Who's in Charge? Discovering the Autonomy in an Institutional Data Repository for Research Data Curation and Sharing

Pei-Ni Chiang[1], Jian-Sin Lee[2], and Wei Jeng[3]†

[1] Indiana University Bloomington, Bloomington, IN 47405, USA
[2] University of Washington, Seattle, WA 98195, USA
[3] National Taiwan University, Taipei 10617, Taiwan
wjeng@ntu.edu.tw

Abstract. To facilitate data sharing, more and more research data infrastructures have been built. However, less attention is paid to the needs of researchers as data producers in the context of traditional OAIS-compliant institutional data repositories. Meanwhile, researchers usually complete data management tasks themselves throughout the research data lifecycle and express a desire to control the data ingestion process. The contradictory between design and the reality suggests a potential need for autonomy in terms of data curation along with frictions between researchers and professional data curators.

In this study, we explore important features of an ideal institutional data repository through designing the NTUData prototype. It is a researcher-centered system that helps integrate the early phases of the data lifecycle into the process of data curation and thus encourage data sharing. Nine participants in the information science field were recruited for a usability test in which the DCP Toolkit was adopted. The results show that researchers prefer to initiate and perform the whole data submission process themselves. They are also concerned about the interoperability to link NTUData to external resources and the interpretability of text labels within this repository. As for their needs towards autonomy, two perspectives with regards to curating and sharing data can be observed, respectively.

Keywords: Research Data Infrastructure, Institutional Data Repository, Data Sharing, Autonomy, Data Curation Profiles.

1 Introduction

Given the rise of e-Research and digital scholarship in the early 2000s, data sharing naturally plays an important role in advancing scientific research as knowledge currency. Releasing data to the public becomes increasingly crucial as promoting reproducibility has reached consensus in many disciplines. Data sharing benefits academic

† This work was done while Pei-Ni Chiang and Jian-Sin Lee were at the Department of Library and Information Science, National Taiwan University.

communities through improving research transparency, enabling data reuse, and serving the purposes of teaching and learning (Jeng, He, & Oh, 2016). As data sharing and management mandates from funding agencies and journal publishers come into play, more and more researchers are aware of the open data issues and opportunities. Also, more and more research data infrastructures (hereafter: RDIs)--including facilities, tools, platforms, training, and services--have been established, aiming to facilitate the preservation and dissemination of data as one of the research outputs.

Today, most RDIs are developed based on the design of a digital archive--the Open Archival Information System (OAIS) reference model (Hockx-Yu, 2006). The OAIS lays great stress on the long-term preservation, dissemination, and access of information through defining the responsibilities of an organization of people and systems and providing a holistic framework to support the implementations of relevant policies and procedures (CCSDS, 2012).

Three roles outside but around the OAIS model are producers, management, and consumers, all of which, along with the OAIS functional model, constitute the OAIS environment (CCSDS, 2012). In the ecosystem of data repositories proposed by Witt (2014), research data are ingested and archived from perspectives of different levels of entities such as publishers (e.g., Dryad), institutions (e.g., University of Michigan's Deep Blue), consortiums (e.g., ICPSR), and nations (e.g., Research Data Australia and UK Data). All of these repositories enable the curation-sharing-reuse process and long-term preservation of research data, and so are institutional data repositories, which are also one of the characters in the ecosystem and usually compliant with the OAIS model.

As an umbrella term used to completely support the entire research data lifecycle for researchers, in this study we scope RDIs to the information infrastructures and services in the following four stages: I) data production, II) pre-data sharing curation, III) data curation & sharing, and IV) data reuse by others. However, while researchers are the dominant actors of Stage I and II who generate and prepare data for sharing, less attention is paid to both of these stages during the current design process of an institutional data repository. Even though data producers (i.e., researchers) can be the most important stakeholder in respect to data-driven research nowadays, their needs may not be properly addressed or even be overlooked with the adoption of the OAIS model, which focuses more on the overall workflow and the OAIS functional model itself other than external producers. Specifically, the support provided by institutional data repositories usually starts from the later stages (i.e., Stage III and IV) and is typically combined with the intervention (e.g., metadata maintenance and content update) of professional curators within the institutions.

In practice, nonetheless, many data management tasks are accomplished by researchers themselves during different stages within the RDI context, mainly Stage I and II. For example, in a survey conducted by Whitmire, Boock, and Sutton (2015), principal investigators (hereafter: PIs) and research assistants, rather than the information service staff, perform the majority of the data-related tasks, including metadata creation, quality control, storage and organization, and even data sharing and archiving. Diekema, Wesolek, and Walters (2014) indicated that approximately three-fourths of the faculty members are themselves responsible for managing their stored data and seldom rely on librarians to perform such tasks. Lassi et al. (2016) also mentioned that researchers would like to control the ingestion process in terms of depositing their data. Among these studies, a need for *autonomy* when curating research data can be observed, indirectly suggesting a friction between researchers as data producers and those professional data curators who often dominate the data curation-sharing-reuse process.

Moreover, Sayogo and Pardo (2013) pointed out that organizational support significantly influences researchers' likelihood to share their data, and their motivation is sometimes affected by subjective predictors, e.g., perceptions and trust (Lin, 2006). Verbaan and Cox (2014) also discussed the tension and jurisdiction struggle among different stakeholders within the context of research data management. It is thus imperative to understand the potential frictions between researchers and professional data curators affiliated with the institutions so as to ensure researchers' willingness to deposit data into a repository.

Consequently, to integrate the stages of data production and pre-data sharing curation into an institutional data repository and to investigate the gap of user needs between various phases of the research data lifecycle, this preliminary study aims to explore 1) what features and characteristics a researcher-centered institutional data repository should be equipped with so as to better support researchers to complete their data production-curation-sharing-reuse process and 2) how such a system can be developed and made intuitive and user-friendly. Additionally, we also intend to examine in what aspects researchers' needs for autonomy would be reflected during their research process.

2 System Overview & Feature Description

In order to validate the feasibility of a researcher-centered institutional data repository, we designed NTUData, a prototype of a project-based collaborative system expected to be built in National Taiwan University. Instead of employing the concept of information packages submitted to professional curators as with most OAIS-compliant data repositories, NTUData allows researchers to manage their datasets throughout the whole data production-curation-sharing-reuse process themselves, which facilitates future archiving while a research project is still undertaken (Ember & Hanisch, 2013). Such an active curation approach is capable of empowering researchers to gain control over the way they intend to curate data and integrating data-related tasks into the execution process of their projects in an autonomous, seamless, and painless manner. This would indirectly enhance the possibility for researchers to deposit (and then probably share) their data.

As a PI-centered repository focusing on the course of research cycle, tasks that users can complete with NTUData include but not limited to creating projects, assigning roles of research teams, submitting project-related documents and research datasets, filling out metadata fields, setting an embargo period, controlling data access, and commenting on files uploaded.

Three core modules are included in the NTUData prototype, namely *My Projects*, *My Data*, and *Data Usage Analysis*. A PI can establish a research project in the *My Projects* module and manage all the data generated along the way towards the project completion in the *My Data* module. Also, the PI is able to set members in the research team as various roles, and internal users are allowed to comment on datasets submitted. Both of the features make it easier for researchers to communicate and document important milestones or specific steps of their projects so that a robust online collaborative network is shaped. As for the *Data Usage Analysis* module, users can check data metrics regarding the usage of their data (e.g., the counts of views, downloads, and citations of the published data) in a visualized way. For a better understanding of NTUData, three example interface screenshots are demonstrated in Appendix 1 and also available on OSF[2].

3 Pilot User Testing

To ensure the design of NTUData functional and user-friendly, an evaluation from our target user group is needed since this repository aims for empowering researchers to curate and share research data themselves. For a pilot usability test, nine participants from the information science field at National Taiwan University were recruited, including four faculty members, four doctoral students, and one master's student in the thesis track. An overview of our participants and their areas of research is provided in Table 1.

Table 1. An overview of the participants' areas of research.

ID	Status	Research Focus
P01	Master's Student	Human-Computer Interaction
P02	Doctoral Student	Natural Language Processing
P03	Doctoral Student	Digital Humanities
P04	Doctoral Student	Social Network Analysis
P05	Assistant Professor	Information Security
P06	Professor	Information Retrieval
P07	Doctoral Student	Data Reuse
P08	Assistant Professor	Digital Humanities
P09	Associate Professor	Information Management

The user testing process was divided into five sessions. In the beginning, a pre-test survey was given to the participants, primarily focusing on data characteristics and their attitudes towards data sharing. Secondly, participants were asked to briefly share their daily data management practices, such as how and where they store their datasets. In the third session, two tasks were assigned to participants and had to be accomplished using the NTUData prototype, including establishing a research project and uploading a data file with some other features, e.g., role assignment and data access control. During this process, participants' perceptions and thoughts were captured by the think-

[2] https://osf.io/nkpy8/?view_only=693fa0d19a9c43d9a6130244e105fcee

aloud technique that helps communicate what satisfied and frustrated them in this repository. Fourthly, participants would complete a system evaluation form designed based on the System Usability Scale (SUS) created by Brooke (1996). As for the last session, a semi-structured post-test interview was conducted to gather information about key features that the participants consider necessary for an ideal institutional data repository and their comments on the current prototype.

It is noted that questions asked in Session 1, 2, and 5 were extracted from the *Data Curation Profiles (DCP) Toolkit* developed by the research team from Purdue University and University of Illinois at Urbana-Champaign (Carlson, 2010). The 13 modules included in the DCP Toolkit not only help professionals understand researchers' data practices and preferences but also encourage researchers to thoughtfully consider their own data needs. For our user testing, questions were selected from nine modules on the basis of relevance to features of NTUData. More details are shown by the table in Appendix 2.

4 Preliminary Findings

Overall, participants' feedback regarding the NTUData prototype is positive. Eight out of nine agreed that the interfaces are user-friendly and easy to learn. "The ability for me to be able to submit the data to a repository myself, meaning that I initiate and perform the whole submission process" was considered the highest priority by eight participants. The main reasons include that submitting data themselves would be more time-saving and that they are more familiar with their own datasets in comparison with professional curators, so they know a better way to curate the data.

Suggestions for improvement with regards to this repository were centered in two aspects. First of all, five participants mentioned the importance of the interoperability to link NTUData to external resources, e.g., using Single Sign-On (SSO) authentication instead of creating a new account, or connecting NTUData with project management platforms of grant funders to reduce the burden of repeatedly typing duplicate fields when establishing a project in NTUData. Secondly, some text labels of the metadata fields containing jargons that are used often in the LIS field were found unclear by four participants, such as "creator" and "relation" of which there were no specific definitions.

Additionally, we found that researchers' needs for autonomy when interacting with RDIs can be described from the following two perspectives.

1) *Curating data when a research project is still undertaken.* Most participants perform their data management tasks without the intervention of professional data curators as discussed previously. We also found that they prefer to store data wherever the data can be easily accessed for them, such as personal devices, cloud storage, git servers, and databases, instead of data repositories. During the participants' research process, the needs for autonomy are reflected in respect of how to organize and describe data, mentioned by six and three participants respectively. More specifically, the participants tend to group and name data files for personal use or collaboration with other researchers rather than taking into account potential data reusers outside their research teams.

2) *When and to what extent researchers share their data.* Eight participants are willing to share data with anyone after their research comes to an end, while before that, seven participants would only share data with immediate collaborators. In addition, there are more conditions mentioned in terms of data sharing, such as whether data consumers are in the same research team or discipline and what their status is (i.e., faculty members or doctoral/master's students). We thus noted a desire expressed by the participants to keep maintaining autonomous control over such subtle adjustments both before and after they complete their research with regards to making research data openly accessible with a data repository.

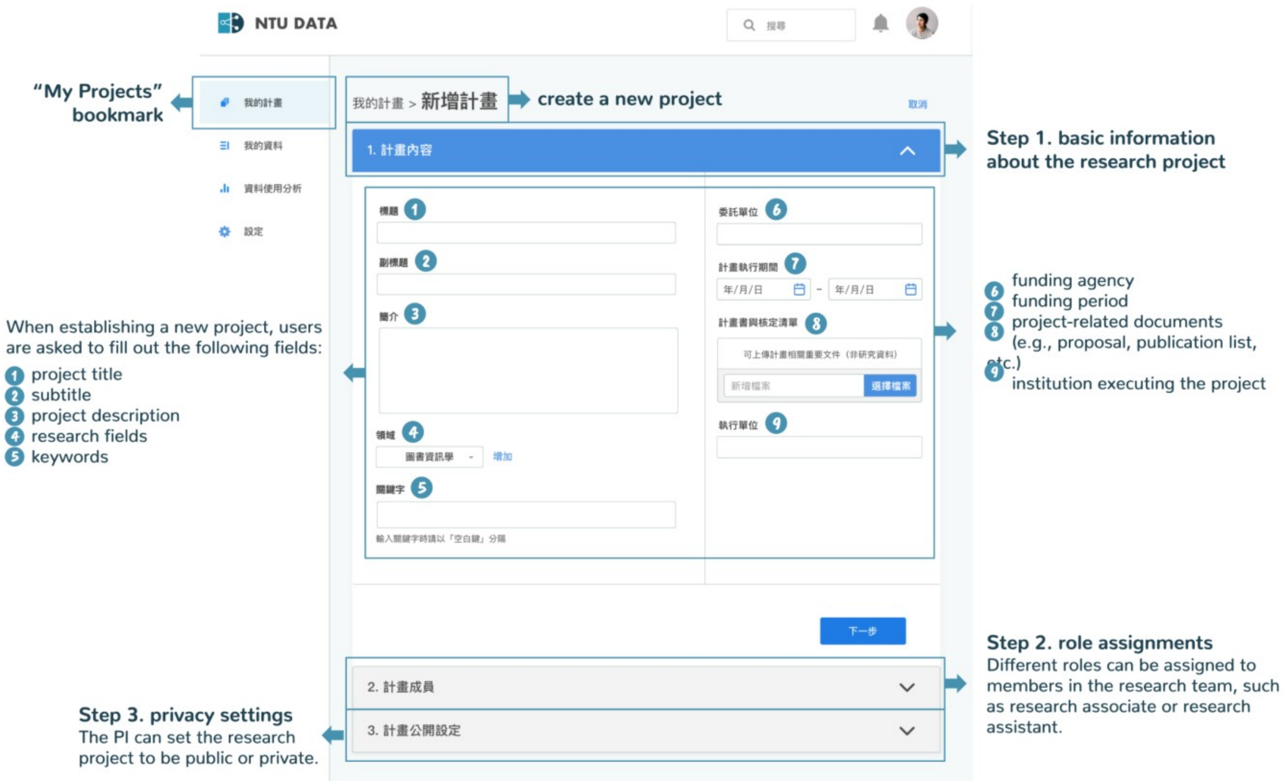
5 Concluding Remarks & Future Work

Aiming to integrate the stages of data production and pre-data sharing curation into researchers' interaction process with RDIs, the NTUData prototype was introduced as an institutional data repository. This researcher-centered system allows users to manage, curate, and share their data themselves, which is different from the traditional OAIS-compliant ones. The adoption of the DCP Toolkit helps identify critical features and characteristics that this repository should be equipped with. Furthermore, the needs for autonomy regarding data curation and sharing were discussed to explore potential frictions between researchers and professional data curators.

Moving forward, we plan to continue creating user interfaces for other modules of NTUData and make it more robust by investigating the needs of researchers from disciplines other than information science and conducting interviews with different stakeholders (e.g., the university librarians or the IT service staff). In addition, more studies are needed to reveal how to encourage researchers to deposit their data into a repository in the early phases of the data lifecycle with the idea of active curation as well as how to establish a trusted relationship between researchers and professional curators so as to maximize the likelihood of being best curated and made open in respect to research data.

Acknowledgement. This work was financially supported by the Ministry of Science and Technology (MOST) in Taiwan, under #MOST 108-2636-H-002-002- and MOST107-3017-F-002-004, and the Center for Research in Econometric Theory and Applications (Grant no. 107L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Appendix 1: Example interface screenshots of NTU Data



"My Projects" bookmark

create a new project

Step 1. basic information about the research project

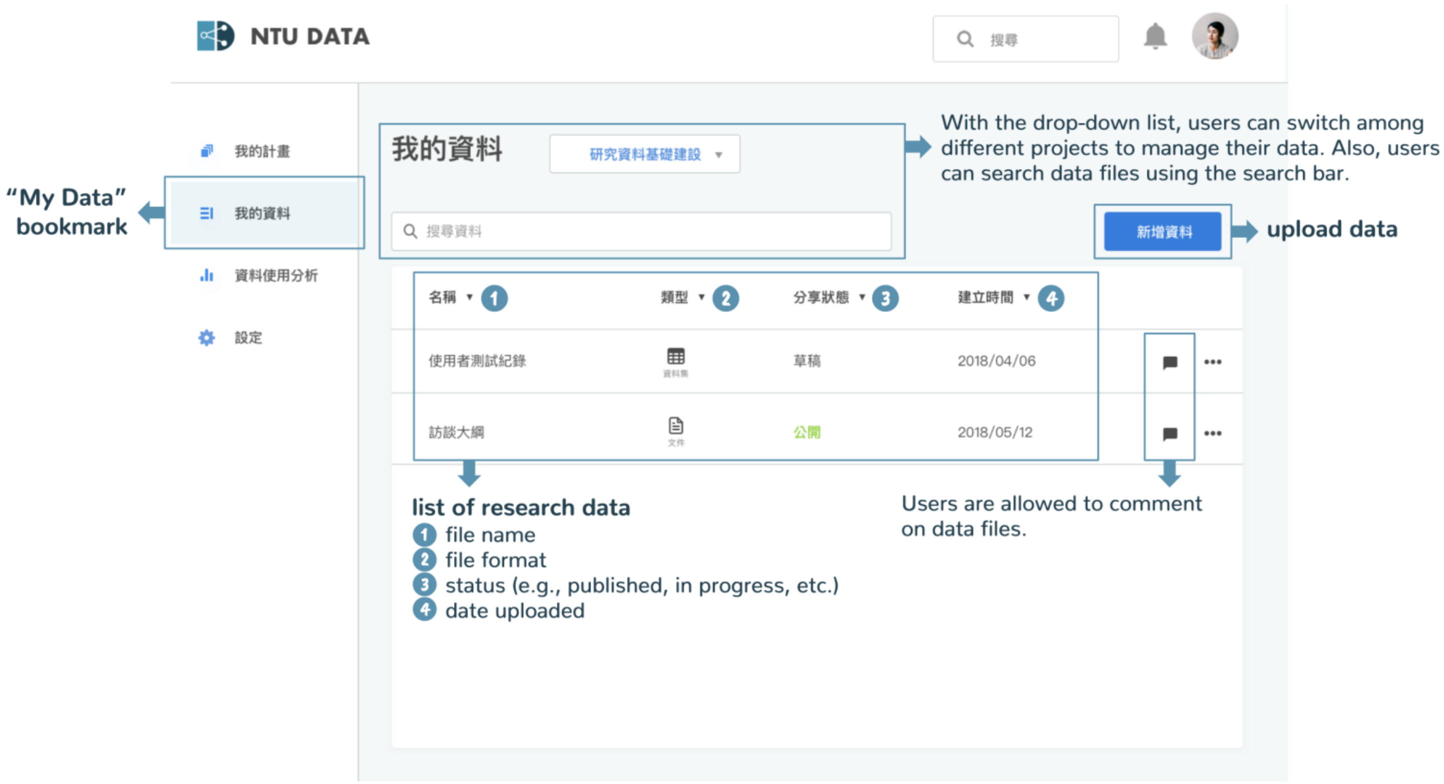
When establishing a new project, users are asked to fill out the following fields:

1. project title
2. subtitle
3. project description
4. research fields
5. keywords

6. funding agency
7. funding period
8. project-related documents (e.g., proposal, publication list, etc.)
9. institution executing the project

Step 2. role assignments
Different roles can be assigned to members in the research team, such as research associate or research assistant.

Step 3. privacy settings
The PI can set the research project to be public or private.



我的資料

名稱 ①	類型 ②	分享狀態 ③	建立時間 ④
使用者測試紀錄	資料集	草稿	2018/04/06
訪談大綱	文件	公開	2018/05/12

"My Data" bookmark

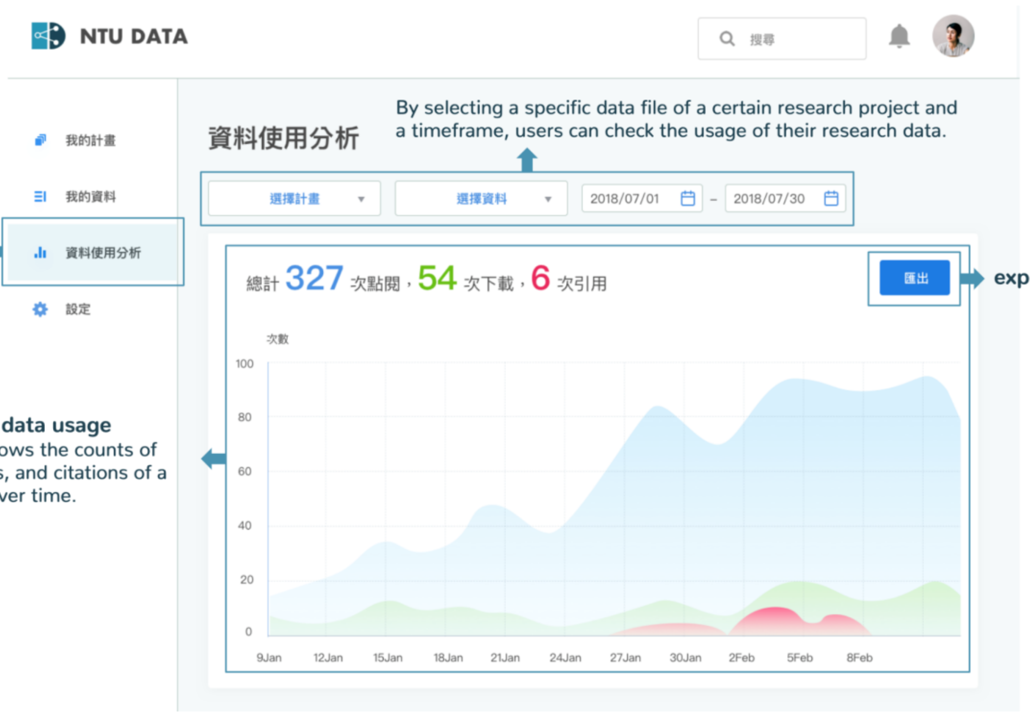
With the drop-down list, users can switch among different projects to manage their data. Also, users can search data files using the search bar.

upload data

list of research data

① file name
② file format
③ status (e.g., published, in progress, etc.)
④ date uploaded

Users are allowed to comment on data files.



資料使用分析

總計 327 次點閱，54 次下載，6 次引用

By selecting a specific data file of a certain research project and a timeframe, users can check the usage of their research data.

“Data Usage Analysis” bookmark

export usage data

visualization of data usage
The area chart shows the counts of views, downloads, and citations of a certain data file over time.

Appendix 2: Modules of the DCP Toolkit used in the pilot user testing

Modules	Session 1	Session 2	Session 5
M1. The Dataset		●	
M2. The Lifecycle of the Dataset			
M3. Sharing	●		
M4. Access		●	
M5. Transfer of Data/Ingest into a Repository			●
M6. Organization and Description of Data			●
M7. Discovery	●		
M8. Intellectual property			
M9. Tools			●
M10. Linking/Interoperability			
M11. Measuring Impact			●
M12. Data Management		●	
M13. Data Preservation			

References

1. Brooke, J.: SUS: A ‘Quick and Dirty’ Usability Scale. In: Jordan, P.W., Thomas, B., McClelland, I.L., Weerdmeester. B. (eds.) Usability Evaluation in Industry. CRC Press, Bristol (1996).
2. Carlson, J.: Data Curation Profiles Toolkit: Interview Worksheet, Paper 3 (2010). http://dx.doi.org/10.5703/1288284315652
3. CCSDS: Recommendation for Space Data System Practices--Reference Model for an Open Archival Information System (OAIS). Magenta Book, CCSDS Secretariat (2012).
4. Diekema, A., Wesolek, A., & Walters, C.: The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories. J. Acad. Libr. **40**, 322-331 (2014).
5. Ember, C. & Hanisch, R.: Sustaining Domain Repositories for Digital Data: A White Paper. White paper, ICPSR (2013).
6. Hockx-Yu, H.: Digital preservation in the context of institutional repositories. Program-Electron. Lib. **40**, 232-243 (2006).
7. Jeng, W., He, D., & Oh, J.: Toward a conceptual framework for data sharing practices in social sciences: A profile approach. Proc. Assoc. Info. Sci. Tech. **53**, 1-10 (2016).
8. Lassi, M., Johnsson, M., Golub, K., Witt, M., & Horstmann, W.: Research data services: An exploration of requirements at two Swedish universities. IFLA J. **42**, 266-277 (2016).
9. Lin, H.: Impact of organizational support on organizational intention to facilitate knowledge sharing. Knowl. Man. Res. Pract. **4**, 26-35 (2006).
10. Sayogo, D. & Pardo, T.: Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. Gov. Inform. Q. **30**, S19-S31 (2013).

11. Verbaan, E., & Cox, A.: Occupational Sub-Cultures, Jurisdictional Struggle and Third Space: Theorizing Professional Service Responses to Research Data Management. J. Acad. Libr. **40**, 211-219 (2014).
12. Whitmire, A., Boock, M., & Sutton, S.: Variability in academic research data management practices. Program-Electron. Lib. **49**, 382-407 (2015).
13. Witt, M.: Defining and Deploying an Institutional Data Repository Service at Purdue (PURR). Libraries Faculty and Staff Presentations, Purdue e-Pubs (2014).