Data Curation Practices in Institutional Repositories: An Exploratory Study

Dong Joon Lee
Besiki Stvilia
College of Communication and Information, Florida State University
142 Collegiate Loop, Tallahassee FL 32306-2100
dl10e@my.fsu.edu; bstvilia@fsu.edu

ABSTRACT
This poster reports on the preliminary findings of a study that examines data curation work in institutional repositories (IR). The study identified three perceived roles played by IR staff (i.e., data curator, IR manager, and metadata specialist) and role-specific sets of activities and skills. The findings of this study can inform the development of best practices and effective infrastructure support for data curation in the context of IRs, as well as teaching data curation in LIS schools.

Keywords
Data curation, research data, institutional repository, data activities, data skills.

INTRODUCTION
Many research universities have operational institutional repositories (IRs) that provide open access to the digital content produced by the universities’ communities. However, if the repositories are filtered by the inclusion of research data objects, the number of universities that have such IRs is dramatically decreased. According to Lee and Stvilia (2012), in 2012 only half of Association of American Universities (AAU) member universities, which are the leading 62 public and private research universities located in the United States and Canada, had IRs that contained research data objects. In addition, the Association of College and Research Libraries (ACRL) published a report about current practices and plans for the future of research data services in academic libraries. According to that report, only a small number of academic libraries in the United States and Canada currently offer research data service, and about 25 to 30 percent of the 351 ACRL-member libraries are planning to provide some research data-related services within the next two years (Tenopir, Birch, & Allard, 2012).

There are growing needs and opportunities to share, reuse, and aggregate data from different contexts. Funding agencies now require applicants to submit plans for disseminating and providing access to research data (IMLS, 2011; NEH, 2011; NIH, 2010; NSF, 2010). In addition, governments and scholarly communities emphasize the access and sharing of research data (Aalbersberg & Kähler, 2011; Office of Science and Technology Policy, 2013; Research Councils UK, 2011; Thomson Reuters, 2012). The increased access to research data improves the impact, as well as the efficiency and effectiveness, of scientific activities and funding. The access, however, is facilitated not just by appropriate policies but also by the employment of effective infrastructure mechanisms, including enhancing data with effective metadata (Lee & Stvilia, 2014; Moore & Smith, 2007; Simmhan, Plale, & Gannon, 2005).

Understanding and providing effective support to data curation practices in the institutions that now implement institutional data repositories is essential to achieve the objectives of IRs which include but are not limited to sharing, accessing, controlling and preserving knowledge and data (Markey, Rieh, St.Jean, Kim, & Yakel, 2007; Westell, 2006). As many research institutions plan to provide some types of research data services (Tenopir et al., 2012), it is important to study the current practices of data curation in IRs. In particular to develop effective infrastructure configuration templates, it is essential to understand data curation related activities in IRs, including different roles played by IR staff, and role specific differences in needs for infrastructure support (Foster, Jennings, & Kesselman, 2004). Furthermore, identifying IR role-specific differences can inform data curation curricula development in LIS schools. This poster reports preliminary findings of a study that contributes towards those goals.

METHODOLOGY
Studying the practices of research data curation requires multifaceted contextual analysis (Borgman, Wallis, & Enyedy, 2007; Stvilia et al., 2014). Hence this study, too, required a research design that could help capture and analyze various sociotechnical and cultural factors that might shape the context of data curation work in IRs. The study used Activity Theory (Engeström, 1987;
Leontiev, 1978) to guide the design of a protocol for semi-structured interviews. Activity Theory suggests general activity structure and context relationships of data curation in IRs. It helps identify and reason about activities generated by subject (e.g., IR curator) and object (e.g., maintaining data in actionable, reusable state) interaction and different mediating factors (e.g., norms, rules, tools, and work environments) of activities’ contexts, including stakeholder communities and their cultures. In the case of IRs, the communities comprise data providers, users, and curators of scholarship and research data managed by those IRs.

This poster reports on data collected from five participants from five different universities in the US. The selection of participants was guided by two criteria. To be eligible for participation in the study, participants had to work for IRs that stored and curated research data objects and had to be operated by AAU member universities. The study relied on findings of an earlier study to identify IRs that met those selection criteria (Lee & Stvilia, 2012). According to that previous research, in 2012, half of the AAU member universities, which are the leading 62 research universities, had IRs that contained research data objects. AAU universities are comparable by size and amount of research conducted. In addition, due to the small size of the study’s population, participants were selected using purposive and snowball sampling techniques. These made the sample relatively homogeneous in terms of data curation resources available and community affiliation. This preliminary research was limited to exploring the following research questions:

• What are the typical data curation-related activities and contexts in IRs?

• How does IR staff perceive their role-related activities and skills?

• What user services do IRs currently support?

FINDINGS

Data Curation Activities and Contexts

The IR objectives mentioned by participants included not only collecting, storing, preserving, and providing access to data, but also supporting entire research data lifecycle from planning to publication. Hence the set of objectives, data curation activities in IRs, also encompassed most actions from general models of research data and related activities (e.g., DCC Curation Lifecycle Model, OAIS). Figure 1 describes data curation activities in IRs and contextual elements around those activities. Interview participants also indicated that they spend significant time on forming an idea of data curation with data providers and creating metadata for the data. One of the interview participants was explicitly telling that:

80-85% of the time that I spend is … going and meeting with people and getting people to coordinate to work. (s1)

The first meetings with data providers to conceptualize and understand their data and curation needs notably affected the later activities, such as receiving data and creating metadata. They tended to decide what types of help are being needed by the providers, who was the right person to help create metadata, and how the data could be organized and stored. Storing datasets from different disciplines may require IRs to support different metadata schemas and knowledge organization systems, and map among those to enable an integrated access to the datasets (Wu, Stvilia, & Lee, 2012).

I think how much assistance each researcher needs, really depends on the first interview. (s4)

Sometimes in month long conversations … DC [metadata] terms are very straightforward terms. However …, they can be interpreted in many ways. (s3)

Participants mentioned various norms, rules, tools, and division of labor, which mediated their activities. Figure 1 includes incomplete lists of those components and factors.

IR Staff’s Role-Related Activities and Skills

The second research question explored IR staff’s perception of their roles, role-related activities and skills needed to curate and manage research data. The analysis identified three primary roles played by IR staff: data curator, IR manager, and metadata specialist. The role-related activities consisted of planning entire data curation governance structures in their IRs, meeting with data providers to move forward curation workflows, managing IRs on a daily basis, and promoting their IRs to bring in as many datasets as they could into their repositories and to increase visibility and use of the IRs. Table 1 shows the relationships between roles and activities. Dotted lines in Table 1 point to the possibilities of overlaps among the role-specific sets of activities. Participants mentioned that communication skills, digital library and metadata literacy, and familiarity with complex and diverse research data were the skill set that they needed to complete their tasks. Table 1 also indicates the relationships and overlaps between skills and roles. In addition, one of the participants mentioned that the
skill set needed was dependent on the levels of complexity and maturity of data curation at a particular IR.

<table>
<thead>
<tr>
<th>Activities</th>
<th>Data Curators</th>
<th>IR Managers</th>
<th>Metadata Specialists</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Build or plan data governance structures</td>
<td>• Manage IRs on a daily basis</td>
<td>• Help data providers to create appropriate metadata for their dataset</td>
</tr>
<tr>
<td></td>
<td>• Consult with data providers and connect them to metadata specialists or IR managers</td>
<td>• Work with data providers to help add metadata and upload data into IRs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Outreach and educate campus community</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skills</td>
<td>• Familiarity with research data (e.g., ability to handle data complexity and diversity)</td>
<td>• Technical skills (e.g., knowledge of digital library architecture and software)</td>
<td>• Metadata knowledge</td>
</tr>
<tr>
<td></td>
<td>• Ability to communicate well with other people and to work in a team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR User Activities / Services</td>
<td>• Searching</td>
<td>• Providing Creative Commons license</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Browsing</td>
<td>• Providing Usage statistics services</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Social networking /Sharing/Tagging</td>
<td>• Running data curation workflow wizard</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Citing</td>
<td>• Supporting project management functions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Storing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Downloading</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. IR staff’s role-related activities and skills and the user activities/services IRs currently support.

I think the skills what I would identify for right now for data curator or data curation professionals, would be split. It's a hard question. I mean, you know I think [it] largely depends on how matured your data curation program is in your university. (S1)

The comparison of the IR staff’s skills identified by this study to the set of data skills needed in genome annotation curation, one of data curation activities, identified by Huang et al (2012) points to some differences and similarities (see Figure 2). The interpersonal skills, such as communication and teamwork abilities, do not have a match in the Huang et al model. On the other hand, the adaptive skills, the skills needed to determine quality (e.g., value and relevancy) of research data, found in the Huang et al model, did not appear in the set of IR staff’s skills. This difference could be caused by differences in the expectations of the levels of data curation provided by IRs and subject specific data repositories. Subject specific data repositories and their staff are expected to provide more comprehensive data curation services including data annotation and quality assessment. Hence, curators of subject-specific data repositories are expected to be subject specialists with good understanding and knowledge of both the subject area and data management.

**IR User Activities and Services**

The literature suggests that current reuse rate of curated data by end-users is relatively low (e.g., Arlitsch & O’Brien, 2012; St. Jean, Rieh, Yakel, & Markey, 2011). The identification of the current set of data user services provided by IRs can serve as a foundation for a related future research, which would explore user satisfaction with the current data services and needs for new user services and tools to increase data reuse. Table 1 presents current user activities or services being supported by IRs.

**CONCLUSION**

This poster reports on the preliminary results of the study that explore the data curation practices in IRs using an Activity Theory lens. IR data curation activities and skills were clustered by the three perceived major roles played by IR staff. Studying data curation work in IRs, identifying different roles played by IR staff and role-specific needs for skills, can inform the development of best practices, infrastructure configuration templates, as well as the teaching of data curation in LIS schools. The future research related to this one will explore metadata needs and practices, particularly identifier needs and uses in the context of data curation in IRs.

**ACKNOWLEDGMENTS**

We express our gratitude to all of the interview participants and the feedback/comments from the 2014 ASIST Research Data Access and Preservation (RDAP) Summit, San Diego, CA.

**REFERENCES**


