IJDC | *Peer-Reviewed Paper*

A Report of Data-Intensive Capability, Institutional Support and Data Management Practices in Social Sciences

Wei Jeng
School of Information Sciences
University of Pittsburgh

Liz Lyon
School of Information Sciences
University of Pittsburgh

Abstract

We report on a case study which examines the social science community's capability and institutional support for data management. Fourteen researchers were invited for an in-depth qualitative survey between June 2014 and October 2015. We modify and adopt the Community Capability Model Framework (CCMF) profile tool to ask these scholars to self-assess their current data practices and whether their academic environment provides enough supportive infrastructure for data related activities. The exemplar disciplines in this report include anthropology, political sciences, and library and information science.

Our findings deepen our understanding of social disciplines and identify capabilities that are well developed and those that are poorly developed. The participants reported that their institutions have made relatively slow progress on economic supports and data science training courses, but acknowledged that they are well informed and trained for participants' privacy protection. The result confirms a prior observation from previous literature that social scientists are concerned with ethical perspectives but lack technical training and support. The results also demonstrate intra- and inter-disciplinary commonalities and differences in researcher perceptions of data-intensive capability, and highlight potential opportunities for the development and delivery of new and impactful research data management support services to social sciences researchers and faculty.

Received 19 October 2015 ~ *Accepted* 24 February 2016

Correspondence should be addressed to Liz Lyon, 135 N Bellefield Ave, Pittsburgh, PA 15260, USA. Email: elyon@pitt.edu

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

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

During the last decade, the practice of research has become increasingly focused on the collection, analysis and visualization of growing volumes of data. This 'data deluge' was described by Hey and Trefethen (2003) who highlighted the importance of data curation and research data management. This data landscape is a complex space with disciplinary differences in behaviours and many community stakeholders involved. Their various roles and responsibilities were described by Lyon (2007), with the data stakeholders including publishers, funding agencies, universities, libraries and data centers, as well as the researchers themselves. The recognition of a new era of data-intensive science was encapsulated in the 'Fourth Paradigm' by Hey, Tansley and Tolle (2009), which illustrated how data-intensive science was transforming research practice in certain domains, such as ocean science and healthcare.

Reviewing a very selective data landscape timeline from 2003 to date, demonstrates that 'data' has been the subject of many influential national strategy reports (Atkins, 2003; ANDS, 2012; Research Data Canada, 2008) and recommendations from leading professional scientific societies (The Royal Society of London). There have been newly-established data journals (e.g. GigaScience[1], GeoScience Data Journal[2]), innovative infrastructure platforms (Dryad[3], figshare[4], DataONE[5]), new centers of expertise (UK Digital Curation Centre[6] launched in 2004), major international conferences (International Digital Curation Conference[7] annually since 2005), novel educational programs (e.g. MSc Data Science at UC Berkeley) and the establishment of a high-profile community-based international organization (Research Data Alliance[8]).

The Community Capability Model Framework (CCMF) shown in Figure 1 was developed by UKOLN Informatics, University of Bath, in consultation with the eScience community to provide a foundation for determining the capability of a given community for data-intensive research (Lyon et al., 2012b). The CCMF used the following definition of a Capability Model as a foundation for its development:

> 'A model for determining whether, how easily, and how well the agent in question could, in theory and in practice, accomplish a given task' (Lyon et al., 2012a).

The concept is often associated with the notion of 'maturity' and a number of existing capability and/or maturity models informed the development of the CCMF. These are briefly reviewed here, with fuller descriptions in the CCMF White Paper and in Lyon et al. (2012a).

1 GigaScience: http://gigascience.biomedcentral.com
2 GeoScience Data Journal: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-GDJ3.html
3 Dryad: http://www.datadryad.org
4 Figshare: https://figshare.com
5 DataONE: https://www.dataone.org
6 Digital Curation Centre: http://dcc.ac.uk
7 International Digital Curation Conference: http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc
8 Research Data Alliance: https://rd-alliance.org

 doi:10.2218/ijdc.v11i1.398

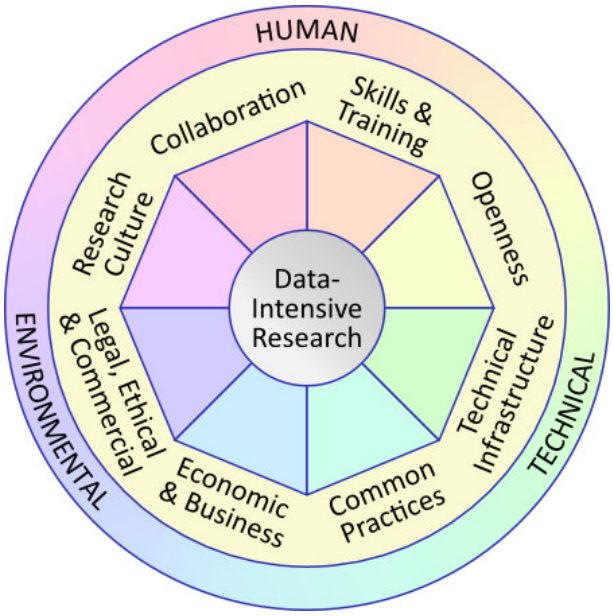

Figure 1. Community Capability Model Framework.

The development of this instrument has closely engaged the data curation communities, especially IDCC community. In IDCC 2013, one of the authors held a workshop for the debut of the Community Capability Model Framework (CCMF), whose capability factors and applicability were then introduced. Later in 2014 the same team held a half-day workshop that focused on the application of the Capability Profile Tool in Agronomy and in Environment Sciences. The Community Capability Model Interest Group (IG) has held several IG meetings at RDA Plenary meetings since 2013.

This study applies the Community Capability Model Profile Tool developed from the Community Capability Model Framework (CCMF), which is described in a later section. We aim to address the following research questions:

2. What are the current statuses of data-intensive practices, institutional support, and individual data management practices within the social sciences?

3. How do these capabilities compare and contrast within the specific disciplines studied?

In summary, this article makes three contributions. First, this article presents an existing data-profiling tool (i.e., CCMF) tailored for social science domains. Second, findings related to the research questions are expected to not only reveal the social science community's capability and infrastructure for supporting data related activities, but also to further our understanding of the discipline as a whole. Third, in addition to these research questions, we also discuss the capability implications and community impacts for a range of data stakeholders, and in particular for education programs, curation infrastructure and institutional support services.

Data Practices in Social Sciences

Although data-intensive disciplines are typically associated with the STEM domains, it has been widely acknowledged that social sciences also need great access to data and transparency (Guest, Namey, and Mitchell, 2012; Tolle, Tansley, and Hey, 2011).

Tenopir et al. (2011) conducted a national survey that recruited 1,329 scientists, including 204 social science scholars. The survey found that social science researchers are "less likely to make their data electronically available to others" when compared to STEM scholars, and only 47 out of 204 (23%) agreed or somewhat agreed that their data could be easily accessed by others. On the other hand, the percentage agreement from scholars in atmospheric science and biology were nearly two times higher or more (38.5% and 49.0%, respectively). In contrast, 162 out of the 204 social scientist participants (79.4%) in the survey agreed or somewhat agreed that they had concerns about data being used in ways other than intended.

The Data Curation Profiles (DCP) project[9] unveiled common standards for social science disciplines, such as preferred file exchange formats, file sizes, and embargo times (where data are not published or shared until a certain date or some certain conditions have been met) for researchers (Cragin, Palmer, Carlson, and Witt, 2010). So far, the DCP website has published five volumes since 2009 and includes seven profiles related to social science and humanities. Researchers Lage, Losoff, and Maness (2011) in University Libraries at University of Colorado-Boulder have also adopted the DCP tool to examine the institution's scientific data curation activities.

Since most systematic and comprehensive studies are based on STEM researchers, less is known about whether or not the social science community manage their data effectively and efficiently. Therefore, there is an imperative need to bridge the gaps between research into STEM disciplines and social science disciplines in terms of data intensive research.

Assessing Data Capability

There are several endeavours within the professional working groups and discipline communities that address the assessment of data capability and practice. The Australian National Data Service (ANDS) has applied a Community Maturity Model to research data management primarily within institutions (ANDS, 2011). Once again, five levels of maturity are applied to four process areas: Institutional Policies and Procedures, IT Infrastructure, Support Services, and Managing Metadata.

A maturity model for research data management in research projects was developed by Crowston and Qin (2011) using five levels which may be summarized as ad hoc, reactive, proactive, standards maintained, improvements made proactively. They also identified four process areas: 1) data acquisition, 2) processing and quality assurance, 3) data description, and 4) representation, data dissemination, repository services/preservation.

The Cornell Maturity Model (Kenney and McGovern, 2003) was developed to describe a higher education institution response to digital preservation and uses a five point approach: Acknowledge, Act, Consolidate, Institutionalize, Externalize. Three dimensions were added: Organization, Technology, and Resources. These became known as the Three-Legged Stool Model by Kenney and McGovern (as cited in McGovern, 2007). This approach was further modified and developed by others e.g. AIDA Project Toolkit[10] and the CARDIO tool[11] from the UK Digital Curation Centre. Each of these initiatives extended the model to include a scorecard tool, based either in Word and Excel (AIDA) or a Web-based tool (CARDIO).

9 Data Curation Profiles: http://datacurationprofiles.org
10 AIDA Project Toolkit: http://aida.da.ulcc.ac.uk/
11 CARDIO: http://cardio.dcc.ac.uk/

 doi:10.2218/ijdc.v11i1.398

The DMVitals toolkit for data management developed at the University of Virginia also uses an Excel-based scorecard with five color-coded levels of maturity, followed with recommendations and action statements (Sallans and Lake, 2014).

In summary, these capability and maturity models display a number of commonalities and differences.

Methodology

Instrument: Community Capability Model Framework (CCMF)

The Community Capability Model Framework (CCMF) in this study provides a foundation for determining the capability of a given community for data-intensive research (Lyon et al., 2012a). The term 'community' was interpreted as the groups of people, i.e. faculty or academics or researchers, defined by a particular discipline or sub-discipline. Further articulation of this aspect is given in Lyon et al. (2012b) and the CCMF white paper (Lyon et al., 2012a).

The CCMF aims to encompass a broad range of aspects of 'data-intensive research capability' as indicated by the categories. Acting as an instrument, CCMF covered eight factors contributing to data management capability (Table 1), which were assessed in order to gain an understanding of data infrastructure issues in social science disciplines.

Table 1. Eight dimensions of the CCMF instrument.

#	Dimension	Description
1	Collaboration	Researchers describe their collaborative cultures between sectors, between themselves and their colleagues, and if their studies engaged the public.
2	Skill and training	Researchers are asked to assess their own skill sets and evaluate their institutional training programs related to data curation.
3	Openness	Researchers are asked to answer the extent of openness regarding their research, methods, data, and research outcomes.
4	Technical infrastructure	Researchers are asked to evaluate their discipline-wide support in terms of data storage, computing, processing, discovering, and accessing.
5	(Data) Common practices	Researchers capture details about their data characteristics and how they describe their data.
6	Economic and business models,	Researchers are asked to answer questions related to funding, in terms of its scale, location, and coverage.
7	Legal, ethical and commercial	Researchers answer regulatory framework, norms, and ethical responsibilities related questions.
8	Research culture	Researchers are asked to answer questions related to reward models and validation framework related to their research

An Excel-based CCM profile tool was developed from the model and made available from the CCM website[12]. The tool has a worksheet for each capability factor with more specific sub-components facilitating a deep analysis of each dimension. The tool can be used as a self-assessment tool by a researcher or can be used in a mediated

12 CCM website: http://communitymodel.sharepoint.com

mode. Five capability levels are used to describe the level of ability or activity within a particular dimension: 1) Nominal activity, 2) Pockets of Activity, 3) Moderate Activity, 4) Widespread Activity, and 5) Complete Engagement. The score for a particular capability factor gives an indication of the perceived position of that community from the viewpoint of the researcher.

Instrument Modification

The instrument of this study was an adapted version of the CCMF Toolkit with discipline-tailored modifications that were designed primarily to enhance comprehension. This was achieved by adding social science friendly descriptions, exemplars, or tools, and providing explanations of technical terminologies.

There were 37 modifications in total; some sample modifications are provided in Table 2. For each survey profile, the participant was asked to work on 16 open-ended questions about the nature of their research data and data-sharing behaviours. They were also asked to complete 55 closed-ended questions based on the CCMF Toolkit. For each closed-ended question, the participants could freely add comments or suggest preferred exemplars that the instrument did not list.

Table 2. Modification examples to CCMF instrument.

Modification Categories	Examples of Original Versions	Examples of Modified Versions
Adding discipline-tailored exemplars and tools	4.2 Tool support for data capture and collection	4.2 Tool support for data capture and collection (e.g. Screencasting tools, digital audio recorder, Web content scripters, Qualtrics, SurveyMonkey)
	5.5 Standard vocabularies, semantics, ontologies	5.5 Standard vocabularies, semantics, ontologies (e.g. LCSH, MeSH)
Providing explanations of technical terminologies	2.11 Data referencing and citation e.g. DataCite DOIs	2.11 Data referencing and data citation e.g. it uniquely identifies an object stored in a repository, such as DataCite DOIs)
	2.12 Data metrics and impact e.g. impact factors, altmetrics	2.12 The concepts of measuring scholarly impacts on data e.g. impact factors of research datasets, altmetrics of datasets such as the number of downloads
Providing discipline-tailored descriptions in social sciences	3.4 Openness of methodologies/workflows (e.g short 'how-tos', scripts for processing, programs for conversions)	3.4 Openness of methodologies/workflows (e.g. steps for preparing an interview or a focus group, how to run different statistical models on a software program)

 doi:10.2218/ijdc.v11i1.398

Data Collection and Limitations

This study uses a convenience sampling method for data collection, recruiting researchers who are conveniently available to participate in this study. The recruitment procedure further ensures that participants represent different domains in social sciences.

Targeted participants include doctoral students, post-doctoral researchers, and faculty members from the Departments of Anthropology, Political Sciences, and the Library and Information Science (LIS) Program at the University of Pittsburgh, USA. A recruitment message was posted on two major social media platforms: Craigslist and Facebook. We asked the potential participants to pass along the recruitment information to others who may be interested in the research study.

Four participants were interviewed (for open-ended questions) and mediated (for closed-ended ones) in July and August 2014. Each interview and mediation session was two to three hours long in total, allowing for a 'deep dive' into the scholars' data practices and capability levels. Each participant was compensated with $20-25 gift cards (USD) for their completion time. Besides the interviews and mediations, the CCMF tool was emailed to a cohort of 15 participants beginning in August 2014, and ten were completed and returned as of October 2015, under a self-assessment approach. For participants who conducted the survey using a self-assessment approach, the announced length of time was 60 minutes. Each participant was compensated with a $15 gift card (USD) for their completion of the survey.

Although it might be effective to use a convenience sampling method at this exploration stage, there are also several short-comings: there might be a selection bias because all the participants are affiliated with the University of Pittsburgh, and they are early-career researchers.

Results

As shown in Table 3, the social sciences cohort included a total of 14 participants sub-divided into PhD students (N=7), post-doctoral researchers (N=4) and Assistant Professors (N=3). The cohort included the disciplines of: Anthropology (N=4), Political Sciences (N=4), and Library and Information Science (N=6).

Social Scientists' Data Practices

On average, participants used 6.8 words, or 2.7 phases to describe their research data. A wide range of data types are reported in Table 4, with a higher proportion of observation field notes (N=8), interview records (N=8), and survey data (N=4). P01, an anthropological researcher, stated that he had been trained to collect data using a holistic approach: he usually deals with complex qualitative data, such as field notes, surveys, interviews transcriptions (categorized as interview records in Table 4), maps, and material samples, such as tickets or leaf samples. P03, a PhD student whose research interest is geography information systems (GIS) and accessibility, stated that her data usually has multiple attributes:

> '...plus space and time. Some attributes are quantitative and some qualitative. There are often classification codes that are needed to understand some attributes' (P03).

However, political science scholars in this study deal with more quantitative data. For example, P06 and P08 stated that they use government statistics and datasets for large-N analysis.

Table 3. Study participants (A = interviewed and mediated, B = self-assessed).

#	Approach	Position	Discipline	Self-Identified Research Topic
P01	A	Post-doctoral researcher	Anthropology	Cultural anthropology
P02	A	PhD student	Library and Information Science	Music metadata
P03	B	PhD student	Library and Information Science	Geospatial information systems (GIS) and accessibility
P04	B	PhD student	Library and Information Science	Information retrieval
P05	A	PhD student	Anthropology	Cultural anthropology, Legal Anthropology (child adoption)
P06	A	Assistant professor	Political Science	Comparative politics
P07	B	Post-doctoral researcher	Political Science	Area studies (South Asia)
P08	B	PhD student	Political Science	Comparative politics, political methodology
P09	B	PhD student	Anthropology	Archaeology
P10	B	Visiting scholar (assistant professor)	Library and Information Science	Public library management
P11	B	Post-doctoral researcher	Anthropology	Medical anthropology
P12	B	Assistant professor	Library and Information Science	Public library management
P13	B	Post-doctoral researcher	Library and Information Science	Information seeking behaviours
P14	B	PhD student	Political Science	International relations

Participants were also asked about the uniqueness of their data. Nine of 14 participants stated that their data could be fully or partially recreated and is therefore not unique. P05, a senior PhD student who studies child adoption culture in Federated States of Micronesia, specified that in regards to partial recreation:

> '[In my study] legal records can be always retrieved, but I am not sure about the interview (data)' (P05).

When the participants were asked to estimate their typical data volumes for one research project, the responses ranged from less than 25MBs (N=2), 200MBs (N=1), 1-10GBs (N=4), to more than 10GB (N=5). Three participants out of the five who claimed to produce more than 10GB of data per project (P01, P02, P05) specified that their data

 doi:10.2218/ijdc.v11i1.398

Table 4. Types of research data in social sciences.

Types of Data	Mentioned by N Participants
The field notes	8
Interview records	8
Survey results (questionnaire)	4
Experimental log/records	2
Historical documents	2
Maps	2
Spatial data	2
Relationship data (e.g. triples of metadata)	2
Government statistics	2
Participant diary	1
Data from focus group	1
Material samples	1
Video or screen-casting	1
Archaeological excavation survey	1
Observational data (did not specify)	1

include video, audio, photos, and screencast videos. Two participants answered 'it depends.' For example, P11 stated:

> 'In terms of computer space, very little. In terms of documents (audio files, video files, transcripts, diaries, surveys, etc.) and researcher-produced data (journal, analytic memos, code books, observation and field notes, etc.), it can be significant, especially if analysing and coding by hand' (P11).

Figure 2 summarizes the responses that we collected in participants' open-ended items. Based on participants' responses, we found that social scientists do have a need to reuse others' data, especially data from institutional repositories or discipline repositories (N=10, 71%). However, on the contrary, only three out of 14 participants (P05, P09, and P10) had deposited their data in repositories. It is worth noting that although only half of the participants had received requests for sharing materials or data, all participants indicated that they are willing to share upon request.

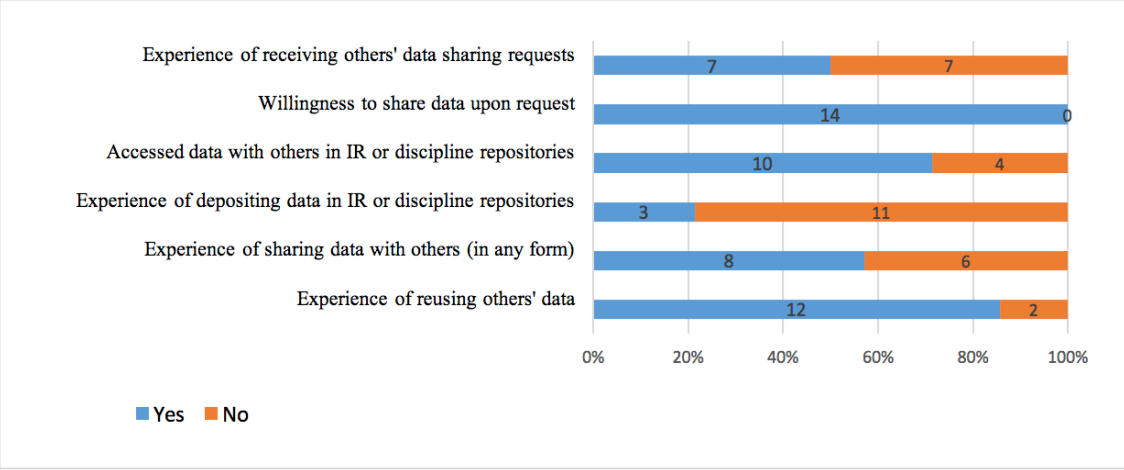

Figure 2. Participants' data sharing practices.

Social Scientists' Data Capability

Figure 3 presents a summary of data capability (shown in medians) in social science disciplines across all capability dimensions. The radar plot demonstrates some inter-disciplinary synergies and differences in data-intensive capability across this sub-section of the social sciences. For example, the dimension of data common practices and technical infrastructure has been highly rated by LIS researchers, even though they work in different sub-disciplines, whereas the anthropologists seem to value the dimension of collaboration more. Political science scholars rank Legal and Ethical and Openness as the highest development, while assigning relatively low scores to other dimensions.

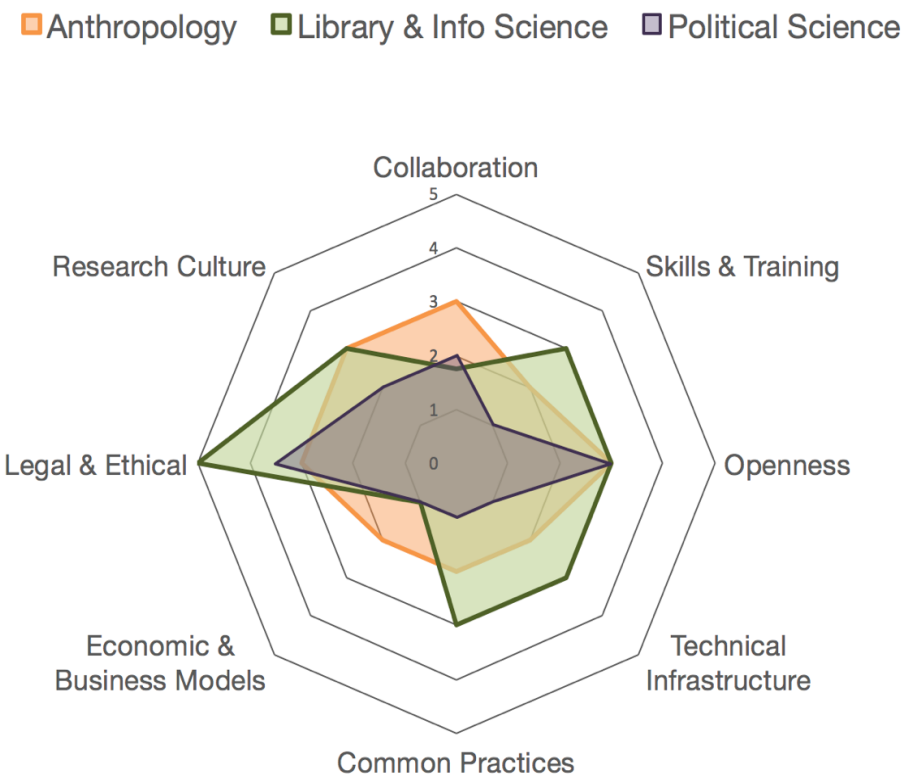

Figure 3. Capability summary for social sciences disciplines (by median).

For the disciplines' differences, we found that anthropology scholars' ratings were relatively evenly distributed across all dimensions. Political science scholars ranked Legal and Ethical and Openness as highest in development, whereas they assigned relatively low scores to other dimensions. LIS scholars gave better scores to Legal and Ethical but in general assigned higher scores to Skill and Training, Technical Infrastructure, and Common Practices than other two disciplines.

By ranking the median, we filtered out the top ten most-developed activities for each discipline and then identified the top activities shared among two or more disciplines. All items that were rated 3.5 or above are illustrated as a Venn diagram in Figure 4, which provides a better visualization for overlapping items.

The most developed activity across three disciplines is openness of published literature. While the legal and ethical responsibilities aspects had been rated highest by both LIS and political science researchers, in anthropology there is a mix of economic, business and collaboration concerns. On the other hand, common practices related to

 doi:10.2218/ijdc.v11i1.398

data curation and analysis (i.e. data collection, visualization, and process workflows) are ranked higher in LIS in comparison to the other two fields, whereas political science has more unique items related to their openness and reuse culture in their top ten list.

Using the same approach, 18 items rated in the bottom ten were visualized in a Venn diagram (Figure 5). The median of all items in Figure 5 are rated one (nominal activity). Data identifier, scale of infrastructure, and the use of a research discovery/networking system (e.g. CRIS) are the common items across three disciplines. The economic and business models capability dimension is most commonly perceived as weakly developed by LIS and political science.

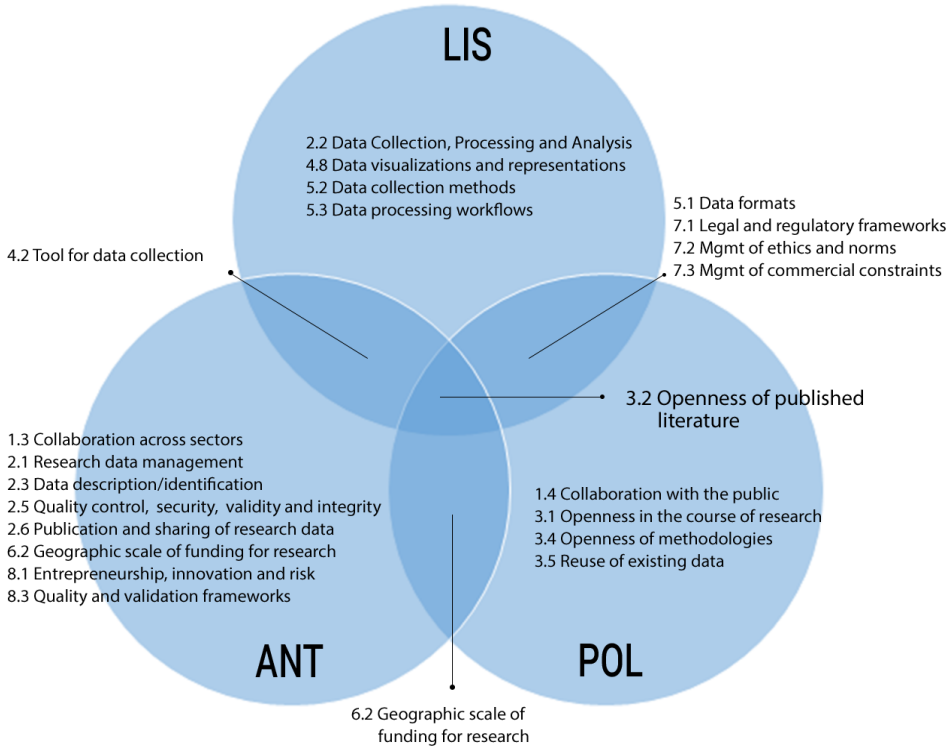

Figure 4. Most developed activities by discipline.

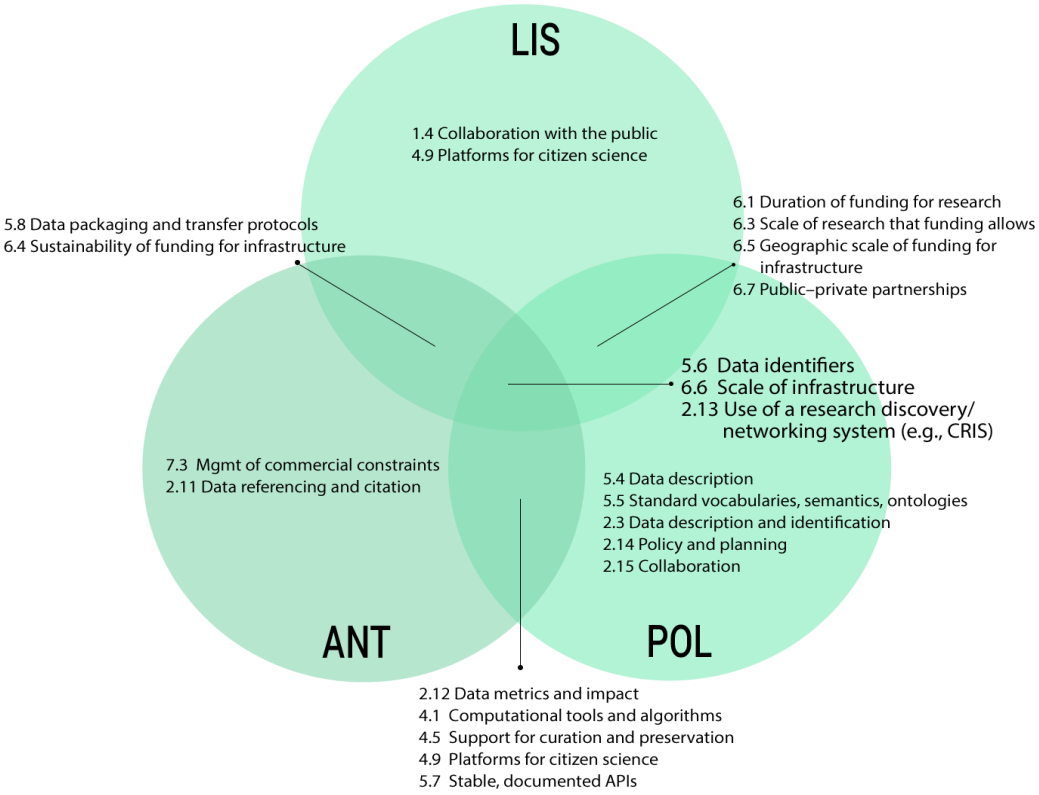

Figure 5. Least developed activities by discipline.

It is worth mentioning that one of the participants also shared their 'know-more moment' with us. P11 stated:

> 'The [CCMF] survey made me realize even more that we have so many technological opportunities that we aren't using and taking advantage of – especially in terms of data sharing and collaboration!' (P11).

Discussion

Insights on the Instrument

The results from this study in selected disciplines within the social sciences demonstrate the effectiveness of the Community Capability Model tool in identifying and measuring capability for data-intensive research. The breadth and depth of the CCM tool dimensions covering technical, human and environmental aspects of 'data-intensive' produces a rich and informative picture of the perceptions and assessments of the

 doi:10.2218/ijdc.v11i1.398

scholars within these disciplines. The CCM tool has proved to be easy to use by practicing researchers, but its effectiveness was enhanced by the customization of the vocabulary and illustrative exemplars to match the disciplinary expectations and context familiarity of the participants.

However, one disadvantage of this tool is that the overall process is time consuming, therefore making it difficult to recruit participants. The combination of open-ended questions plus closed ones with commentary provides essential opportunities for the researcher to extend and explain their opinions in terms of the capability levels selected in the various dimensions. Whilst there is scope to further extend the open-ended section, there is a delicate balance between receiving a successfully-completed and rich response to the instrument, and the respondent not replying because the instrument is perceived to be too complex or to take too much time to complete. We are very aware that researcher time is precious, however the CCM methodology appears to achieve the right balance and to produce high-quality information.

The diverse composition of the cohort participants (from PhD students to Assistant Professors) and their varied sub-disciplines (from music metadata to geospatial data), has resulted in evidence of them handling a broad range of data types from interview records to government statistics to video. This heterogeneity creates a particular challenge to libraries and data centres in providing research data management support services.

Insights on the Social Scientists' Data Practices

The majority of participants stated they were not creating or collecting unique data, such as environmental observations, however survey and interview records may be considered unique in one sense; whilst they can be repeated, the participants' views and answers may change over time. The perception of the value of specific datasets may help to determine later decisions on selection, appraisal and ingest into a data repository. The participants dealing with video described data volumes of >10GB per project. Whilst these are not huge data volumes compared to certain disciplines, such as astronomy or high energy physics, there are storage and management implications when the total numbers of scholars working with this type of data are considered together across an institution. The data volume results listed in Table 3 suggest that social science researchers are positioned in the long tail of research data; this is an Interest Group topic within the Research Data Alliance.

Reviewing the social sciences together (before examining the three disciplines studied), a very small group of activities (N=14) were considered to be well-developed practices in data-intensive research. However, the participants perceived the Legal, Ethical and Commercial dimension to be a strength and this may reflect the robust Institutional Review Board processes in place, which provide a well-established foundation for good data practices when working with human subjects. The social scientists also ranked aspects of openness as highly developed; this may reflect well-established community behaviours in publishing in open access archives, such as RePEc.[13] In contrast there were far more items identified by the social scientists as not well-developed activities (nominal or pockets), with a clear gap in Economic and Business Models capability. This dimension explores aspects of data-intensive research funding (duration, geographic scale, size of investment, sustainability, partnerships). The pilot results indicate a strongly-perceived lack of investment in this domain. The

13 RePEc: http://repec.org

second dimension which is perceived as weak is that of Skills and Training for data-intensive research. This dimension includes training for research data management tools, data management plans, data description, data publication, data citation and data metrics. This result indicates a distinct opportunity for support services from data stakeholders, such as libraries and IT services, to provide timely advocacy, guidance materials and training programs for research data management directly to researchers, perhaps through graduate schools, doctoral training centres or as elements in the promotion of digital scholarship initiatives.

Comparing the disciplines of Anthropology, Political Sciences, and Library and Information Science (LIS), there are limited commonalities between the top ten most developed (high capability) items in the three areas, beyond agreement on Legal and Ethical aspects, which are considered a strength. The varied mix of activities listed in Figures 4 and 5 is perhaps a reflection of the distinctiveness of the disciplines investigated; whilst they can be collectively described as social sciences, the perceptions of researchers reflect their diversity and independence. Similarly, there is a rich mix of activities viewed as least developed (low capability), but with clear agreement on the poor research funding situation (see Figure 5). The detail for each discipline again highlights opportunities for new and impactful professional services focused on 'data' to be delivered to researchers. Additional support for data curation and preservation, for data standards and for data management planning tools are identified by participants.

Conclusion

This study confirms that the economic and business model, skill and training activities, and technical infrastructure were the least-developed activities for social science scholars. The diverse composition of the cohort participants (from PhD students to Assistant Professors) and their varied sub-disciplines (from music metadata to geospatial data), has resulted in evidence of them handling a broad range of data types from interview records to government statistics to video. This heterogeneity creates a particular challenge to libraries and data centres in providing research data management support services. The results also suggest that social scientists have developed more maturely in terms of legal and ethical aspects, and have positive attitudes on data openness and sharing. In future work, it will be worthwhile to deepen the understanding of the disciplines' similarity and deviation in data practices and capabilities. In conclusion, the results from this CCM Profile study suggest that there is much work to be done to help to equip researchers in the social sciences with the data-intensive capability and skills need for 21st century science.

References

Atkins, D.E. (2003). Revolutionizing science and engineering through cyber-infrastructure. Report from the NSF Blue-Ribbon Taskforce. Retrieved from https://www.nsf.gov/cise/sci/reports/atkins.pdf

Australian National Data Service. (2012). Towards the Australian data commons. ANDS Technical Working Group. Retrieved from http://www.ands.org.au/__data/assets/pdf_file/0009/386982/ardc_eif_annual_report_2011-12.pdf

Australian National Data Service. (2011). Research data management framework: Capability maturity guide. ANDS Guides. Retrieved from http://ands.org.au/guides/dmframework/dmf-capability-maturity-guide.html

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

Crowston, K., & Qin, J. (2011). A capability maturity model for scientific data management: Evidence from the literature. *Proceedings of the American Society for Information Science and Technology, 48*(1), 1-9. doi:10.1002/meet.2011.14504801036

Guest, G., Namey, E.E., & Mitchell, M.L. (2012). Collecting qualitative data: A field manual for applied research. Sage.

Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G.C. Fox, and T. Hey, (Eds.) *Grid computing: Making the global infrastructure a reality.* Wiley, New York.

Hey, A.J.G., Tansley, S., & Tolle, K.M. (Eds.). (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, Washington: Microsoft Research.

Kenney, A.R. & McGovern, N.Y. (2003). The five organizational stages of digital preservation. In P. Hodges, M. Sandler, M. Bonn, and J. P. Wilkin, (Eds) *Digital Libraries: A Vision for the 21st Century*. University of Michigan Scholarly Publishing Office, Ann Arbor, MI. Retrieved from http://hdl.handle.net/2027/spo.bbv9812.0001.001

McGovern, N.Y. (2007). A digital decade: Where have we been and where are we going in digital preservation? *RLG DigiNews, 11*(1).

Lage, K., Losoff, B., & Maness, J. (2011). Receptivity to library involvement in scientific data curation: A case study at the University of Colorado Boulder. *portal: Libraries and the Academy, 11*(4), 915-937. doi:10.1353/pla.2011.0049

Lyon, L. (2007). Dealing with data: Roles, rights, responsibilities and relationships. Consultancy Report. UKOLN. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

Lyon, L., Ball, A. Duke, & Day, M. (2012a). Community capability model framework white paper. UKOLN, University of Bath, Bath. Retrieved from http://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-24042012.pdf

Lyon L., Ball, A. Duke, M. & Day, M. (2012b). Developing a community capability model framework for data-intensive research. In iPres 2012: Proceedings of the 9th International Conference on the Preservation of Digital Objects, Toronto, Canada.

Research Data Canada. (2008). Stewardship of research data in Canada: A gap analysis report. Retrieved from http://publications.gc.ca/collections/collection_2009/cnrc-nrc/NR16-123-2008E.pdf

Sallans, A., & Lake, S. (2014). Data management assessment and planning tools. *Research Data Management: Practical Strategies for Information Professionals*. Purdue University Press. Retrieved from http://www.jstor.org/stable/j.ctt6wq34t.7

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PloS ONE, 6*(6), e21101. doi:10.1371/journal.pone.0134826

Tolle, K.M., Tansley, D.S. W., & Hey, A.J. (2011). The fourth paradigm: Data-intensive scientific discovery [point of view]. *Proceedings of the IEEE, 99*(8), 1334-1337.