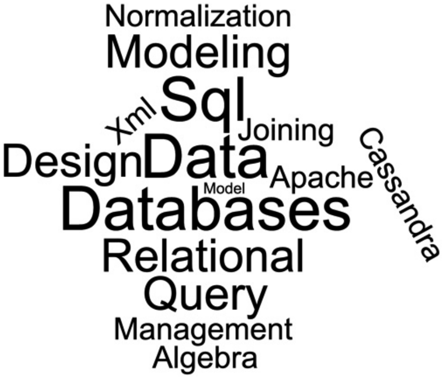

Figure D1: T1-CC Database Design (32.96%)

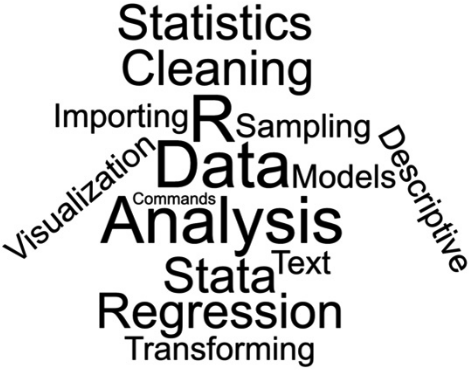

Figure D2: T2-CC Data Analysis and Statistics (21.84%)

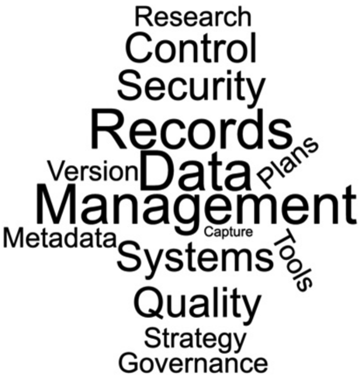

Figure D3: T3-CC: Data Quality Assurance and Quality Control (19.97%)

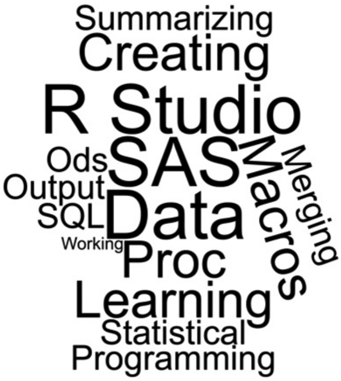

Figure D4: T4-CC Programming (15.45%)

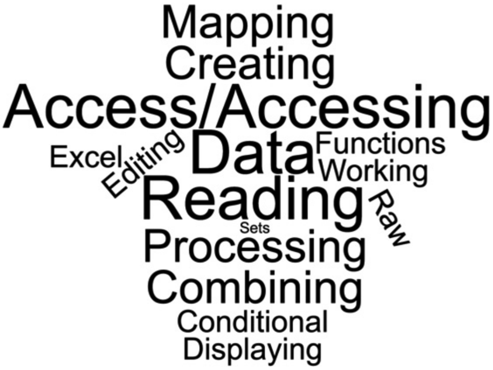

Figure D5: T5-CC Data Processing (9.78%)

Appendix E: ALA top 15 terms (course objectives)

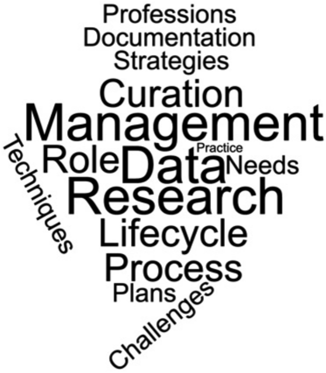

Figure E1: T1-ALA Research Data Management (27.26%)

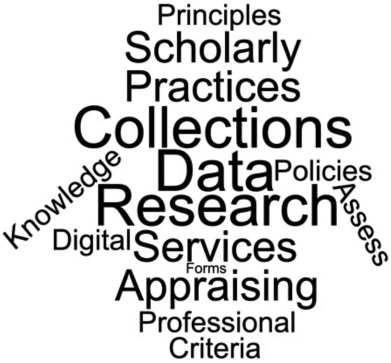

Figure E2: T2-ALA Data Services (26.13%)

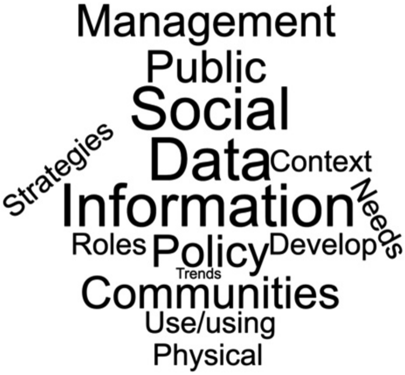

Figure E3: T3-ALA User and Community Focused (25.75%)

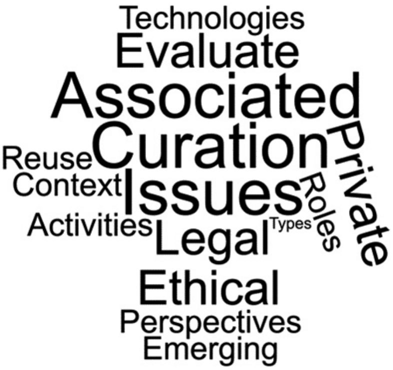

Figure E4: T4-ALA Legal, Ethical, and Context Focused (14.85%)

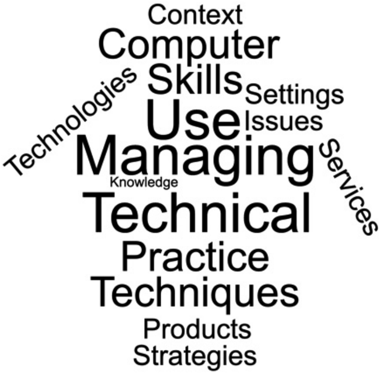

Figure E5: T5-ALA Technology Focus (6.02%)

Appendix F: CC top 15 terms (course objectives)

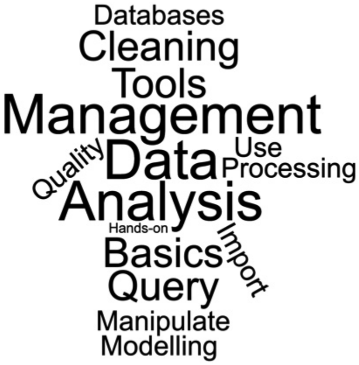

Figure F1: T1-CC Database Management Foundations (25.89%)

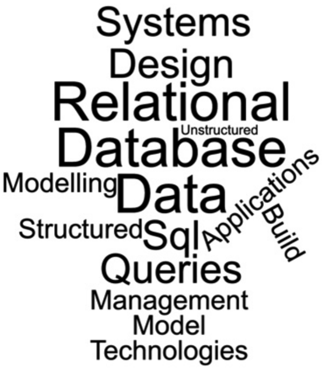

Figure F2: T2-CC Database Design (22.81%)

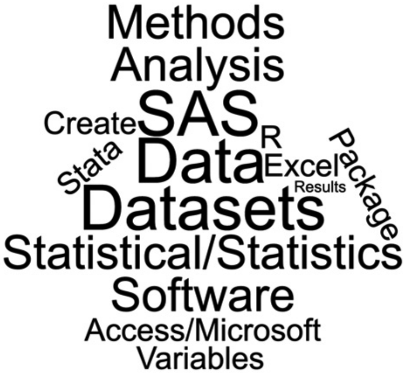

Figure F3: T3-CC Data Analysis and Statistics (20.84%)

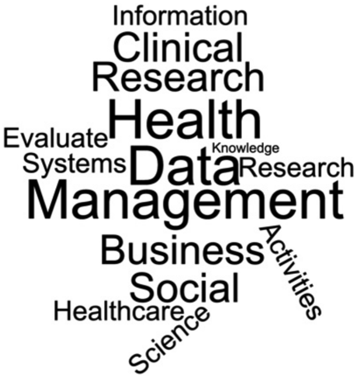

Figure F4: T4-CC Domain-Specific (17.09%)

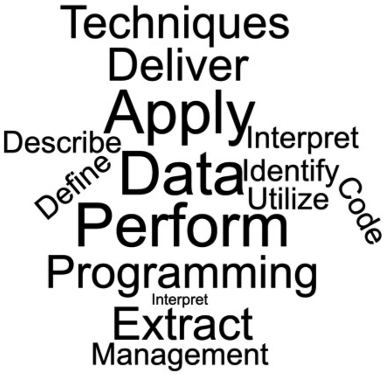

Figure F5: T5-CC Data Processing (13.74%)