Heather Moulaison Sandy, Heather Froehlich, Cynthia Hudson-Vitale, Denice Adkins. 2019. Topic Modeling and Facet Analysis of an Emerging Domain: Research Data Management and Data Curation. NASKO, Vol. 7. pp. 63-76.

Heather Moulaison Sandy — University of Missouri
Heather Froehlich — Pennsylvania State University
Cynthia Hudson-Vitale — Pennsylvania State University
Denice Adkins — University of Missouri

Topic Modeling and Facet Analysis of an Emerging Domain: Research Data Management and Data Curation

Abstract

Research data management (RDM) is often seen as the overarching field that permits research data to be managed, and is related to the field of data curation (DC), a subset of digital curation. Together, RDM and DC (RDM/DC) allow information professionals to work with clients and each other to make data available in support of the research enterprise. An emerging area of scholarly communication, RDM/DC represents a rich area of study from the perspective of knowledge organization (KO). This paper explores the following research question: What can facet analysis tell us about the emerging field of RDM/DC?

First, the MAchine Learning for LanguagE Toolkit (MALLET) implementation of Latent Dirichlet Allocation (LDA) is used for topic modelling of abstracts of the RDM/DC scholarly literature. A preliminary analysis of this empirical data by the research team yields a number of topics and, when possible, their relevant aspects or contexts. Facet analysis principles are next applied to these results, producing four general facets: Practice, Stakeholders, Resources, and Study of RDM/DC; however, complex notions infused throughout the field such as "services" and "metadata" do not appear outright in the analysis. Each facet is then further explored through logical division, and the resulting system is encoded in Protégé and visualized using WebVOWL. We conclude that the major areas of emphasis in this data-intensive field will be fundamentally of interest to those in LIS, in scholarly communication, and perhaps increasingly, in KO and other fields that manage and make available data of all kinds.

Introduction

Data is essential to supporting open science and ethical research, allowing for reproducibility of research initiatives and for reuse, potentially, of data by others (e.g., Darch and Knox 2017). Research data management (RDM) is often seen as the overarching field that requires research data to be managed in accordance with funder mandates, journal requirements, disciplinary practices, etc. (Whyte and Tedds 2011). The closely related area of data curation (DC) is a sub-set of digital curation (Johnston 2017), an area that likewise supports the management of research data through efforts relating to digital preservation, handling, and sharing of the data according to established best practices in the information professions.

Together, these aspects of RDM and DC (RDM/DC) allow information professionals to work with their clients and with each other to make data available in support of the research enterprise. An emerging area of scholarly communication in terms of both research and practice, RDM/DC is a rich area of study from the perspective of knowledge organization (KO). RDM/DC is a field that has been systematically investigated for a number of years, especially from the policy and case study perspective. At the present point in this, a holistic overview of the field remains under-surveyed, leaving a gap in our understanding of the field and its evolution to date.

RDM/DC Defined

Interest in research data management has been voiced repeatedly over the years, including in works designed to support research methods (e.g., Denzin and Lincoln 2000). When considering data, "Stakeholders include researchers themselves, their institutions, their discipline-based communities and repositories, publishers and the general public" (Steinhart 2013, 19). Information professionals play a key role in supporting research data management. Research data management was arguably first studied in a systematic way by Tenopir, Birch, and Allard (2012) through a grant received from the IMLS. RDM "concerns the organisation of data, from its entry to the research lifecycle through to the dissemination and archiving of valuable results" (Whyte and Tedds 2011, 1) As such, it includes aspects of digital curation and preservation, as well as an understanding of the scholarly communication process, the research process, and other aspects relevant to the domain's data and its preservation and future use by specialized scholars and researchers.

The related field of data curation can further be seen as a subset of digital curation (Johnston 2017), an area closely related to research data management. Digital curation and its correlate, digital preservation, are mature fields that have been systematically studied by information since archives have been collecting digital objects for long term access and use. Meanwhile, the sub-field of data curation can be considered from two distinct perspectives: the user's perspective (describing policies and practices that pertain to researchers) and from the information professional's point of view (including the technologies, software packages, standards, practices and tools to support access) (Johnston 2017).

Because of the relatively recent emergence of these dynamic and growing areas of study in LIS, a gap exists in our understanding of the emerging field of RDM/DC.

Research question

In light of the limited understanding of the emerging field of RDM/DC, the current research project explores the following research question:

RQ: What can facet analysis tell us about the emerging field of RDM/DC?

Facet Analysis

Faceted classification is a longstanding traditional approach to classification in KO (e.g., Vickery 2008). Facet analysis, one aspect of the process of creating a faceted classification scheme (a priori, a scheme promoting retrieval on the part of users), has recently been analyzed and debated (Hjørland 2013) in the KO literature.

Applying Facet Analysis in KO

The KO literature is rich with approaches for implementing facet analysis in support of practical knowledge organization systems (KOSs) to promote retrieval. Cheti and Paridisi (2008) use facet analysis to revise a verbal subject indexing system used in Italian libraries as a way of producing a fully-faceted thesaurus. In a similar vein, Spiteri (2000) investigates the use of facet analysis in the creation of a number of commercial thesauri, finding that there was "as yet no consensus amongst thesaurus designers about the best way in which to apply

facet analysis" (45) despite its ubiquity. Facets and facet analysis are currently of interest in computer science due to the emergence of the world wide web and new ways for generating and displaying the classification (Priss 2008) and ways in which faceted classification supports retrieval (Vickery 2008) in online environments (Slavic 2008; Slavic and Davies 2017). It can also be beneficial in interdisciplinary environments given the traditional approach of standard universal schemes such as Dewey Decimal Classification, Universal Decimal Classification, and Library of Congress Classification which are discipline-based (Gnoli 2008).

In addition to its ability to structure KOSs for retrieval, facet analysis can also be used in a somewhat novel way to explore and intellectually map new fields of study and inquiry. Facet analysis is closely tied to faceted classification and retrieval; it can also be used as a tool for understanding the relationship between classes and between things (Buchanan 1980). Ranganathan is credited with the formalization of the analytico-synthetic approach that FA supports. Ranganathan, according to Hjørland (2013), expresses the following views, among others: "That the discovery of new knowledge implies the need for new classes, which cannot be anticipated by an enumerative system" and "newly discovered knowledge can be expressed in FC designed before the discovery is made by combinations of preestablished categories" (548). For these reasons, facet analysis can be used to explore an emerging area. To this end, Shiri (2014) uses facet analysis to assess the emerging field of big data, and to draw conclusions about relationships. Ultimately, his paper seeks to "create conceptual and concrete links between information science and knowledge organization methods and traditions and the emerging area of big data" (367). Shiri's usage of facet analysis in this way is consistent with one of the reasons for which ontologies are created, "to analyze domain knowledge" (Tonkin, Pfeiffer, and Hewson 2010, 1).

In this spirit, the current research project continues in this new tradition and adopts an approach similar to that of Shiri (2014), along with that of Tonkin, Pfeiffer, and Hewson (2010)'s quest for domain knowledge through the use of facet analysis; we do this as a way to investigate the emerging field of RDM/DC.

Steps to the Creation of a Faceted Classification Scheme Using Facet Analysis

Buchanan (1980) indicates the stages necessary for the construction of a faceted classification scheme. These stages require the classificationist to:

1. examine a representative sample of the literature, to discover the elemental classes its authors deal with;
2. group these 'isolates' ([Ranganathan, 1959]) (that is, as-yet-unorganised elemental classes) into facets when they become foci;
3. if necessary, apply different characteristics of division to facets to produce subfacets;
4. place the foci in each facet or in each subfacet into order, using broader and narrower order for foci in that relationship or in an appropriate order in array for coordinate foci;

5. place the subfacets in order into facets (the foci are collateral classes of the second type, so that we can use the principle of inversion in deciding this);
6. choose a filing order between facets (their foci are collateral classes of the first type, so again we can use the principle of inversion).

At this state we shall have done enough to produce a scheme which will result in a preferred collocation and systematic order of documents or of records of documents; however, to make its use easier we must do two more things:

7. add a code to each class which will act as its address showing its filing position; this code is called 'notation';
8. produce an alphabetical index to the order classes, using their notation as a link. (46-47)

These basic stages are used in the creation of a number of KOSs. For example, La Barre (2010) proposes similar stages, based on the literature, in support of facet analysis. Likewise, the ANSI/NISO Z39.19-2005 (R2010) Guidelines for The Construction, Format, and Management of Monolingual Controlled Vocabularies proposes stages very much like these for the construction of thesauri.

Method

This paper adopts a topic modeling approach (a useful strategy in KO for assessing a domain (e.g., Joo, Choi, and Choi 2018)) to automate the first part of the facet analysis process. Topic modeling techniques identify high-level categories or classes (or "facets" (Vickery 2008)). These categories are, as Ranganathan termed them, "basic subjects" without "isolates" – that is to say, the categories are top-level concepts and do not have implicit in them "isolates" such as space or time (see Hjørland 2013, 547). This project therefore adopts an empirical approach, specifically, the deductive method as defined by ANSI/NISO (2010, 91), in the establishment of the topics. Several software packages are used, including MALLET for automatic topic clustering and Protégé for encoding the resulting ontology in OWL and WebVOWL for facet-relationship visualization.

The Topic Modeling Process

Like Buchanan (1980) and a number of others (e.g., ANSI/NISO 2010; Vickery 2008), we begin by identifying a representative sample of the scholarly literature in the RDM/DC field -- a corpus of abstracts of research articles containing the terms research data management or data curation. Scholarly articles in (peer-reviewed) journals (i.e., no professional journals) were collected on December 19, 2018 through the Library Literature and Information Science Full Text database. The term "research data management" (in quotations) retrieved 106 scholarly articles in which the search term appeared, as a phrase, somewhere in the article's metadata (e.g., in the title, abstract, or keywords)[1]; "data curation" (in quotations) yielded

[1] Given the fact that the term may be found anywhere in an article's metadata, we are not able to explicitly infer that any given article is about one subject or the other. Specifically, we built our corpus to investigate the sematic

111 scholarly articles which were combined for analysis. A publicly accessible version of these searches is available through our Zotero group[2]. The database search returned 217 articles in total, all of which include an abstract.

Because abstracts contain economically-written information regarding the aboutness of a research article, its results, and its conclusions (Cremmins 1996; NISO 2015), abstracts represent a coherent and complete style of academic writing devoted to a specific form of publication. Their prescribed word count presents a constraint by which all authors must adhere, avoiding huge variation in length or form. Abstracts therefore are concise, stylistically defined representations of the intellectual content of documents (NISO 2015; Hudon 2013; Cremmins 1996). Only the text of the abstracts was retained for analysis.

Topic modeling is next applied to the corpus for the identification of salient themes in the abstracts. Topic modeling works by presenting lists of words that appear together in statistically meaningful ways. Across all documents included in a corpus, terms are grouped together using a measure of lexical co-occurrence, rather than simply by random chance. Thus, terms which have a statistical likelihood of appearing together within a certain number of points or spaces are therefore semantically related (e.g. Lund and Burgess 1996; Blei 2012; Blei et al. 2010; Murakami et al. 2017). This is occasionally discussed as the "bag of words" (e.g. Burton 2013) model, which ignores grammatical structure in favor of statistical relationships between lexical items for each document. The computer thus seeks strongly-associated vocabulary to assign to discrete 'topics', and one can make comparisons across each collection of words representing each individual topic to pull out machine-generated similarity across a population of vocabulary. As an example, Jockers and Mimno (2012) use topic modeling to differentiate between 18th century novels authored by men and by women, based on sets of topics that appeared more frequently in their novels, making it possible to predict author gender for anonymously-authored works.

The MAchine Learning for LanguagE Toolkit (MALLET) implementation of Latent Dirichlet Allocation (LDA) was used for topic modelling. MALLET is a package designed for statistical natural language processing, document classification, clustering, topic modeling, and information extraction applications for text analysis. We used the following specifications: 50 topics, generated without stopwords, with 200 iterations printing 20 words per topic based on the topic proportion threshold of 0.05. We settled on these specifications after iterating on the ideal number of topics for our purposes. Any set of topics under 10 was too few and any set of topics above 75 was too large to make coherent interpretations of the resulting word-groupings. After much deliberation, we ultimately settled on 50 topics as our ideal number of topics: they were granular enough to show coherent meaning, with a minimum of topics fully representing so-called 'junk topics' of noisy data. We did not use a

relationship of terms within abstracts that include "research data management" and "data curation" in their provided metadata.

[2] https://www.zotero.org/groups/2246238/rdm_in_ko

customized stopword list, as we were interested in the potential overlap between potentially synonymous vocabulary[3].

Thus, thematically-driven vocabulary, such as research, library, and methodology occasionally appear in potentially polysemous ways throughout the individual topics. From the resulting 50 topics identified using LDA, topics were subsequently interpreted by the research team using an iterative process. Each member named and summarized as many topics as possible, identifying an overarching theme and a secondary theme (Chang et al. 2009). Validation was achieved by the four members of the research team investigating topics manually and reaching agreement on the general function and purpose of as many individual topics as possible[4]. The team adjudicated and cross-referenced the topics a total of six times over the course of several weeks. A full list of the results, including topics and names, is available (Froehlich, Hudson-Vitale, Adkins, and Moulaison Sandy 2019). This process influenced our approach to defining topics for subsequent analysis, as discussed below.

The Facet Analysis Process

Starting from the topics and the related aspects and contexts that were identified, we follow Buchanan (1980)'s process of grouping the topics into classes and identifying sub-classes. Part of the process of facet analysis involves paying attention to additional principles of logical division. "In general, any class will have many subclasses, but logical division is typically interested only in collections or families of subclasses that 'divide up' the original class i.e. the subclasses resulting from a step of division need to be disjoint and not have members in common" (Frické 2016, 540). In effect, this is slightly different from the grouped concepts that were created as part of the facet analysis process, wherein some terms appear in multiple clusters (as evidenced in table 1 below). We follow the standard conventions for logical division and ensure that, for our primary classes, they be mutually exclusive in regards to the basis for division. We acknowledge an element of bottom-up classification as the terms supplied in a "topic" yielded the naming of a "bucket", although that empirically-derived bucket is subsequently divided logically to produce a fully divided class.

Exclusive and exhaustive subclasses of the top-level classes (Frické 2016) were created. These subclasses exhibit the traits of the parent class while being exclusive and, to the best of our understanding, exhaustive. Theorists have differed in their approaches to the number and specifics of facets, but have agreed on the need for them to be both exclusive and exhaustive (Frické 2016; Spiteri 2000). The resulting ontology was marked up in Protégé using OWL and visualized in WebVOWL (See below).

[3] The standard stopword list included in MALLET which covers the top 1000 most frequent words in English, was used for this analysis. See https://www.wordfrequency.info/free.asp?s=y.

[4] Some topics we could not reach a conclusion on, which is to be expected: topic modelling presents the computer's understanding of lexical similarity, which does not always match up to human understanding of semantics or salient groupings.

Results

In this section, the results of the topic modeling, facet analysis, and visualization are presented.

RDM/DC Topic Modeling

Topics identified as meaningful, based on the words grouped through the topic modeling process, were established through consensus by the research team. Table 1 shows the results of the topic modeling relating to Training/Education, which are common topics. When possible, sub-facets that function as aspects of the topic (to supply context) were also identified to facilitate the facet analysis step. In the case of the Training/Education facet, aspects were related to the audience, the content, and the contexts. Words used as the basis for the topics were also provided; these also served to inform the facet analysis process.

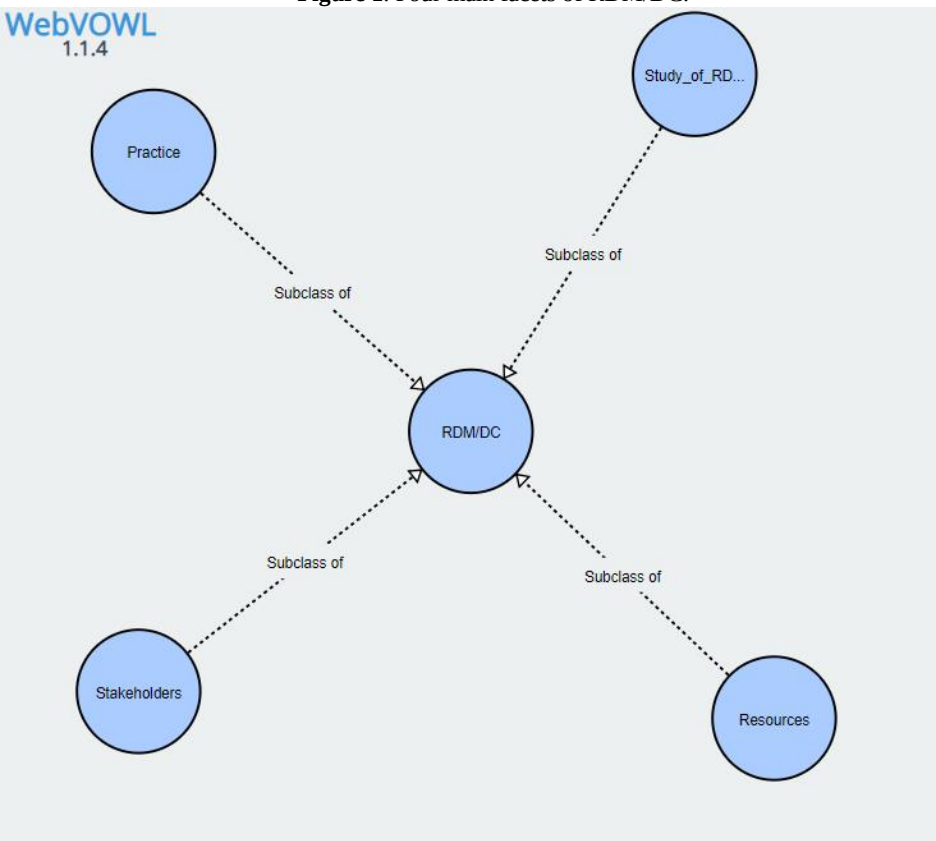
Table 1. RDM/DC topic modeling of "Training/Education" (Froehlich, Hudson-Vitale, Adkins, and Moulaison Sandy 2019).

Topic	Aspect/Context	Words
Training/Education	Audience: Librarians	data scientific requirements curation lis specialists order counts courses schools objectives ability selected evaluate curriculum usage status position details promote
Training/Education	Behavior and practices	data literacy related importance important relationship behaviour citation information vital review subjects nature close expected roles sr real workers literate
Training/Education	Cultural contexts	education results uk emerging studies trends higher developments current change capabilities bibliometrics major focused culture zealand limited landscape resourcing ireland
Training/Education	Audience: Librarians	librarians skills include future data planned small extension preparing teams africa semi train principles terms funders medical california exists endeavor
Training/Education	Audience: Someone else	management data training funding outreach resource comprehensive learning engage readiness mission required require reasons demand ongoing considered connection organizing preserved
Training/Education		library discusses program offer users instruction create established communications improve california continue discipline responded berkeley costs san carolina meeting convenience
Training/Education	Audience: Researchers	quality skills study specific understanding scientists needed found aspects genomics relevant roles perspective priorities constructs genome highly criteria undertaken taxonomy

RDM/DC Facet Analysis

Next, the topics were assessed by the research team, with four facets ultimately being identified. See figure 1 for a graphical representation of the OWL-encoded results in WebVOWL.

Figure 1. Four main facets of RDM/DC.[5]



The principle of logical division was applied in the creation of sub-facets. Based on the salient terms in the topic identified during the topic modeling as well as ones used the field itself, the following facets and sub-facets (see table 2) are presented, with example values for each. The presentation of the facet analysis mirrors Shiri (2014)'s presentation of big data.

[5] Available online at http://www.visualdataweb.de/webvowl/#iri=http://moulaison.net/NASKO/RDM-DC_facets.owl

Table 2. Facets and sub-facets of RDM/DC.

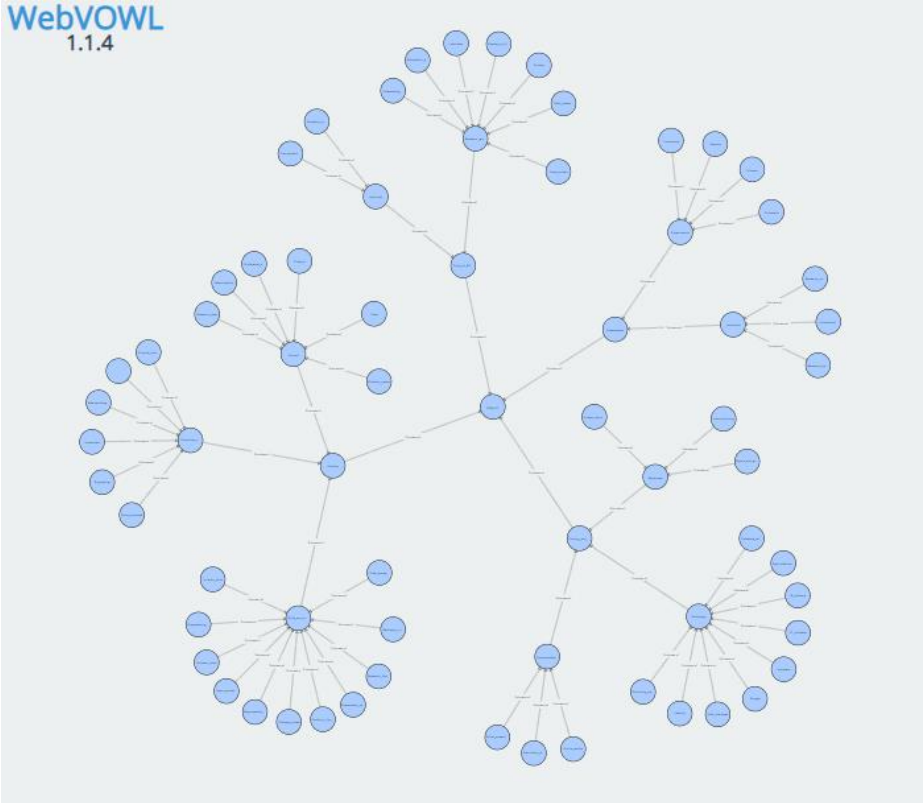
Facets	Sub-facets	Values
Practice	Skills and knowledge	-Data management skills
		-Curation skills
		-Preservation skills
		-Outreach skills
		-Software suites
		-Repository platforms
		-Programming languages
		-Requirements (funders, etc.)
		-Research lifecycle
		-Open access publishing
		-Trends in the field
	Content	-Research data
		-Scholarly papers
		-Conference proceedings
		-Books/reports
	Disciplines supported	-Sciences/medical field
		-Social sciences
		-Humanities
Stakeholders	Individuals	-Researcher producers
		-Research consumers
		-Librarians/information professionals
	Institutions	-Universities
		-Publishers
		-Funders
		-Vendors
Funding and support	Resources	-Internal funding
		-Human resources
		-Grants and external funding
Study of RDM/DC	Research methods	-Surveys
		-Case studies
		-Bibliometric analyses
		-Reviews of the literatures
	Instruction	-Training/workshops
		-Education programs

Visualization of the Resulting Ontology

To visualize the results of the facet analysis process, the classes and their subclasses are encoded in Protégé (https://protege.stanford.edu/) (Musen 2015) and then visualized through

the WebVOWL data visualization tool available through Visual Data Web (http://www.visualdataweb.org/). See figure 2.

Figure 2. Visualization of the RDM/DC field using WebVOWL[6]



Discussion

Ultimately, the facet analysis yields a structure that feasibly could be used in the creation of a faceted classification scheme to support organization and retrieval of research in the RDM/DC area. We will examine the results in light of the field, and explore the meaning and implications.

Using topic modeling to support the establishment of main topics and their aspects was useful in identifying ideas presented in the literature and exploring their relationships to the

[6] Retrieved from http://www.visualdataweb.de/webvowl/#iri=http://moulaison.net/NASKO/RDM-DC.owl. Set Filter "Degree of collapsing" to 0.

field and to each other. The empirical nature of the topic modeling process provided structure that otherwise would not have been present. The requirement that facet analysis adhere to principles of logical division as practiced in KO made for a more challenging and labor-intensive approach to carrying out the facet analysis than anticipated. The research team expected "plug-and-play" ability to move topics from the topic modeling analysis directly into the top-level classes of the facet analysis.

To adhere to the principles of exhaustivity and exclusiveness that are required for logical division, and not simply to "divide" the topics as would be fitting to a "classification" (see Hjørland 2013) required a good deal of unexpected grouping and arranging. Shiri (2014) mapped his high-level facets to Ranganthan's PMEST. This project chose specifically not to do that, in the interest of allowing topics/classes to emerge as guided by the empirical topic modeling process. We chose not to fit this emerging and specialized field to a scheme that came to be criticized and that was intended for faceting universal schemas.

Given this project's interest in letting the top-level classes emerge from the data, some concessions had to be made in the presentation of the facets. Obeying principles of logical division, for example, meant that not every aspect of the topic modeling was able to be represented in the facet analysis. Complex notions infused throughout the field, specifically "services" and "metadata," do not appear outright in the facet analysis this study produced. This is because principles of logical division require that facets and sub-facets be exclusive and exhaustive (Frické 2016), and there was no way to make these overarching topics fit in one and only one of the classes that emerged. We believe that there is not one area to accommodate services and metadata for the very reason that it is impossible to talk about RDM/DC without services and metadata being included in that discussion. This reinforces our initial assertion that RDM/DC is primarily about the behavior of users regarding data, and is, for this reason, a major finding of this study.

Limitations and Future Work

Initially, due to the data-intensive aspects of the RDM/DC field, the original intention was to seek points of convergence with Shiri (2014)'s facet analysis of big data as a way to evaluate the results of this study. Very little overlap was found between the two ontologies, potentially because of the differing approaches to facet analysis that were employed. To evaluate this work, usability testing (ANSI/NISO 2010, 95) will be employed to compare author-supplied keywords (i.e., warrant) included with the published articles to the classes and sub-classes generated through this research project.

Conclusion

This paper intends to examine the emerging field of research data management and its companion field, data curation, in a novel way. This paper also demonstrates the limitations of using topic modeling of journal literature for creating a facet analysis of a field of research. There are aspects of this emerging field that are yet uncovered in research. Through the examination of relationships between central notions in the field enabled by of topic modeling and facet analysis techniques, we have been able to show the major areas of

emphasis in this data-intensive field that, not coincidently, continues to be of interest to those in LIS, in scholarly communication, and perhaps increasingly, in KO and other fields that rely on data of all kinds.

Acknowledgement

This work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

References

ANSI/NISO. 2010. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Baltimore, MD: National Information Standards Organization. ISBN: 1-880124-65-3. Retrieved from https://groups.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf.

Blei, D.M. 2012. "Probabilistic Topic Models". *Communications of the ACM* 55(4):77–84.

Blei, D. M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Buchanan, B. 1980. Theory of Library Classification. London: New York: Shoe String Press, Inc.

Burton, M. 2013. "Topic Modeling for JDH." May 21, 2013. Retrieved from http://mcburton.net/blog/joy-of-tm/.

Chang, J., S. Gerrish, C.Wang, J.L. Boyd-Graber, and D.M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Information Processing Systems* 22. Retrieved from https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.

Cheti, A., and F. Paradisi. 2008. "Facet Analysis in the Development of a General Controlled Vocabulary." *Axiomathes* 18(2): 223–41. doi:10.1007/s10516-008-9033-4.

Cremmins, E. 1996. The Art of Abstracting. 2nd edition. Arlington, VA: Information Resources Press.

Darch, P.T., and E.J.M. Knox. 2017. "Ethical perspectives on data and software sharing in the sciences: A research agenda." *Library & Information Science Research* 39(4): 295-302.

Denzin, N.K. and Y.S. Lincoln. 1994. Handbook of Qualitative Research. Thousand Oaks, CA: Sage.

Frické, M. 2016. "Logical Division." *Knowledge Organization* 43(7): 539–49. doi:10.5771/0943-7444-2016-7-539.

Froehlich, H., C. Hudson-Vitale, H.M. Sandy, and D. Adkins. Data Set for Topic Modeling and Facet Analysis in an Emerging Domain: Research Data Management and Data Curation. https://doi.org/10.26207/hrz9-9963.

Hjørland, B. 2013. "Facet analysis: the logical approach to knowledge organization." *Information Processing & Management* 49: 545-57. doi:10.1016/j.ipm.2012.10.001.

Hudon, M. 2013. Analyse et représentation documentaires: introduction à l'indexation, à la classification et à la condensation des documents. Québec: Presses de l'Université du Québec.

Jockers, M.L. and D. Mimno. 2013. "Significant Themes in 19th-Century Literature." Poetics 41(6): 750-769.

Johnston, L.R., ed. 2017. Curating Research Data: Volume One: Practical Strategies for Your Digital Repository. Chicago, Illinois: ACRL. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf.

Joo, S., I. Choi, and N. Choi. 2018. “Topic Analysis of the Research Domain in Knowledge Organization: A Latent Dirichlet Allocation Approach.” *Knowledge Organization* 45(2): 170-183. doi:10.5771/0943-7444-2018-2-170.

La Barre, K. 2010. “Facet Analysis.” *Annual Review of Information Science and Technology*, 44: 243-84.

Lund, K. and C. Burgess. 1996. "Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence." Behavior Research Methods, *Instruments, & Computers* 28(2): 203-208.

Murakami, A., P. Thompson, S. Hunston, and D. Vajn. 2017. “‘What Is This Corpus about?’: Using Topic Modelling to Explore a Specialised Corpus.” *Corpora* 12(2): 243–77. doi:10.3366/cor.2017.0118.

Musen, M. A. June 2015. The Protégé Project: A Look Back and a Look Forward. AI Matters. *Association of Computing Machinery Specific Interest Group in Artificial Intelligence* 1(4). doi:10.1145/2557001.25757003.

NISO (National Information Standards Organization). (2015). Guidelines for Abstracts (No. ANSI/NISO Z39.14-1997 (R2015)). Baltimore, MD. Retrieved from https://groups.niso.org/apps/group_public/download.php/14601/Z39-14-1997_r2015.pdf.

Owens, L.A., and P.A. Cochrane. 2004. “Thesaurus Evaluation.” *Cataloging & Classification Quarterly* 37(3–4): 87–102. doi:10.1300/J104v37n03_07.

Packalén, S. and P. Henttonen. 2016. “Ambiguous Labels: Facet Analysis of Class Names in Finnish Public-Sector Functional Classification Systems.” *Knowledge Organization* 43(7): 490–501. doi:10.5771/0943-7444-2016-7-490.

Priss, U. 2008. “Facet-like Structures in Computer Science.” Axiomathes 18(2): 243–55. doi:10.1007/s10516-007-9023-y.

Shiri, A. 2014. “Making Sense of Big Data: A Facet Analysis Approach.” *Knowledge Organization* 41(5): 357–68. doi:10.5771/0943-7444-2014-5-357.

Slavic, A., and S. Davies. 2017. “Facet Analysis in UDC: Questions of Structure, Functionality and Data Formality.” Knowledge Organization 44(6): 425–35. doi:10.5771/0943-7444-2017-6-425.

Slavic, A. 2008. “Faceted Classification: Management and Use.” *Axiomathes* 18(2): 257–71. doi:10.1007/s10516-007-9030-z.

Spiteri, L. 2000. “The Essential Elements of Faceted Thesauri.” *Cataloging & Classification Quarterly* 28(4): 31–52. doi:10.1300/J104v28n04_05.

Spiteri, L. 1998. “A Simplified Model for Facet Analysis: Ranganathan 101.” *Canadian Journal of Information and Library Science* 23(1/2): 1–30.

Steinhart, G. 2013. “Partnerships between Institutional Repositories, Domain Repositories and Publishers.” *Bulletin of the American Society for Information Science and Technology* 39(6): 19–22. doi:10.1002/bult.2013.1720390608.

Tenopir, C., B. Birch, and S. Allard. 2012. Academic Libraries and Research Data Services: Current Practices and Plans for the Future: An ACRL White Paper. Chicago: Association of College and Research Libraries. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf.

Tonkin, E., H. Pfeiffer, and A. Hewson. 2010. An evidence-based approach to collaborative ontology development. In: Workshop on Matching and Meaning 2010, 2010-03-31 - 2010-04-01, Leicester. Retrieved from https://core.ac.uk/download/pdf/2805296.pdf.

Vickery, B. 2008. “Faceted Classification for the Web.” *Axiomathes* 18(2): 145–60. doi:10.1007/s10516-007-9025-9

Weil, Ben H. 1970. "Standards for Writing Abstracts." *Journal of the American Society for Information Science* 21(5): 351-357.

Whyte, A. and J. Tedds. 2011. “Making the Case for Research Data Management.” DCC Briefing Papers. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm