IJDC | *Peer-Reviewed Paper*

Identifying Topical Coverages of Curricula using Topic Modeling and Visualization Techniques: A Case of Digital and Data Curation

Seungwon Yang
School of Library and Information Science
Center for Computation and Technology
Louisiana State University

Boryung Ju
School of Library and Information Science
Louisiana State University

Haeyong Chung
Department of Computer Science
The University of Alabama in Huntsville

Abstract

Digital/data curation curricula have been around for a couple of decades. Currently, several ALA-accredited LIS programs offer digital/data curation courses and certificate programs to address the high demand for professionals with the knowledge and skills to handle digital content and research data in an ever-changing information environment. In this study, we aimed to examine the topical scopes of digital/data curation curricula in the context of the LIS field, using a semi-automatic approach. We collected 16 syllabi from the digital/data curation courses, as well as textual descriptions of the 11 programs and their core courses offered in the U.S., Canada, and the U.K. The collected data were analyzed using a probabilistic topic modeling technique, Latent Dirichlet Allocation, to identify both common and unique topics. The results are the identification of 20 topics both at the program- and course-levels. Comparison between the program- and course-level topics uncovered a set of unique topics, and a number of common topics. Furthermore, we provide interactive visualizations for digital/data curation programs and courses for further analysis of topical distributions. We believe that our combined approach of a topic modeling and visualizations may provide insight for identifying emerging trends and co-occurrences of topics among digital/data curation curricula in the LIS field.

Received 09 August 2018 ~ *Revision received* 21 November ~ *Accepted* 27 January 2019

Correspondence should be addressed to to Seungwon Yang, School of Library & Information Science and Center for Computation and Technology, Louisiana State University, 267 Coates Hall, Baton Rouge, LA 70803. E-mail: seungwonyang@lsu.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Curricula related to digital and/or data curation, is not new to the field of library and information science. Yet, the advancement of information technologies and cyber-scholarships, as well as the mandatory requirements of data management plans by research funding agencies, has led to an exponential increase of digital content and research data. This, in turn, leads to an ever-growing need for the teaching and learning of digital/data curation (Higgins, 2011; Tibbo, 2015; Yakel, Conway, Hedstrom and Wallace, 2011). Several education programs for digital/data curation, exist to address such needs, including specializations in Masters' programs, certificates, conferences, and workshops. In addition, Massive Open Online Courses (MOOCs) relevant to digital/data curation are also offered, such as:

- Research Data Management and Sharing, taught by Helen Tibbo at the University of North Carolina, Chapel Hill and Sarah Jones at Digital Curation Center, U.K. (offered by Coursera.com)
- Data Management for Clinical Research, taught by Stephany Duda and Paul Harris at Vanderbilt University, U.S. (offered by Coursera.com)
- Introduction to Digital Curation, taught by Jenny Bunn at University College London, U.K. (offered by class-central.com)

There also exists a large number of online resources for digital/data curation and other similar courses, such as research data management or preservation. For example, a YouTube search using the key phrase "digital curation" returns approximately 31,400 results at the time of this writing. The key phrase "data curation" returns 7,420 results, "research data management" returns 1.13 million results, and "digital preservation" returns approximately 227,000 video resources.

While such a proliferation of digital/data curation curricula has provided benefits to the community in curating and managing traditional digital objects and digital research data, the broad scopes and overlapping nature, of digital and data curation curricula has also increased the potential complexity and confusion in selecting topics that can cover each curriculum comprehensively. The wide coverage of topics in both digital and data curation curricula may hinder instructors from having a "big picture" of the topical distribution in the field. This, in turn, can lead to selecting an unbalanced set of topics, or missing important topics altogether when they are preparing their digital or data curation course materials. Furthermore, students who are taking those courses may have unbalanced knowledge in digital or data curation.

In spite of the potentially negative impacts from an unbalanced or lacking set of topics in the curriculum, a few studies exist that examine and analyze topics across multiple digital/data curation curricula.

The purpose of our study is to examine the broad topical scopes of digital/data curation curricula by addressing the research questions (RQs) below:

- RQ.1. Which topics can we generate from digital/data curation curricula?
- RQ.2. What are the similarities and differences between the two topic groups?

 doi:10.2218/ijdc.v14i1.586

- RQ.3. What is an effective way to interactively visualize the topics for comparison?

In order to address RQ.1., we applied a topic modeling approach, Latent Dirichelet Allocation (LDA), to 11 certificate program descriptions (two data curation and nine digital curation programs in the U.S., Canada, and U.K.), as well as 16 individual course descriptions (four data curation and 12 digital curation). With the understanding that certificate programs include multiple core courses, which may contain the main theme of the program, we merged descriptions of core courses with the program description prior to applying LDA. To address RQ.2., the topics generated for the digital/data curation programs were tabulated and compared for between-program topical similarities and differences. The same approach applied to the course descriptions as well. This resulted in another topic table for between-course topical comparisons. We then developed visualizations of topical distribution for both certificate programs and courses in order to address RQ.3. Our visualizations provided interactions such as the sorting of topics based on their frequency across programs or courses, or the re-grouping of topics within each program or course.

We hope that, by examining the topical coverage of curation curricula, the findings of this study will contribute to the digital/data curation community in several ways. First, there exist only a few studies which attempt to investigate the topical scopes of digital/data curation curricula systematically. In an effort to fill this gap, we collected the data from current program and course descriptions, as well as course syllabi, and then analyzed the data by employing a probabilistic topic modeling method. Second, by offering a comprehensive topical "map" of both digital and data curation curricula, instructors who are preparing courses may develop more comprehensive courses by including the major topics provided by most digital/data curation curricula. Third, instructors who have been teaching such courses may upgrade their courses by including more topics into their existing classes. As future studies analyze more digital/data curation curricula, a comprehensive picture may be created to allow for more effective teaching and more consistent learning by the future digital/data curators.

Curricula for Digital and Data Curation

Literature Review

Both academics and practitioners have been working to understand and define the scope of digital/data curation for a few years now. Considering that it is such a complex concept with many facets, it is no surprise that the results are varied. With the advent of massive storage capacities, such as cloud computing and big data processing, data-intensive research fields require effective data management and the ability to reuse large amounts of scientific data. Along with this phenomenon, federal funding agencies, such as the National Institute of Health (NIH) and the National Science Foundation (NSF), have mandated data management plans and data deposits. In addition, the Federal Research Public Access Act, which was reintroduced by a number of senators in 2009, requires open public access to journal articles produced by research that are funded by 11 U.S. federal government agencies (Ogburn, 2010; Hiedorn, 2011). The Library and Information Science (LIS) field has seen a rising demand for professionals who are equipped with the knowledge and skills to cope with this new trend of data production

and management. In 2008, Swan and Brown pointed out that only a few LIS schools teach required digital curation skills to future librarians and "the need for education and training of digital curation is more pressing than ever, due to massive increase in digital data" (Tibbo, 2015).

Data curation is often used interchangeably with data management, digital curation, or data science in the LIS field, with many attempts to define these concepts. The term "curated" comes from the Latin curare – to care for (Buneman, 2008). Therefore, data curation refers to "the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation."[1]

According to Yakel (2007), digital curation is defined as "the umbrella term for digital preservation, data curation, and digital asset and electronic records management" (p. 338). The term 'digital preservation' refers to a 'series of managed activities necessary to ensure continued access to digital materials for as long as necessary' (Digital Preservation Coalition, 2009). Hirtle (2010) pointed out that "the concept of digital preservation originally developed in libraries, not archives, as an aid to ongoing library analog preservation efforts" and that "it initially did not concern itself with the preservation of information that was 'born digital'" (p. 124). In the same venue, 'digital curation [involves] maintaining and adding value to digital research data throughout its lifecycle.'[2] The active management of research data through curation "reduces threats to [the data's] long-term research value and mitigates the risk of digital obsolescence" while making the data more sharable. Therefore, digital curation is "the creation, management, and use of digital materials… for a wide range of activities…" The term digital curation is increasingly being used for the actions needed to add value to and maintain these digital assets over time, for current and future generations of users (Beagrie, 2008).

Harris-Pierce and Liu (2012) examined 52 LIS schools in the United States and Canada to analyze the number of data curation courses, at the graduate or graduate certificate level. They conducted a qualitative content survey to determine if the offered courses are adequate enough to address the needs of coping with "data deluge" (Hockx-Yu, 2006). They identified the gap between such needs and the LIS education in data curation at that time. In order to be considered as a data curation course, the term "data curation" had to be included in the course title, course description, or course objectives; or one or more of data curation's programs were the main or partial focus of the course. The findings reported which LIS schools offer data curation courses; course titles (digital vs. data curation); course descriptions and objectives; the prerequisites for the courses; course readings and assignments; and the topics covered in the courses. The authors identified various topics covered in the courses and suggested a need for continuing collaborative work to determine the optimal course objectives and learning outcomes.

Previous literature has reviewed several prominent degrees and/or certificate programs in the data and digital curation areas of the LIS field. Many of these programs identified the imminent need for incorporating data curation and management education in the LIS field (Yakel, 2007; Brown, 2009; Ogburn, 2010; Hiedorn, 2011). The University of Michigan, School of Information proposed a

1 Definition of data curation by The University of Illinois' Graduate School of Library and Information Science: https://www.clir.org/initiatives-partnerships/data-curation

2 Digital Curation Centre – What is digital curation?: http://www.dcc.ac.uk/digital-curation/what-digital-curation

 doi:10.2218/ijdc.v14i1.586

specialization, Preservation of Information (PI), to emphasize digital curation and preservation issues. The school's original Masters' level course in the administration of archives and records (which began in the 1960s) and archives and record management (ARM) specialization (which began in the mid-1990s) has evolved and extended to include this new specialization. The PI specialization has a strong digital curation curriculum, which consists of three components: courses, practice-based internships, and a solid technological infrastructure (Yakel, Conway, Hedstrom and Wallace, 2011). Courses include Preserving Information, Advanced Preservation Management, Digital Preservation, Web Archiving, Digitization for Preservation, Preserving Sound and Motion, and Economics of Sustainable Digital Preservation.

Harvey and Bastian (2012) introduced the digital curation curriculum at Simmons College, which focuses on cultural heritage informatics. In order to cope with a fast changing and intensifying virtual information and preservation environment and demand for state-of-the-art curriculum offering, Simmons developed the Digital Curriculum Laboratory (DCL) in 2008. This web-based space "provides integrated access and is available for experiment in the context of archives and preservation" (p. 27). The DCL is designed to teach digital curation courses, such as Archiving and Preserving Digital Media and Digital Stewardship, and facilitate students' learning with hands-on experience and related theories.

University of Arizona's DigIn (Digital Information Management) certificate program, initially began in 2005 (Fulton, Botticelli and Bradley, 2011) to address the shortage of professionals for the upcoming digital data and information environment. Their certificate program includes courses in technology, management, and policy issues and provides hands-on and case-based learning opportunities, through their digital lab environment. Their courses, such as Introduction to Digital Collection, Introduction to Applied Technology, Managing the Digital Environment, and Preservation of Digital Collections, highlight a focus on digital curation education.

At the University of North Texas, the digital curation curriculum was designed by using a competency-based approach (Kim, 2015). Using a combination of digital curation job analysis, competency standards, and curriculum benchmarking, they developed four competencies for the digital curation program: Content curation competency; Curation technologies competency; Curation models and modeling competency; and Curation services competency (e.g., Matrix of Digital Curation Knowledge and Competencies developed by University of North Carolina's DigCCur Project (Lee, 2009)). The resulting four courses were developed to reflect this competency model: Digital Curation Fundamentals, Digital Curation Tools and Applications, Preservation Planning and Implementation for Digital Curation, and Advanced Topics in Digital Curation. Lee presented a six-dimensional matrix for identifying and organizing the material to be covered in a digital curation curriculum from the DigCCurr project at the University of North Carolina, Chapel Hill (Lee, 2009; Lee, Tibbo and Schaefer, 2007).

Several other major LIS schools, such as the University of Illinois at Urbana-Champaign, the University of Texas at Austin, Syracuse University, University of California Los Angeles, and University of Tennessee, also offer graduate certificate programs or graduate degree programs in digital curation (Ray, 2012; Tibbo, 2015). Although, Tibbo pointed out, the distinction of exclusively "digital" curation could be dropped from the title in the future, due to the possible prevalence of the digital domain while the notion of "curation" could be sustained in digital curation (Tibbo, 2015). This obvious lack of a coherent topical domain regarding digital/data curation warrants a further systematic examination of the scope of the digital/data curation domain.

Methodology

Collecting Data

Over the last decade, LIS programs have foreseen the advent of the digital information environment and recognized increasing demands for information professionals equipped with the knowledge and skills for curating digital data. Hence, existing archival or record management programs have been expanded, or new digital curation degrees and certificate programs have been created along with corresponding curricula. Currently, approximately one-half of LIS programs offer the digital/data curation-related programs and/or courses.

Our data collection was focused on finding the descriptions of digital/data curation post-graduate certificates and programs and their required courses, as well as collecting digital/data course syllabi online. We examined a list of 59 American Library Association (ALA)-accredited LIS Programs. Our data sets were confined to the material written in English, to avoid the possible ambiguities of words found in the collected texts caused by different languages. A couple of the identified LIS programs from Spanish-speaking institutions were excluded from our data. Thus, our data included digital/data curation programs in the U.S., Canada, and the U.K.

The data, both at the program-level and at the course-level, were collected to examine the two different topical scopes, respectively. For the program-level data collection, we visited the websites of the above-mentioned LIS programs during Fall 2015 and Spring 2016. For the course-level data collection, we searched for courses in each LIS program, as well as its parent institution's graduate school catalogs, in order to find the course descriptions and the syllabi. If we could not find a course syllabus online, we contacted the instructors of the courses and asked for a copy of the syllabus. In total, sixteen syllabi of digital/data curation were gathered for data analysis.

Extracting Topics

In order to uncover topical scopes from the program and course descriptions in our data, we applied a topic modeling (Blei, Ng and Jordan, 2003; Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2004; Steyvers and Griffiths, 2007; McCallum, 2002). Topic models automatically process multiple documents, and then generate latent topics by considering the features of words in texts such as their frequency and relationships with other words. In this way, texts can be processed and summarized efficiently for effective human understanding.

Of the various topic model approaches, we used a probabilistic model, called Latent Dirichlet Allocation (LDA), to uncover common, as well as unique topics that were present in the program and course descriptions (Blei, Ng and Jordan, 2003; Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2007). LDA, a frequently-used probabilistic topic model, is based on the idea that a text document consists of a mixture of topics that are latent in the document. Different topic models make different assumptions regarding how those topics are mixed (i.e., distributed) in a document. In LDA, a Dirichlet distribution is introduced as the prior for topic distribution, and this further simplified the problem of the Bayesian statistical inference of topics (Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2007).

 doi:10.2218/ijdc.v14i1.586

LDA requires its users to set up several parameters, including the number of topics and the number of words in each topic. Each topic in LDA is then represented as a group of words. For this current research, we set the number of topics to be 20 and the number of words in each topic to be ten. These parameters were found empirically by balancing the generality and uniqueness of the topics. For example, selecting the number of topics to be ten would produce topics that are too general and broad across program or course descriptions. This would not identify the distinct topics in each program/course. If we set the number of topics to be 30, LDA would produce multiple unique topics which are part of only a few programs and course descriptions. The model would fail to uncover the common and broad topics that exist across programs or courses. Setting the number of topics at 20 generated both the broad and common topics across curricula, as well as the unique topics that are specific to each individual curriculum.

The LDA model also requires setting up two hyperparameters, α and β, which affect the topic and word combinations respectively. In general, setting $\alpha = 50/T$ and $\beta = 0.01$ is known to work well with various types of text collections (Steyvers and Griffiths, 2006). T denotes the number of topics, so we set $\alpha = 50/20 = 2.5$ and $\beta = 0.01$ for this research. Before applying LDA text files were created. For program-level topics, we merged each program description and that entire program's required courses' descriptions into a single text file. This generated a total of 11 text files. For the course-level topic extraction, each of the 16 course descriptions was added to their own text file.

As explained above, topics generated by LDA are considered latent in a document. In other words, they are not observable directly in the content of the document. Instead, LDA represents each topic as a group of multiple words, from which a topic can be inferred. From the study of Chang et al. (2009), we could expect that human intervention in interpreting LDA topics could be a means to improve the quality of topics generated from the LDA algorithm. Thus, we manually assigned a single "inferred" topic that may represent each word group – an LDA topic – for simplicity purposes. As an effort to reduce subjectivity in assigning those inferred topics, an external collaborator worked together with us to generate inferred topics independently, and then merged them together. Whenever an inferred topic was not agreed upon, there was a short discussion to resolve it.

Developing Interactive Visualizations

Once latent topics from curricular descriptions were modelled by LDA, we needed to understand how these topic models could be employed by the schools and programs. Talley et al. (2011) and Hall, Jurafsky, and Manning (2008) emphasized that it is important for human analysts to be able to identify and validate the latent topics discovered within a large collection of text documents more efficiently through the use of visualizations. Accordingly, we extended existing visualizations for a more effective presentation of topics (Chuang, Ramage, Manning and Heer, 2012; Chuang, Ramage, McFarland, Manning and Heer, 2012; Ramage, Hall, Nallapati and Manning, 2009).

The visualizations that we employed in this study are based on a tabular layout primarily inspired by both a simple bubble matrix and Termite (Chuang, Manning and Heer, 2012). The tabular layout was designed to facilitate comparing and seriating co-occurring courses/programs both within and across latent curriculum topics. This visualization consists of two primary components, the Matrix and Bar chart views.

For this visualization, each row found in the Matrix view corresponds to a curriculum topic model identified by LDA and each column corresponds to a different

digital/data curation program. Accordingly, each circle represents an occurrence relationship between curriculum topic (row) and a program or school (column). Additionally, the color of each circle indicates different types of curricula between the data curation (red) and digital curation programs (blue). Each bar on the Bar chart view on the right indicates the total number of programs corresponding to a latent curriculum topic. Thus, by examining this view, we can understand how many schools/programs belong to specific topics. Additionally, the entire visualization view can be reorganized by sorting the rows and columns for the Matrix view and sorting the Bar chart view in ascending or descending order.

Using our interactive visualizations of topical distributions allows us to observe the co-occurrence of the curriculum-derived topics in digital/data curation programs. In addition, we can also evaluate emerging patterns in datasets and digital curation programs based on the occurrences of topics. For instance, our visualizations enable users to quantify and assess the over- and under-utilized sub-topics for each overarching curriculum topic (see the Discussion section). In particular, our visualization approach enables us to verify whether one or more latent curriculum topics are being used by additional universities and programs, as well as to refine the topic models through the use of "human-in-the-loop" supervision (Chuang, Gupta, Manning and Heer, 2013).

Results

Anatomy of Collected Data

As described in the Methodology section, we collected data in two levels: (1) program-level and (2) course-level. Figure 1 illustrates programs and courses included in our dataset as boxes that are organized in two rows. The 11 boxes in the first row denote the certificate programs included in our data. Each program has a prefix "PGM_." There are three small, dark gray boxes attached to three programs (one data curation and two digital curations) in the first row, which indicates that those programs have associated course syllabi that are part of our course data. The 16 boxes in the second row show individual courses in digital/data curation curricula. Each course has a prefix "CRS_." Four syllabi were identified from three programs as indicated by the arrows.[3]

3 The data used in this study can be downloaded from: https://github.com/seungwonyang/IJDC

 doi:10.2218/ijdc.v14i1.586

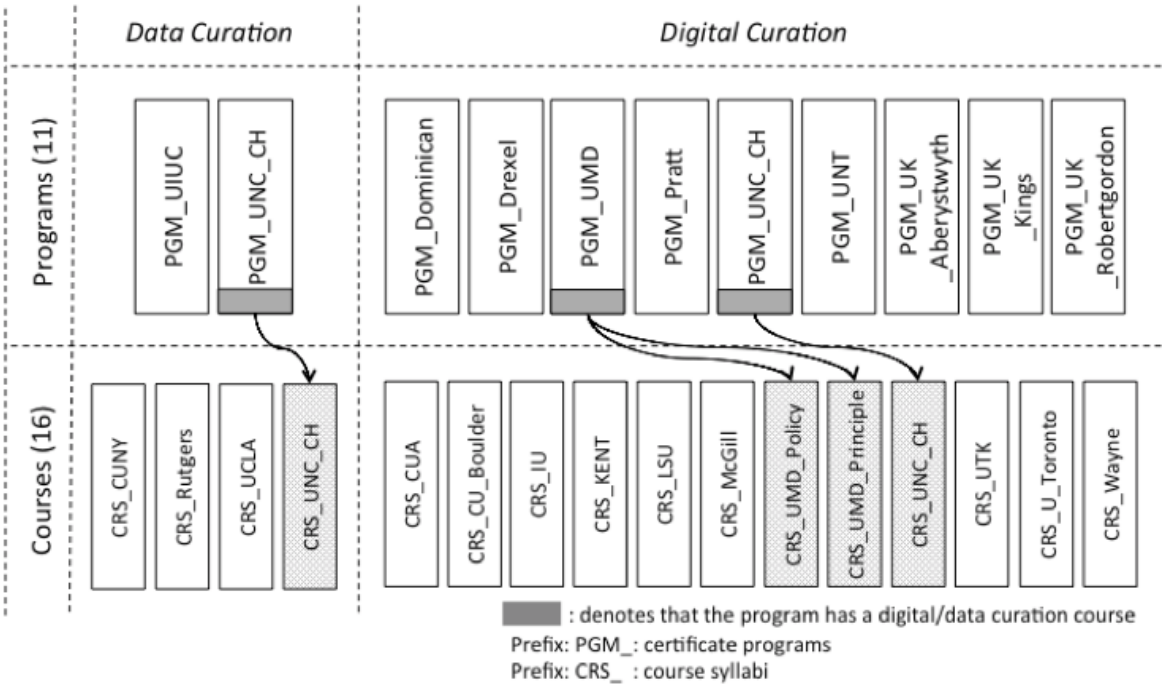

Figure 1. Post-graduate certificate programs and courses for digital/data curation. The courses with gray rectangles have their associated programs.

Program-level descriptions. In total, 11 certificate programs in digital/data curation fell under this category. Three digital curation programs were offered in the U.K., and 8 were offered in the U.S.:

- Data curation programs: University of Illinois Urbana-Champaign (PGM_UIUC), and University of North Carolina, Chapel Hill (PGM_UNC_CH)
- Digital curation programs: Dominican University (PGM_Dominican), Drexel University (PGM_Drexel), University of Maryland (PGM_UMD), Pratt Institute (PGM_Pratt), University of North Carolina, Chapel Hill (PGM_UNC_CH), and University of North Texas (PGM_UNT)
- UK Digital curation programs: University of Aberystwyth (PGM_UK_Aberystwyth), Kings College, London (PGM_UK_Kings), and Robert Gordon University, Aberdeen (PGM_UK_Robertgordon)

Our criteria for including digital/data programs in our data were that the program titles should have keywords "digital" or "data", and "curation." For example, the School of Information at Pratt Institute, offers an advanced certificate program called "Conservation and Digital Curation." Since this program contains the keywords "digital" and "curation", we included it in our data. Descriptions for these programs, their required courses and/or recommended courses, were collected.

Among the 11 certificate programs illustrated in Figure 1, three of them have a small gray box attached to them with arrows pointing to their corresponding digital/data curation course(s). The other eight programs offer various other courses such as "digital preservation," "database management," "digital libraries," or "privacy and security in the networked world." However, those programs did not include digital

or data curation courses specifically. The number of data curation versus digital curation programs was not balanced. Compared to the nine digital curation programs, we found only two data curation programs: PGM_UIUC and PGM_UNC_CH.

Course-level descriptions

In total, 16 courses fell under this category. Similar to the selection criteria for the Program-level descriptions mentioned above, courses in this category consisted of the digital/data curation courses that have "digital" or "data" and "curation" in their title. Four courses – CRS_CUNY, CRS_Rutgers, CRS_UCLA, and CRS_UNC_CH – were data curation courses, and the other 12 were digital curation courses (Figure 1) showing that the number of digital curation courses were three times more than that of data curation courses. Two digital curation courses – CRS_McGill and CRS_U_Toronto – were offered by Canadian institutions.

Depending on the needs and interest, institutions may offer individual digital/data curation courses or offer such courses as part of their digital/data curation certificate programs. Four courses – CRS_UNC_CH (data curation), CRS_UMD_Policy, CRS_UMD_Principle, and CRS_UNC_CH (digital curation) - were part of their associated digital/data curation programs, and these relationships were visualized with the arrows connecting the related programs and courses in Figure 1. More specifically, the University of Maryland provided two digital curation courses with two different foci – CRS_UMD_Policy and CRS_UMD_Principle – under their PGM_UMD certificate program. University of North Carolina at Chapel Hill provided one course in data curation, and another in digital curation. The other 12 courses in white rectangle boxes did not have any associated programs, and thus no arrow connections from those boxes were present.

Topics from the Digital/Data Curation Programs

As explained in the Methodology section, we set the number of topics to be 20, and the number of words in each topic to be ten. These parameters were selected empirically, to ensure both the generality and uniqueness of the generated topics (See Table 1 in the Appendix section, LDA Topics). To help understand the latent meaning of the LDA topics and to alleviate the limitation of applying LDA to small dataset, we manually developed the corresponding inferred topics by working together with a collaborator, who is a graduate student in the field of Library and Information Science, to reduce subjectivity in assigning those inferred topics. Our collaborator accessed all the original data (i.e., syllabi, course/program descriptions) used in this study and had fair understanding of the subject matter prior to assignment of inferred topics. These topics were examined by the authors for accuracy, and then refined by the collaborator in case the inferred topic did not represent the LDA topic well.

Once the inferred topic assignments were finalized, they were placed under Inferred Topics column in Table 1 (See Appendix) next to the LDA Topics column.

These inferred topics also cover the five broad topical categories below:

- **System**: archive/repository systems
- **Technology**: archives/repositories and their technologies
- **Content**: type and preparation of multimedia/digital content

 doi:10.2218/ijdc.v14i1.586

- **Policy and strategy**: policies and strategies for accessing, acquiring, and managing the system
- **Future trend:** emerging skills and future career in digital/data curation

Topics from the Digital/Data Curation Courses

An application of LDA to the course descriptions resulted in 20 topics as well (see Table 2 in the Appendix section). Again, we manually developed the corresponding inferred topics and placed them next to LDA topics to help understanding of the readers. Similar to the case of program-level topics in the previous section, these course-level topics also cover the five broad topical categories again. However, some topics only appeared in the course-level topic list. The comparison of topics between the program- and course-level is illustrated in the next section.

Comparing the Program-Level and Course-Level Topics

Although the program-level (Table 1) and course-level (Table 2) topics have somewhat different topical scopes, we attempted to compare them to identify common, as well as unique, topics (Figure 2). The detail steps for creating Figure 2 are as follows:

1. Application of LDA to generate 20 LDA topics
2. Generation of inferred topics by working together with an external collaborator
3. At this point, we have Tables 1 and 2 with inferred topics
4. Further refinement and/or merging of inferred topics in Tables 1 and 2 to compare topics between programs and courses more effectively
5. Creation of Figure 2 by using the refined inferred topics from the Step 4 above

Due to slightly varied terms between inferred topics from Tables 1 and 2, we first had to further refine those topics in both tables using the same vocabularies. The topics in each table were merged together if needed. For example, two inferred topics in Table 1, *3. Archival resources and tools* and *6. Archival repositories*, were combined into Archive/collection. In addition to the topic merging within each table, we also identified semantically-similar topics across Tables 1 and 2, and assigned a common vocabulary for them. For example, *16. Access management* in Table 2, became Accessibility, and this became the common topic between the two topic tables. Also, *13. Multimedia content* and *20. Digital content design and marketing* in Table 1 were identified as a semantically-similar topic to *8. Library resources* in Table 2, thus a common vocabulary multimedia/digital content/library resource replaced those three topics. Finally, once the refining of topics for both tables was completed, unique as well as common topics were identified as shown in Figure 2.

The two-thirds of the program-level topics were overlapping with the two-thirds of the course-level topics as listed in the intersecting area of the two circles, each representing program-level and course-level. However, there were five topics that were unique in each of program- and course-level.

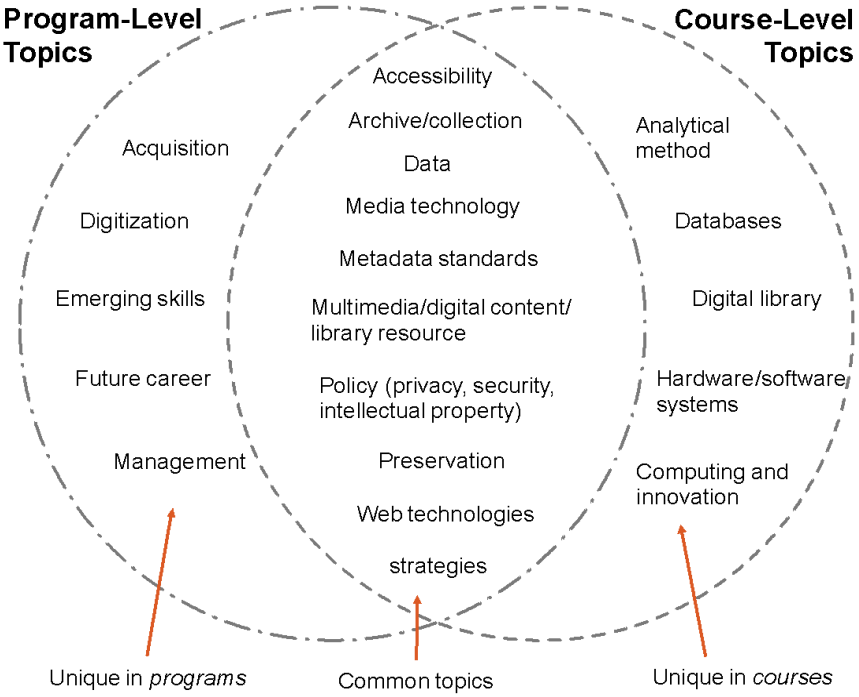

Figure 2. Common and unique topics identified from the program-level and course-level descriptions after manual refinement process.

Visualizing and Comparing Topics from Digital and Data Curation Domains

In order to portray the relationships between topics and programs/courses, we developed two interactive visualizations: (1) program-level topic view; and (2) course-level topic view. Each visualization has a Matrix view and a Bar chart view as discussed in the Methodology Section. The Matrix view presents the distribution of 20 topic IDs (numbered 1-20) across programs or courses. The Bar chart view lists the topic IDs with corresponding topic phrases along with their frequency across programs/courses. Our interactive visualizations offer a macroscopic view of the topical distribution in the program- and/or course-level descriptions. In addition, the visualizations also allow a program- and course-level detail investigation of the topics using the sorting features integrated in the visualization.

Interactive Visualization of Program-Level Topics

Figure 3 illustrates a distribution of 20 topics among the digital/data curation programs in its Matrix view on the left. The topics are organized in the order of their frequency across programs in its Bar chart view on the right.[4] The Matrix view shows the 11 program names in its columns and corresponding topic IDs in its rows. Data curation program names and their circular topic markers are colored in red, and digital curation programs are in blue. The Bar chart view lists the 20 inferred topics along with their frequency across program descriptions. The most frequent topic 'data, systems, and their

4 We also made this visualization accessible online at: http://bagua.cct.lsu.edu/vis/topic/program_topics.html

 doi:10.2218/ijdc.v14i1.586

applications' appears in seven programs. Other frequent topics include: acquisition, description, preservation, and access methods; digitization and other practical issues; archival resources and tools; or metadata standards. These topics appeared six times across programs. Except for the topic 'archival resources and tools', these frequent topics appear both in the data curation and digital curation programs. In contrast, three topics, 'concentrations for future careers', 'collection development', and 'understanding the Web', appear only once in a digital curation programs.

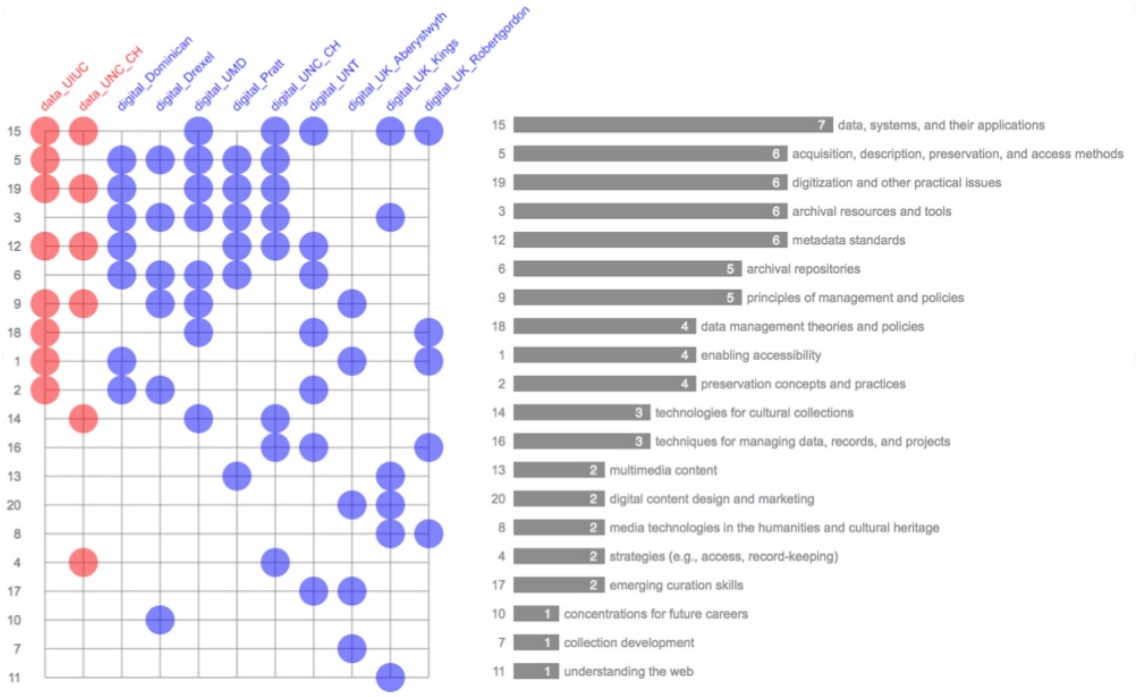

Figure 3. An interactive visualization for inferred topics from program descriptions.

The sorting feature of the Matrix view allows the topic IDs in row headings, or program names in column headings to be realigned by clicking them. For example, all the topics of data_UIUC are grouped together when the program name, data_UIUC, is clicked (see Matrix view in Figure 4). Clicking a program name also highlights the corresponding topic IDs and topics in the Bar chart view in red color (see Bar chart view in Figure 4), making it easier to read them. If the program name is clicked again, it toggles and all the topics that are not in data_UIUC will be grouped together closer to the program name. In addition, the Bar chart view provides a drop-down menu for sorting the listed 20 topics in an ascending or descending order of their frequency.

Once all the topics of a program were grouped together, as shown for the data_UIUC topics in Figure 4, the topics can be compared against those in other programs to identify the most (dis)similar data or digital curation programs. In our example, data_UIUC shared five topics with digital_Dominican and digital_UMD programs. However, it only shared a single topic, *data, systems, and their applications*, with digital_UK_Kings. data_UIUC shared three to five topics with the digital curation programs in the U.S., but it shared one to three topics with the programs in the U.K. In this way, the topical similarities between programs can be identified using the visualization.

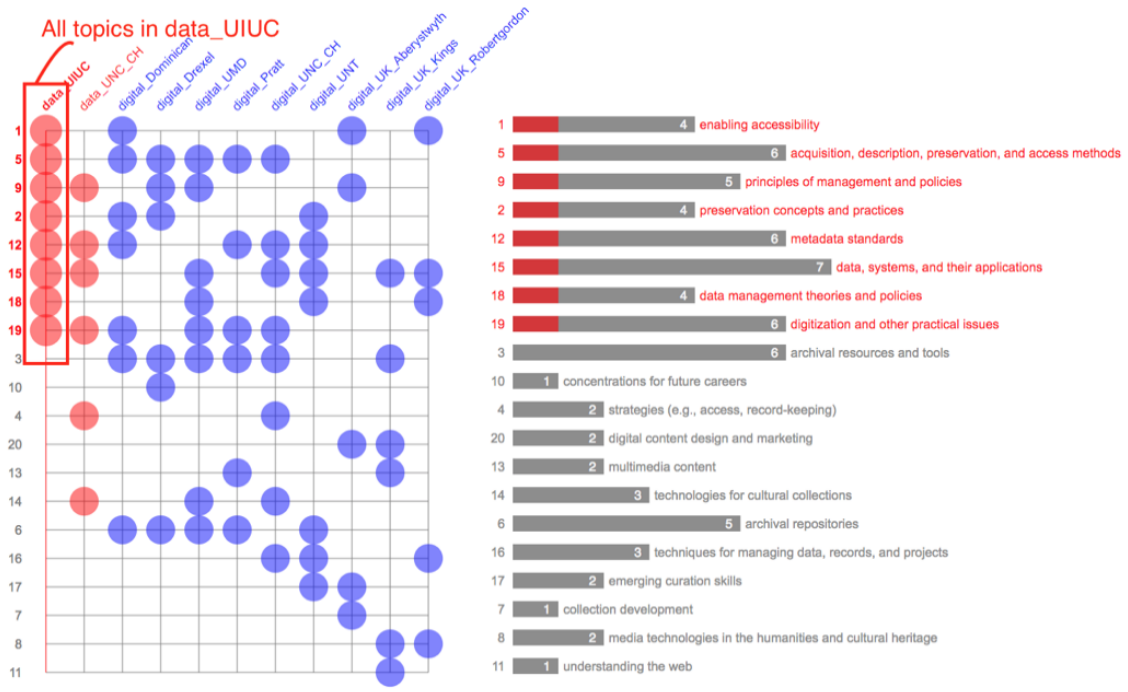

Figure 4. The program-level visualization showing all the topics of *data_UIUC* grouped together.

Interactive Visualization of Course-Level Topics

We applied the same visualization approach for the digital/data curation course descriptions as well.[5] Figure 5 presents a distribution of 20 topics among the digital/data curation courses. The topics are organized in a descending order of their frequency across courses in the Bar chart view. The Matrix view shows course names in its columns and corresponding topic IDs in its rows. Following the similar color coding approach from the previous program-level visualizations, the data curation course names and their topic IDs are colored in red and digital curation courses in blue. The most frequent topic found at the course level was *access management*, which appeared in ten courses. Other frequent topics include: *strategies for repositories* (appeared in nine courses); *digital library services and properties* (eight courses); *application of analytical methods* (eight courses); *preservation* (eight courses); or *library resources* (eight courses). Three topics, *computing and innovation*, *software and storage system*, and *media archives*, appear only once across digital/data curation courses.

5 The course-level topic visualizations are also accessible online at: http://bagua.cct.lsu.edu/vis/topic/course_topics.html

 doi:10.2218/ijdc.v14i1.586

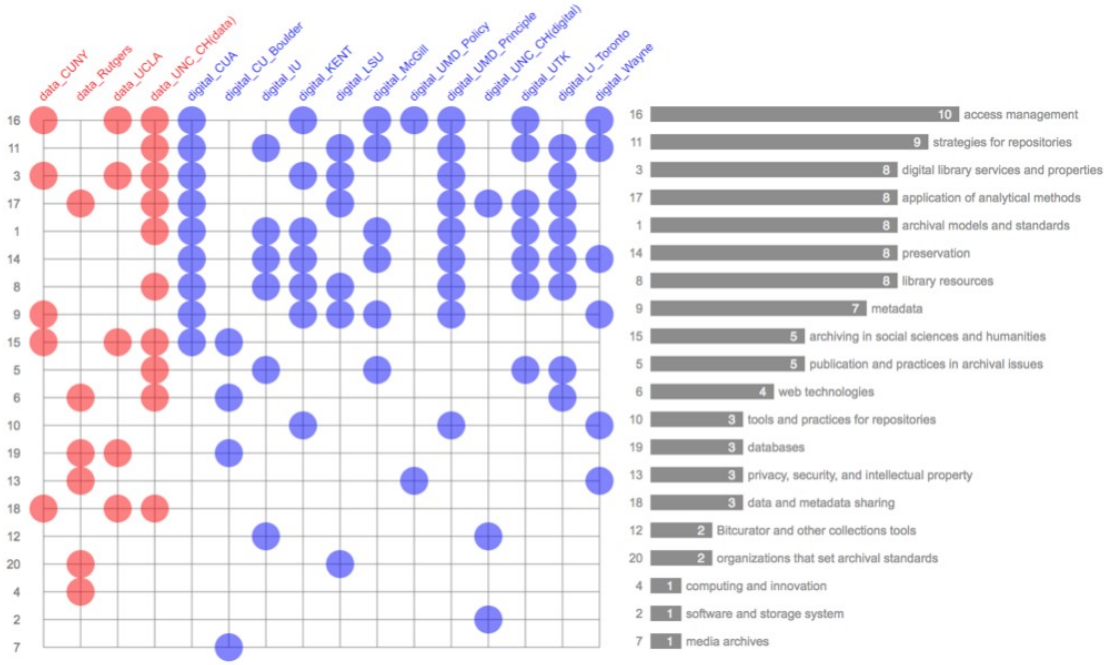

Figure 5. An interactive visualization of inferred topics from the course-level descriptions.

Topics can be sorted in the course-level visualizations as well. In the Matrix view of Figure 6, all the topics of digital_LSU are grouped together when the course name is clicked. Clicking the course name also highlights the corresponding topic IDs and topics in the Bar chart view in blue (Figure 6), making it easier to read them. Again, the Bar chart view provides a drop-down menu for sorting the listed 20 topics based on their frequency.

The course topics are compared again to identify the most (dis)similar data or digital curation courses using the Matrix view. In Figure 6, digital_LSU shared five topics with digital_CUA, digital_UMD_Principle, and digital_U_Toronto courses. However, it only shared a single topic, *application of analytical methods*, with digital_UNC_CH(digital). If compared to four data curation courses, data_UNC_CH(data) was the most topically similar course to digital_LSU sharing four topics. With data_UCLA, digital_LSU shared only one topic, *digital library services and properties*. digital_LSU shared one to five topics with the digital curation courses, and one to four topics with the data curation courses in the U.S.

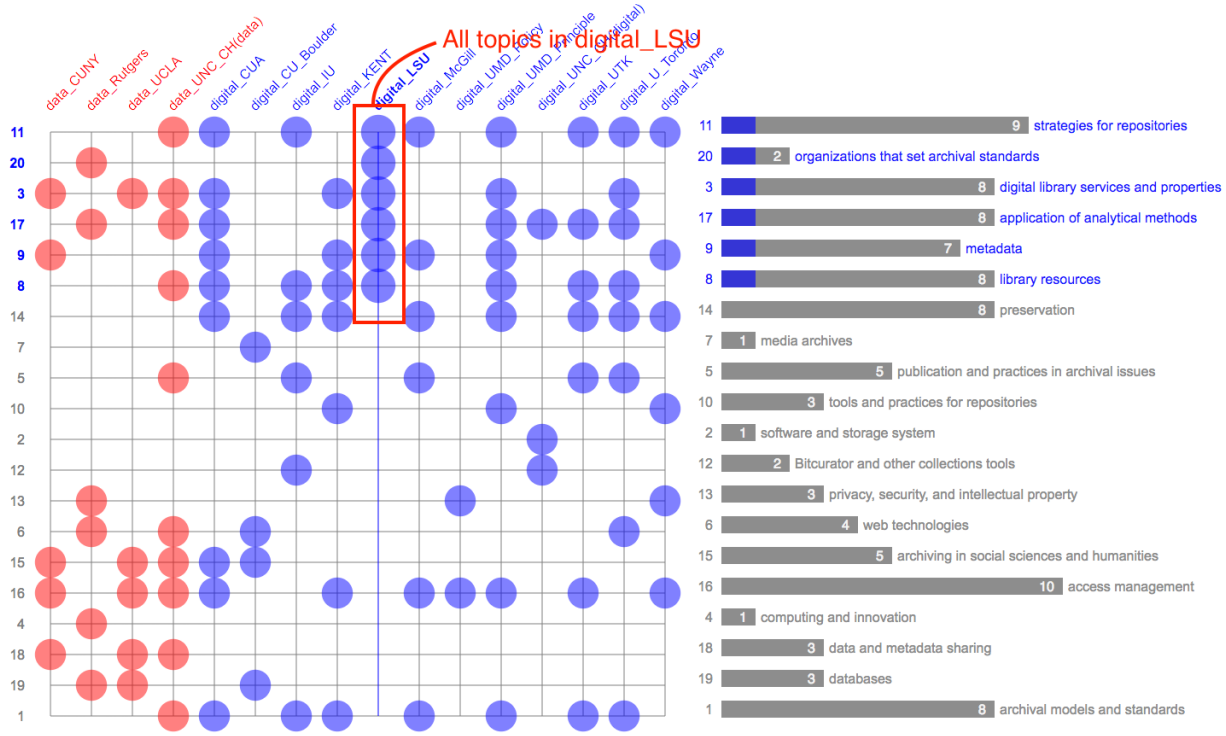

Figure 6. The course-level visualization showing all the topics of digital_LSU grouped together.

Discussion

Joint Approach of Topic Modeling and Visualization

We found that our integrated approach of topic modeling and visualization using the program- and course-level descriptions was an effective approach for uncovering and presenting the topical scopes of digital/data curation programs and courses currently offered. The topic modeling using LDA allowed us to extract common and unique topics from multiple textual documents. LDA topics are supposed to be latent, and each topic consists of ten observable keywords in this study. Therefore, it was not always convenient to infer a single keyword or phrase from each LDA topic. Our approach for addressing this issue was to infer a single phrase that may represent the ten terms in each LDA topic. Once the inferred topics were developed, we could directly use them in our visualizations. After a further refinement of the inferred topics from program and course descriptions using common vocabulary, topics between the program-level and course-level were also compared, as shown in Figure 2.

Our visualizations allowed us to conduct a comprehensive analysis of topics in both macroscopic and microscopic ways. Macroscopically, we observed the overall distribution of curriculum topics across all programs and courses. Microscopically, we could drill down to examine specific curricular topics and associated programs and courses. By clicking the row or column heading in the Matrix view (or by selecting a menu item), we could sort the order of the topics by their frequencies across programs

 doi:10.2218/ijdc.v14i1.586

and courses. Similarly, we could also group the columns (i.e., according to programs or courses) by the occurrence of topics. Sorting features in both Matrix and Bar chart views enabled us to repeatedly reorganize the rows and columns so that we could observe emerging patterns that show how specific curriculum topics occur in a set of digital/data curation programs and courses. This capability for reorganizing the screen view was essential for transforming seemingly random patterns, at least initially, into increasingly coherent and meaningful co-occurring groups across a range of curriculum topics.

Specifically, an emerging pattern of topics revealed interesting clusters and the distribution of topics (red and blue dots) on the Matrix view. Observing such a pattern is valuable for determining which topics might be essential in designing future digital/data curation courses. Other topic circle clusters can also be useful for illuminating some of the under-utilized curriculum topics that have yet to be adopted by certain programs and courses.

Visual Analysis of Topics

Our visualization approach enabled us to see how latent topics are being employed by both program-level and course-level curricula. As we introduced in the Results section, the 20 topics from digital/data curation curricula can be broadly classified into the five high-level categories: *System*; *Technology*; *Content*; *Policy and strategy*; and *Future trend*.

As illustrated in Figures 7 and 8, we were able to observe distinct visual patterns of topics (circles on the visualizations) which indicate how frequently or infrequently different sets of topics were being adopted among the programs and courses. Note that in Figures 7 and 8, one can distinguish clusters of circles, where a grouping of circles represents topics that are populated more densely (Figures 7.A and 8.B) than the other sparse groupings of topics (Figures 7.B and 8.C). Specifically, these patterns allow us to quantify and assess consistently-used, as well as under-adopted curriculum topics.

Consistently-used topics

We were able to observe a set of curriculum topics that were employed more consistently across the program- and course-level descriptions. At the program level, five topics in the *Policy and strategy* category were utilized (Figure 7.A). Additionally, we observed that the topics in the *Technology* category were also employed consistently across different programs. However, at the course level, the *Policy and strategy* category represents the most popular topics among the digital/data curation courses, and four topical categories (including *System*, *Technology*, *Content*, and *Policy and strategy*) were somewhat spread over the digital curation courses (Figure 8.A), indicating that they were adopted consistently among many courses (the exception being the digital curation course at the University of Colorado, Boulder).

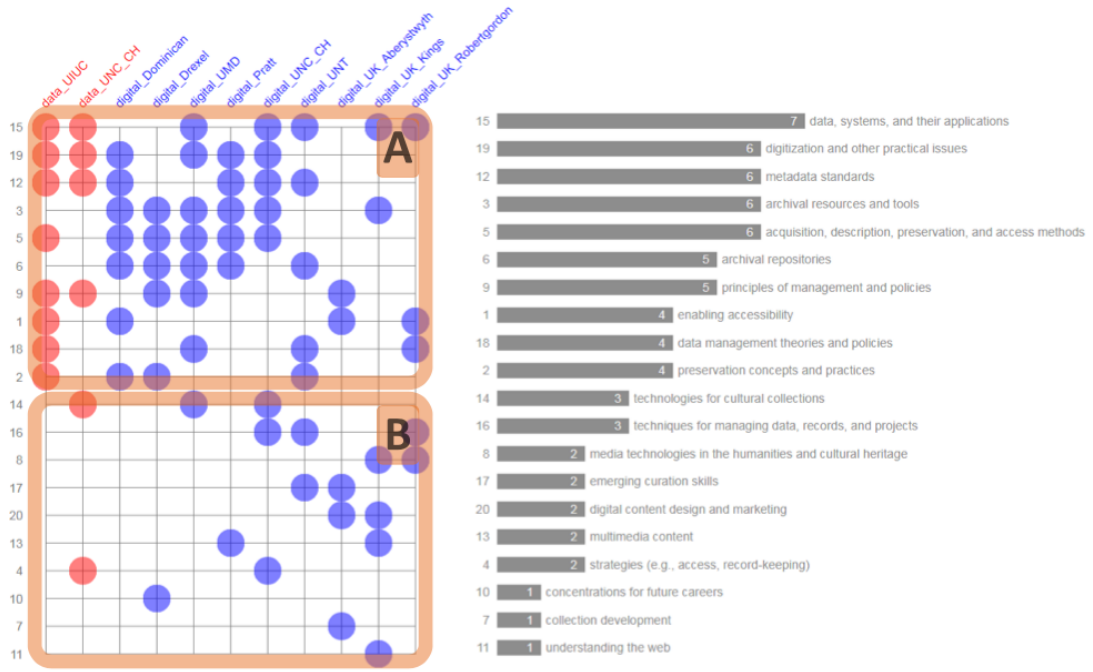

Figure 7. Program-level topic visualization for digital/data curation programs. (A) Consistently-used topics; (B) Under-adopted topics.

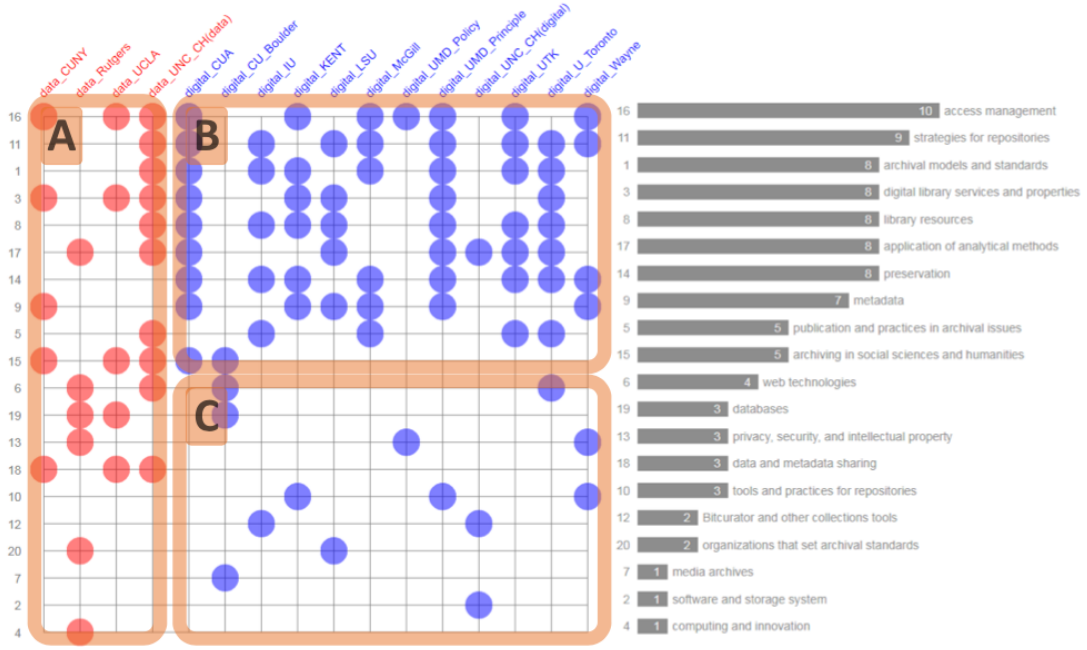

Figure 8. Course-level topic visualization. (A) Adopted topics for data curation courses; (B) Consistently-used topics for digital curation courses; (C) Under-adopted topics for digital curation courses.

 doi:10.2218/ijdc.v14i1.586

Under-adopted topics

As shown in Figures 7.B and 8.C, a set of topics was adopted by only a smaller number of programs (one to three programs), which refers to the under-adopted curriculum topics. In terms of the program-level descriptions, the emerging curation skills and future career topics (i.e., the *Future trend* category) were covered by only two of the 11 programs (Figure 7.B). Additionally, the content of curated data (e.g., multimedia content or humanities and cultural heritages) and the digital concept design were adopted by only two programs.

It is also of note that the topics in the *Technology* and *System* categories (Web technologies, database, software tools, data sharing methods, etc.) were rarely mentioned in the course-level descriptions. This finding may indicate that current digital/data curation curricula tend to focus more on the aspects of policies, strategies, or services, as opposed to both system and technology topics. Interestingly, in the same course-level visualization (Figure 8.A), the data curation curriculum (the red circles) appeared to be rather spread out over the four data curation courses – with the exception of the data curation course at Rutgers University, which had many unique technical topics, probably due to its focus on technical aspect of digital/data curation curriculum.

Limitations of this Study

One of the limitations of this study might be our use of program descriptions (merged with their core course descriptions), as well as individual course descriptions for topic modeling. We assumed that the program and course descriptions would include most of the topics that were taught in the core courses of a program or in the individual courses. However, it is possible that some topics were included only in the weekly schedules in the syllabus, reading lists, or in the lecture slides, and thus we missed them in our topic modeling results and visualizations.

Another limitation might be a subjectivity issue in inferring topics from the LDA-generated topics. In order to reduce subjectivity as much as possible, we worked with an external collaborator to develop a single inferred topical word or phrase for each LDA-generated topic of ten words. However, there still exists the possibility that semantically-different inferred topics could have been identified, assigned, and agreed by all of the researchers involved in the topical inference.

This study was also limited by our selection of programs and individual digital/data curation courses and syllabi written in English due to our own language limitations. There exists huge potential in collecting and analyzing digital/data curation data written in European or Asian languages. However, we reserve this as one of our future studies.

Applicability to Other Academic Fields

Our joint approach of topic modeling and visualization allowed us to identify and present the topical scopes of digital/data programs and courses. Thus, it has a potential for analyzing and visualizing topics from other academic programs and courses as well. For example, social media/network analysis is a multi-disciplinary field taught in the Information Science, Computer Science, Communications, or Social Science fields. Depending on the field, however, the main topics in the course may vary. The course offered in Computer Science might be focused more on the technical and data analytical aspects, whereas those offered in Information Science or Social Science may focus more

on information use, communications, or social perspectives. By applying topic modeling and visualizations to collected descriptions and syllabi of social media/network analysis courses, we may develop a broader and unified topical scope that might be used in multiple relevant fields. Using this broad topical scope, instructors, who are currently teaching related courses, may identify topics that are missing from their courses and add them to provide more comprehensive courses. The syllabus analysis for finding the popular textbooks and readings may also help instructors who are designing or upgrading their courses with newer resources.

Conclusions and Future Work

In this study, we presented an analysis of topical scopes for digital/data curation programs and courses by applying topic modeling and visualization approaches to the descriptions of 11 certificate programs and 16 courses. After identifying and inferring topics by applying LDA topic modeling, we developed interactive visualizations for a visual analysis of the topical patterns. Using the sorting feature provided by the Matrix view and Bar chart view in each visualization, a cluster of shared topics, as well as unique topics from data/digital curation programs and courses, were able to be quickly viewed in a macroscopic perspective. The Matrix view also enabled the comparison of a pair of courses, based on their topical similarities in a microscopic perspective.

Our approach was effective and meaningful in that the identified topical scope for digital/data curation curricula uncovered fundamental components of the curatorial tasks in the curriculum. This approach may also be applicable to other types of educational opportunities such as workshops, massive open online courses, or webinars in digital/data curation or any other academic domain.

For researchers, we share the data used in this study, which include descriptions of courses and programs, a combined list of readings and textbooks, as well as the HTML and JavaScript visualization codes[6]. We plan to extend this study by including analyses for more curricula and other types of opportunities. Our future research may lead to a more accurate and comprehensive picture of topical scopes for digital/data curation and their topical relationships.

References

Beagrie, N. (2008). Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*, *1*(1), 3-16.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Brown, D.J. (2009). International Council for Scientific and Technical Information (ICSTI) Annual Conference – Managing data for science. *Information Services & Use*, *29*(4), 103-21.

6 See: https://github.com/seungwonyang/IJDC

Buneman, P., Cheney, J., Tan, W., & Vansummeren, S. (2008). Curated database. PODS'08, June 9–12, 2008, Vancouver, BC, Canada. *Council on Library and Information Resources*. Data curation. Retrieved from http://homepages.inf.ed.ac.uk/opb/papers/inv.pdf

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., & Blei, D.M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems* (pp. 288-296).

Chuang, J., Gupta, S., Manning, C. & Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on machine learning (ICML-13)*, 612-620.

Chuang, J., Manning, C.D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 74-77. ACM.

Chuang, J., Ramage, D., Manning, C.D., & Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the ACM Human Factors in Computing Systems* (CHI). Retrieved from http://vis.stanford.edu/papers/designing-model-driven-vis

Chuang, J., Ramage, D., McFarland, D.A., Manning, C.D., & Heer, J. (2012). Large-scale examination of academic publications using statistical models. In *Advanced Visual Interfaces Workshop* (AVI Workshop 2012).

Digital Preservation Coalition (DPC). (2009). *Introduction to digital preservation handbook*. London: Digital Preservation Coalition, Retrieved from https://www.webarchive.org.uk/wayback/archive/20090317142641/http://www.dpconline.org/docs/handbook/DPCHandbook.pdf

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

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228-5235.

Hall, D., Jurafsky, D., & Manning, C.D. (2008). Studying the history of ideas using topic models. *EMNLP*, 363-371.

Harris-Pierce, R. & Liu, Y. (2012). Is data curation education at library and information science schools in North America adequate? *New Library World, 113*(11-12), 598-613.

Harvey, R. & Bastian, J. (2012). Out of the classroom and into the laboratory: Teaching digital curation virtually and experientially. *IFLA Journal*, *38*(1), 25-34.

Heidorn, P. (2011). The emerging role of libraries in data curation and e-science, *Journal of Library Administration*, *51*(7/8), 662-72.

Higgins, S. (2011). Digital Curation: The emergence of a new discipline. *International Journal of Digital Curation*, *6*(2), 78-88.

Hirtle, P. (2010). The history and current state of digital preservation in the United States. In *Metadata and Digital Collections: A festschrift in honor of Thomas P. Turner*, 124. Ithaca, NY: Cornell University Library.

Hockx-Yu, H. (2006). Digital preservation in the context of institutional repositories. *Program: Electronic Library and Information Systems*, *40*(3), 232-243.

Kim, J. (2015). Competency-based curriculum: An effective approach to digital curation education. *Journal of Education for Library and Information Science*, *56*(4), 283-297.

Lee, C. (2009). *Matrix of digital curation knowledge and competencies*, Version 13. Retrieved September 10, 2016, from http://ils.unc.edu/digccurr/digccurr-matrix.html

Lee, C., Tibbo, H., & Schaeffer, J. (2007). Defining what digital curators do and what they need to know: The DigCCurr project. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries*, 49-50.

McCallum, A.K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from http://mallet.cs.umass.edu/

Ogburn, J. (2010). The imperative for data curation. *Portal: Libraries and the Academy*, *10*(2), 241-246.

Ramage, D., Hall, D., Nallapati, R., & Manning, C.D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248-256. Association for Computational Linguistics, Stroudsburg, PA.

Ray, J. (2009). Sharks, digital curation, and the education of information professionals. *Museum Management & Curatorship*, *24*(4), 357-368.

Ray, J. (2012). The rise of digital curation and cyberinfrastructure. *Library Hi Tech*, *30*(4), 604-622.

Steyvers, M. & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*, *427*(7), 424-440.

Swan, A. & Brown, S. (2008). The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs. Report to the JISC, Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf

Talley, E.M., Newman, D., Mimno, D., Herr, B.W., Wallach, H.M., Burns, G.A.P.C., Leenders, A.G.M., & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, *8*(6), 443-444.

Tibbo, H. (2015). Digital curation education and training: From digitization to graduate curricular to MOOCs. *International Journal of Digital Curation*, *10*(1), 144-153.

Yakel, E. (2007). Digital curation. *International Digital Library Perspectives*, *23*(4), 335-340.

Yakel, E., Conway, P., Hedstrom, M. & Wallace, D. (2011). Digital curation for digital natives. *Journal of Education for Library and Information Science*, *52*(1), 23-31.

Appendix

Table 1. Inferred and LDA-generated topics from the program-level descriptions.

Topic ID	Inferred Topics	LDA Topics
1	Enabling accessibility	knowledge professional libraries long core enable opportunity integrity accessibility range
2	Preservation concepts and practices	provide preservation concepts science disciplines degree practice competencies info hours
3	Archival resources and tools	archives theoretical resources tools studies electronic museums based society implement
4	Strategies (e.g., access, record-keeping)	strategies include special dr taken find instructor professor master faculty
5	Acquisition, description, preservation, and access methods	preservation library access description needs methods material acquisition focusing organization
6	Archival repositories	materials organizations records archival overview repositories learn introduces objects study
7	Collection development	development required collection support explore semester organizations organized public architecture
8	Media technologies in the humanities and cultural heritage	media technologies field humanities assets challenges heritage developing dynamic industry
9	Principles of management and policies	principles specific management system policies experience services future discovery computer
10	Concentrations for future careers	work relevant manage create concentration working well careers preserve projects
11	Understanding the web	web understanding social structured internship texts process publishing text case
12	Metadata standards	metadata standards program university prerequisite curriculum preserving essential offered prepares
13	Multimedia content	content art approaches range history

 doi:10.2218/ijdc.v14i1.586

		image roles implementation video critical
14	Technologies for cultural collections	collections professionals cultural technology creation technological variety school individuals specialization
15	Data, systems, and their applications	data systems current academic plan applications intellectual quality explores activities
16	Techniques for managing data, records, and projects	develop techniques apply project managing evaluate ability provision north completing
17	Emerging curation skills	skills ensure emerging understand processes organization curating technical responsible design
18	Data management theories and policies	management application representation settings theory address focuses planning user policy
19	Digitization and other practical issues	issues practical practices analysis digitization introduction organizational database historical institutions
20	Digital content design and marketing	content design includes marketing asset practice domain successful culture security

Table 2. Inferred and LDA-generated topics from the course-level descriptions.

Topic ID	Inferred Topics	LDA Topics
1	Archival models and standards	model archival planning cultural magazine technology standards heritage collections approaches
2	Software and storage system	software disk system investigation Addison Garfinkel code image Richard building
3	Digital library services and properties	libraries services article social jisc plans needs plan properties foundation
4	Computing and innovation	nature computing climate paul historical volume study times innovation text
5	Publication and practices in archival issues	publications reports pub council issues pubs practice article London content
6	Web technologies	web online center work eds challenges future en network sources

7	Media archives	media archive archives storage internet optional memory archaeology culture blogs
8	Library resources	library resources national review American archiving centre sites default initiative
9	Metadata	metadata archives technical lifecycle risk paper libraries understanding study selection
10	Tools and practices for repositories	content repositories tool practice communities building cloud manual see Schuman
11	Strategies for repositories	repositories repository documents open strategies oais storage formats clir reference
12	Bitcurator and other collections tools	bitcurator computer collections images academic access Wesley lab tools guide
13	Privacy, security, and intellectual property	privacy security policy copyright intellectual issues focus property records cases
14	Preservation	preservation materials activities objects case oclc trusted certification assignments focus
15	Archiving in social sciences and humanities	practices policies sciences role board social archiving knowledge humanities society
16	Access management	management access institutional humanities guide Washington discuss arl criteria development
17	Application of analytical methods	read analysis conference files introduction proceedings watch based concepts media
18	Data and metadata sharing	sharing points tasks provide post activity metadata contributions point feedback
19	Databases	public time databases projects north learn place economics archivists today
20	Organizations that set archival standards	archive icpsr web de ukdataservice oai pmh weibel google basic