Toward engaging researchers in research identity data curation

Abstract

This collaborative planning grant project, addressing the IMLS priority of establishing a shared, distributed, national digital platform, will explore researcher participation in research identity management systems. *In particular, it will address the need to have greater knowledge of how to design scalable and reliable solutions for research identity data curation by examining researchers' perceived value of research identity data and services; motivations to participate in and commit to online research identity management systems, and contribute to research identity data curation.* Accurate research identity identification and determination are essential for effective grouping, linking, aggregation, and retrieval of digital scholarship; evaluation of the research productivity and impact of individuals, groups, and institutions; and identification of expertise and skills. The reliability and scalability of those services will be critical to the success of national, distributed, digital information infrastructure that IMLS strives to build. There are many different research identity management systems, often referred to as research information management (RIM) or current research information systems (CRIS), from publishers, libraries, universities, search engines and content aggregators with different data models, coverage, and quality. Although knowledge curation by professionals usually produces the highest quality results, it may not be scalable because of its high cost. The literature on online communities shows that successful peer curation communities which are able to attract and retain enough participants can provide scalable knowledge curation solutions of a quality that is comparable to the quality of professionally curated content. Hence, the success of online research identity management systems may depend on the number of contributors and users they are able to recruit, motivate, and engage in research identity data curation.

The government, funding, and accrediting agencies requiring universities to curate and share research information and data, as well as the surge of interest on academic campuses in open access, and the use of research information and scholarship for expertise identification and overall institutional reputation management, make the curation of research identity data a priority for academic libraries. Although there is a significant body of literature on authority control in library databases, automated entity extraction, determination and disambiguation on the Web, and the design and management of online peer-production communities, there is still a dearth of research on researcher participation in and commitment to online research identity data management systems and communities. This project will address that need.

The outcomes of this exploratory research will include but not be limited to a qualitative theory of research identity data and information practices of researchers, quantitative model(s) of researchers' priorities for different online research identity data and services, the factors that may affect their participation in and commitment to online research identity management systems, and their motivations to engage in research identity data curation. The study's findings can greatly enhance our knowledge of the design of research identity data/metadata models, services, quality assurance activities, and, mechanisms for recruiting and retaining researchers for provision and maintenance of identity data. Design recommendations based on this study can be adopted in diverse settings and can produce improved services for multiple stakeholders of research identity data such as researchers, librarians, students, university administrators, funding agencies, government, publishers, search engines, and the general public. To ensure as broader impact as possible results of the study and related datasets will be openly disseminated through a project website, data repositories, conferences, blogs, and open access peer reviewed journals.

Toward engaging researchers in research identity data curation

1. Statement of Need

This collaborative planning grant project, addressing the IMLS priority of establishing a shared, distributed, national digital platform, will explore researcher participation in research identity management systems. In particular, it will address the need to have greater knowledge of how to design scalable and reliable solutions for research identity data curation by examining researchers' perceived value of research identity data and services, motivations to participate in and commit to online research identity management systems, and contribute to research identity data curation.

Scientific research, as well as the evaluation of research productivity and impact of individual researchers and institutions and related policy development have become increasingly data driven. There are growing needs as well as opportunities to share, reuse, and aggregate data from different contexts. The Institute for Museum and Library Services (IMLS, 2015), the National Endowment for the Humanities (NEH, 2015), the National Science Foundation (NSF, 2015), and the National Institutes of Health (NIH, 2015) require applicants to submit data management plans, including plans for disseminating and providing access to digital scholarship, research data and related metadata. To maintain that research data and scholarship in a usable/reusable and discoverable state for ongoing research, education, reporting, verification, and evaluation, it is essential to curate research entity data or metadata. Entities are distinguishable objects that can be concrete or abstract (Elmasri & Navathe, 2000). Examples of entities are books, authors, geographic locations, proteins or genes. A set of important attributes that characterize a particular entity constitutes the entity's metadata profile, which can be included in reference databases (e.g., authority databases) and used for entity determination and disambiguation. In biology, taxonomists may need to determine whether a particular specimen belongs to an established taxon or if it represents a new taxon. Genomics researchers may need to distinguish the sample's identity in order to identify genotype-phenotype relationships at the individual or population level. Librarians, in particular catalogers, may need to resolve different entities in bibliographic descriptions in order to link and collocate related works and publications. Administrators and bibliometrics / scientometrics researchers may need to resolve author names to evaluate the productivity and impact of individual scientists, groups, or institutions; identify potential collaborators, experts, and research community structure; and track alumni careers for reporting, planning and fundraising (Cucerzan, 2007; Hinnant et al., 2012; OCLC Research, Task Force on the Registering Researchers, 2014; Stvilia et al., 2011; Wu et al., 2012). Search engines, social media platforms, and intelligence agencies may need to resolve multiple email and social media accounts to an identity to get more accurate understanding of individual users' or groups' web behaviors, preferences, conversations, sentiments, or a user's social network structure and dynamics. Effective curation and aggregation of data, however, may require knowledge of community, disciplinary, and cultural differences in data and metadata quality requirements, rules, norms, and references sources (Atkins et al., 2003; Stvilia et al., 2007). Hence, institutional repositories (IRs) may need the participation of subject specialists, librarians, and most importantly researchers themselves in data curation activities to ensure the quality and reliability of their metadata and data services (Lee & Stvilia, 2014; Lee, 2015; Tenopir et al., 2012).

There have been distinct domain-specific approaches to entity metadata management. Libraries have been controlling metadata for bibliographic entities for very long time (Svenonius, 1989). They have used a set of standards and trained professionals to produce and curate authority metadata to ensure its quality. The problem with those standards, however, is that it is difficult to achieve widespread adoption, consistent interpretation,

and use (Stvilia et al., 2005). In addition, there could be more than one standard for the same entity and more than one database could curate knowledge about an entity instance. Libraries have tried to address this issue with aggregation mechanisms such as the Virtual International Authority File (VIAF), which aggregates authority metadata produced by large libraries around the World. Currently, VIAF also links to entity instance metadata from open crowdsourced authority databases such as Wikidata. Its scope, however, is determined by the scopes of authority databases of participating libraries. Libraries, traditionally, have not curated authority data of researchers who authored journal papers or conference proceedings only.

Online peer or socially curated knowledge databases, such as Wikipedia and Wikidata, have become one of the most important aggregators and sources of knowledge on Web. The world's largest encyclopedia – Wikipedia – comprising more than 200 language specific encyclopedias is a major source of general reference knowledge. Likewise, another Wikimedia project – Wikidata – aggregates factual knowledge on various entities in multiple languages and makes it accessible in a machine processable, structured format to both human and computational agents. Still, these databases are far from being comprehensive with regard to research/scholarly identity data as their scopes are shaped by Wikipedia's notability criteria and the preferences of individual editors who seed biography articles and/or identity records for a particular scholar.

Reliable and *scalable* determination and disambiguation of research identity are essential services that the National Digital Platform needs to provide to enable distributed grouping, linking, aggregation, and retrieval of scholarship; evaluation of the research productivity and impact of individuals, groups, and institutions; and identification of expertise. There are many different research identity management systems, often referred to as research information management (RIM) or current research information systems (CRIS), from publishers, libraries, universities, search engines and content aggregators with different data models, coverage, and quality (e.g., ExpertNet.org, Google Scholar, ORCID, Reachnc.org, ResearchGate). These databases employ different approaches and mechanisms to curating research identity information: manual curation by information professionals and/or users, including the subjects of identity data; automated data mining and curation scripts (aka bots); and some combination of the above. With universities engaging in the curation of digital scholarship produced by their faculty, staff, and students through IRs, some of these universities and IRs try to manage research identity profiles of their contributors locally (e.g., Expertnet.org, Stanford Profiles). Some large academic libraries use the VIVO[1] ontology to make their data, including researcher identity information, discoverable and linkable for cross-institutional retrieval, processing and analysis both by human and computational agents. The use of ontologies and Semantic Web technologies can make data machine processable and "understandable" and hence may reduce the cost of data aggregation and analysis. Ultimately, however, the completeness and accuracy of data make RIM systems reliable and successful. While knowledge curation by professionals usually produces the highest quality results, it is costly and may not be scalable (Salo, 2009). Libraries and IRs may not have the sufficient resources to control the quality of large scale uncontrolled metadata often batch harvested and ingested from faculty authored websites and journal databases (Salo, 2009). They may need help from IR contributors and users to control the quality of research identity data.

The literature on online communities shows that successful peer curation communities which are able to attract and retain enough participants can provide scalable knowledge curation solutions of a quality that is comparable to the quality of professionally curated content (Giles, 2005). Hence, the success of online research identity management systems may depend on the number of contributors and users they are able to recruit, motivate, and engage in research identity data curation. There is a significant body of research on what makes peer

[1] http://www.vivoweb.org/

knowledge creation and curation groups and communities successful. Some of the issues and factors that may affect the success of peer curation of knowledge are peer motivations to contribute, the effectiveness of work articulation and coordination, task routing, and quality control (e.g., Cosley et al., 2006; Nov, 2007; Stvilia et al., 2008). Most of the previous research, however, has focused on encyclopedia, question answering and citizen science communities. There has been little investigation of the peer curation of research identity data.

The National Digital Platform, in addition to shared content, software, and hardware modules, may need to provide shared research based knowledge for effective design, configuration, and management of those resources. In particular, a shared knowledge base is necessary to design effective sociotechnical mechanisms to recruit, build, and manage user communities around library resources, and engage them in library events and activities, which may include the curation of digital research identity and authority data. Although there is a significant body of literature on authority control in library databases, automated entity extraction, determination and disambiguation on the Web, and the design and management of online peer-production communities, there is still a dearth of research on researcher participation in and commitment to online research identity management information systems and communities. In particular, it is important to have greater understanding of what researchers' perceived value for different research identity data and services is; and what affects researcher's decision to participate in research identity data curation in online research identity management systems. This study will address those needs by examining the research questions specified in the subsections below.

1.1 Needs for and Uses of Research Identity Data

There have been considerable deliberations on the needs for and uses of research identity data and how to manage that effectively in LIS research and practice communities (e.g., NISO Altmetrics Initiative[2]; Research Data Alliance[3], OCLC Research, Task Force on the Registering Researchers, 2014). An OCLC task force identified 5 stakeholder groups of research identity data: researcher, funder, university administrator, librarian, and aggregator (OCLC Research, Task Force on the Registering Researchers, 2014). For the researcher stakeholder group, the task force formulated five needs: disseminate research, compile all publications and other scholarly output, find collaborators, ensure network presence is correct, and retrieve others' scholarly output to track a given discipline. It is important to mention that this set of needs was compiled based on expert opinions of task force members, supplemented with a scenario based analysis. It would be valuable to test this typology empirically as well as to investigate what could be some of the disincentives for researchers to participate in online research identity data sharing and curation.

As different units in universities (e.g., office of research) are increasingly interested in collecting and analyzing research output for the purposes of reporting, accreditation, and/or organizational reputation management, those activities and interests overlap with the traditional interests of academic libraries. Hence, academic libraries have to better align their digital services with those broader organizational needs and priorities not to see their role and image diminished in their institutions (Dempsey, 2014; Tenopir et al., 2012). One straightforward approach would be to add research identity management services to institutional repositories (Palmer, 2013). Indeed, there is evidence from the practice that adding research identity management services or RIM to an IR might increase researchers' interest in the IR (Dempsey, 2014; Tate, 2012). The increased interest in an IR,

[2] http://www.niso.org/topics/tl/altmetrics_initiative/
[3] https://rd-alliance.org/groups

however, might not always translate in increased use of an IR and/or increased engagement in research identity data curation as multiple global research information systems offer similar services and strive for researcher's attention and contributions (e.g. ResearchGate, Academic.edu). Libraries need researcher engagement in research identity data curation to provide scalable, high quality, research identity management services. Relying solely on automated mining, extraction and aggregation of research identity data might result in poor quality. Salo (2009) discussed metadata quality problems in IRs caused by the lack of authority control and of involvement of trained professionals. The author also pointed to the failure of automated batch ingest scripts to control names and often dumping uncontrolled, poor quality metadata from faculty authored webpages and journal websites in IRs without proper identity determination and disambiguation.

Indeed, one of the main inhibitors of data sharing and use/reuse is concern about the quality of data. The data owners may be concerned about the quality and potential misuse or misinterpretation of data. The users, on the other hand, may not have sufficient resources and/or access to the process that generated the data to evaluate its quality, and hence may mistrust and not use the data. Quality of data determines the quality of findings and decisions (Stvilia et al, 2007). High quality data is essential if one is to make high quality decisions or to justify, validate and evaluate existing decisions and results. For example, the quality of research identity data is essential to evaluate research, and/or make related policy decisions. Data and related metadata can be used for different purposes and can represent different levels of importance in different activities and to different stakeholders of that data (Greenberg, 2001; Stvilia & Gasser, 2008). Furthermore, researchers may rely on different properties and cues of data to assess its relevance, quality, and value (Faniel & Jacobsen, 2012; Stvilia et al., 2015). Thus, to facilitate research identity data use, and to engage researchers in its quality assurance, it is important to identify the stakeholders' value structures, and priorities for different research identity data elements and align RIM data structures and quality assurance activities with those priorities. In addition, as with curation of any data, the curation and quality assurance of research identity data have cost. While some of the costs can be alleviated through automated mining, extraction, disambiguation, and aggregation of research data and information, semantically heavier tasks such as skill and expertise identification or answering expert level questions may require researchers' involvement. Researchers' perceived value and importance of RIM services must outweigh the cost of a curation task(s) for them to get engaged and contribute to that task. Hence, this study will examine the following research questions:

1. *Why do researchers use an online research identity management system(s)?*
    a. *How do researchers use an online research identity management system(s)?*
    b. *What are the data models of scholar/researcher entity of different research identity management systems?*
    c. *How do those data models compare to each other? What are the differences and similarities?*
    d. *What are the perceived values of the different elements of those identity models among researchers?*
    e. *What are some of the data services in those research identity managements systems and what are the perceived values of those services among researchers?*

1.2 Motivations to Participate in Research Identity Data Curation

Effective knowledge creation and curation require high quality contributors and curators, effective work organization, and tools. Some research identity management systems strive to build communities of users around their databases and engage those communities in the curation of research identity profiles, knowledge

sharing through Q&A boards, and recruitment of new users (e.g., ResearchGate). The challenge is how to motivate users to contribute to those activities.

The online community literature shows that volunteer knowledge curators in open peer-production systems like Wikipedia are mostly driven by intrinsic motivations such as their interests in specific areas, which are often shaped by their organizational and ethnic affiliations, hobbies, professional experiences and expertise, and beliefs (Nov, 2007; Stvilia, 2006; Stvilia et al., 2008). For example, Nov (2007) found enjoyment or having fun as the top motivation to contribute in English Wikipedia. Interestingly, these intrinsic motivations can produce both constructive and disruptive behaviors (i.e., trolling and vandalisms) towards the community's objectives. The challenge then is how to encourage good, positive behaviors and contributions and discourage disruptive behaviors. This problem could be mitigated by decreasing the cost and increasing the benefits of system use and making contributions for good members, and increasing the cost of disruptive actions (Stvilia et al., 2008). In particular, a community may deploy specific system architecture components and mechanisms to reduce the cost of system navigation and content discovery and evaluation, as well as to increase the satisfaction of the needs shaping members' intrinsic motivations, such as the ability to self-select for tasks based on personal interest and level of competence (Resnick et al., 2011; Stvilia et al., 2008). Stvilia et al. (2008) argued that some of the factors that determined the success of English Wikipedia were the low barrier of entry (i.e., simplicity of Wiki software), the ability of making changes in an article directly without an intermediary, and autonomy in deciding what kind of contribution to make to the community and when. Identifying user expertise and interests and routing and recommending tasks to users based on those characteristics is another example of using an external intervention to enhance intrinsic motivation (Cosley et al., 2006).

In addition to Wikipedia, researchers examined user motivations to contribute in other online communities. Ames and Naaman (2007) interviewed 13 'heavy' users of a Flickr application and identified four types of motivations for tagging: self-organization, self-communication, social-organization, and social-communication. A study of Flickr collections by Stvilia and Jorgensen (2007) listed eight motivations members might have when organizing photographs into groups: (1) to enable easy finding, (2) for easy sharing, (3) for archiving, (4) vanity, (5) "bibliographical," documenting a particular subject or concept (e.g., a sunrise), (6) supporting group or community activities, (7) supporting an individual activity (e.g., documenting a process of setting up a computer for later use), and (8) no particular motivation - the collection was a product of the sum of many individual one-time activities. Nov et al. (2010) found a positive relationship between the motivation of building reputation in the community and the amount of metainformation (i.e., tags) provided. Similarly, in an earlier study examining an online network of legal professionals Wasko and Faraj (2005) found a significant positive effect of motivation of building reputation on the quality and volume of knowledge contribution. In addition, longer tenure in the field was linked with higher volume of contributions. Lakhani and Wolf (2005) identified self-identification with a community and sense of obligation to contribute as a motivation to participate in F/OSS projects along with sense of enjoyment and receiving rewards as other motivations.

One of the indirect indicators of researcher reputation in scholarly communication is scholarly impact measured by the number of citations the researcher's works receive. Hence, participation in research identity databases could also be motivated by gaining higher citation counts on publications or other alternative metrics of impact. Moed (2007) suggested that providing an early view to a publication by posting its preprint in a preprint database like arXiv could result in a higher citation count.

There is a significant body of research on knowledge curation work organization in online communities, including activity systems, tools used, division of labor and roles played, norms, conventions, and policies.

Studies also examined the issues of career development, work coordination, conflict resolution, and recruitment. Arazy et al. (2015) examined functional roles played and the evolution of member career paths in Wikipedia. Crowston et al. (2013) discussed the growth dynamics and recruitment efficiency of Wikipedia communities. Resnick et al. (2012) used an analysis of relevant social sciences literature and cost-benefit analysis to propose 48 design claims to bootstrap a new online community, identify its niche, recruit members, achieve a critical mass, and retain and grow membership against competition. The claims range from content navigation and selection principles, using professionally created content and services provided by paid staff to attract initial members, to the principles of site presentation and framing. Since the claims are based on little empirical data, it would be valuable to investigate whether these design principles are valid for online research identity management systems and communities. Quality assessment, prediction, and intervention are essential aspects of knowledge curation. Stvilia et al. (2009) examined and contrasted quality assessment models of English, Arabic, and Korean Wikipedias and found the communities had different understandings of and models for article quality. In addition, it was found that in small peer-production communities errors may persist longer than in large communities (Tan et al., 2014). Finally, to encourage data use and researcher engagement in data curation, it is important is to convey quality and credibility through online RIM system and community design mechanisms (Choi & Stvilia, 2015; Schneiderman, 2000).

The online communities and data curation literatures provide valuable insights for designing research identity management systems and building and maintaining user communities around those systems. More empirical research, however, is needed to understand what motivates researchers to participate in online research identity management systems; what research identify information they consider important and are willing to share and maintain, and, alternatively, what discourages them from having an online research identity profile and/or engaging in its quality maintenance. Hence, this study will address the following research questions:

2. *Why do researchers participate or do not participate in online research identity management systems and related communities?*
    a. *How do researchers participate in online research identity management systems and communities?*
    b. *What are some of the perceived benefits or advantages associated with having an identity profile in a research identity management system?*
    c. *What are some of the perceived disadvantages or costs associated with having a research identity profile in a research identity management system?*

2. Impact

The outcomes of this exploratory research will include but not be limited to a qualitative theory of research identity data and information practices of researchers, quantitative model(s) of researchers' priorities for different online research identity data and services (e.g., data hosting, article recommendation, citation services, Q&A, potential collaborator and expert recommendations, impact metrics and altmetrics), the factors that may affect their participation in and commitment to online research identity management systems (e.g., seniority level, organizational and community norms and policies), and their motivations to engage in research identity data curation (e.g., to ensure the accurate and complete representations of their research profiles and data; to enhance their reputation). The study's findings can greatly enhance our knowledge of the design of research identity data/metadata models, services, quality assurance activities, and mechanisms for recruiting and engaging researchers for provision and maintenance of research identity data.

Design recommendations and researcher recruitment strategies based on this study can be adopted in diverse settings and can produce improved services for multiple stakeholders of research identity data such as researchers, librarians, university administrators, funding agencies, government, publishers, search engines, and the general public. Due to the widespread use of research identity data and services and the interdisciplinary nature of related research, the findings of this project will be relevant and generalizable to many online research data and knowledge curation communities outside research university libraries (e.g., citizen science communities, science and research Q&A communities on StackExchange), and to web RIMs with global scopes (e.g., ResearchGate and Academia.edu). Furthermore, the study will make novel theoretical contributions to the literatures of data curation and online communities and advance our understanding of online research identity information and data practices of researchers.

The outcomes of this project will also inform academic and continuing professional education curriculum development in digital data curation and online peer curation community design. To ensure as broader impact as possible, the results of the study and related datasets will be freely and openly disseminated through a project website, the Dryad data repository, Florida State University's (FSU) IR, conferences, blogs, and open access peer reviewed journals.

The success of this one-year planning research project will be determined by the reliability and usefulness of the project's target outcomes (i.e., the qualitative theory and quantitative models). The progress towards those targets will be measured by the response rate achieved on a survey the project will carry out, by the statistical measures of reliability and generalizability of the project's findings, the levels of attendance of conference panels and open discussions organized by the project staff, and the acceptance rates of peer reviewed presentations and publications resulted from the project.

3. Project Design

The theoretical framework used for this research will consist of Self Determination Theory (SDT; Ryan & Deci, 2000) and Diffusion of Innovation Theory (DIT; Rogers, 2002). SDT can provide insights on different types of motivations users might have when they use a particular RIM system, and/or participate in research identity data curation. According to SDT, there is a motivation continuum ranging ranging from amotivation (i.e., the state defined by complete lack of motivation and self-determination) to intrinsic motivation, which is inherently autonomous and self-determined due to the person finding the activity s/he performs interesting and/or pleasant (Ryan & Deci, 2000). Between those two extremes, SDT places external motivation, which is externally induced. The SDT model defines four increasingly autonomous types of extrinsic motivations ranging from the type of external motivation which is most controlled and contingent on reward and/or punishment, and the type of external motivation which is most autonomous, as the person internalizes and integrates the imposed behavior's goals and values as his or her own (Gagné & Deci, 2005). In addition, SDT postulates that the satisfaction of three basic needs – feeling competent, autonomous, and related to others in a social world – is essential for the maintenance of intrinsic motivation and support of external motivation (i.e., internalization of the goals and values of imposed activities). Consequently, some of the challenges that CRIS/RIM system designs may face are how to support those three needs (i.e., competence, autonomy, and relatedness) through design mechanisms and features.

DIT (Rogers, 2010) can provide insights into the adoption of information systems, including scholarly identity management systems. According to DIT, the rate and process of innovation diffusion can be affected by the characteristics of innovation, the communication channels used, time, and the social system. The characteristics

or properties of innovation are (a) relative advantage, (b) compatibility, (c) complexity, (d) trialability, and (e) observability. Thus, an information system with greater perceived advantage over the existing systems, higher compatibility with the existing needs, values and expectations of a targeted user group, lower complexity, higher availability of trial experimentation, and greater observability of results might receive a higher adoption rate. The other factors of the innovation diffusion process model, such as the degree of effectiveness of the communication channels used to spread information, and/or the effectiveness of exploiting the social system's structure (e.g., the use of opinion leaders to promote the system) may influence the success of adoption as well (Rogers, 2002). Hence, the theory suggests that libraries and other institutions might need to make research identity management systems less complex, more compatible with the existing knowledge organization practices and infrastructure while providing unique data, value added services and making the system use visible and triable to potential users.

The study will use a mixed-methods approach. The project's research activities will include literature analysis, in-depth semi-structured interviews, and a large scale survey. In particular, it will use the above two theories and literature analysis to guide the development of a semi-structured interview protocol, a survey instrument, and coding schemas for data analysis. In addition, the study will adapt and use in the survey instrument the existing scales of different types of motivations found in the literature (e.g., Nov et al., 2010; Oreg & Nov, 2008; Wasko & Faraj, 2005; Venkatesh, 2000).

The research population of this study is defined as employees and students of institutions classified as Research Universities (very high research activity, RU/VH) in the Carnegie Classification of Institutions of Higher Education[4] which, by the start of data collection, will have an institutional repository (IR). The institutions classified as RU/VH are comparable on their resources for research activities and overall sociotechnical contexts for digital scholarship curation. That would allow the researchers to reduce the possible effects of extraneous variables (e.g., disparity in resources and/or needs for research identity data management) on the study's outcomes. The researchers will update the sampling frame they used in the previous studies (Lee, 2015; Lee & Stvilia, 2012) for those universities to make it accurately reflect their current states with regard of having operational IRs. In addition to the requirement of their home institutions to have an IR, to be invited to participate in this study, candidates must have at least one peer-reviewed research publication by the time of data collection. Contact information of potential participants will be obtained from publications deposited in IRs and cross validated or disambiguated using universities' faculty directories and personal homepages. A stratified sampling approach will be used to allow as equal and diverse participation as possible from qualified universities.

The project will start with an analysis of the data models and services of three research identity management systems (ORCID, ResearchGate, and Google Scholar). The lists of research identity profile elements and services identified through this analysis will be used to develop a set of items for interview and survey protocol questions.

Next, to gain an initial understanding of researchers' perception of, participation in, and/or avoidance of research identity management systems and related contexts, the project staff will interview three researchers with an identity profile and without in each of the three databases (a total of 18 participants). The audio recordings of the interviews will be transcribed and content analyzed.

[4] http://www.niso.org/topics/t/altmetric l/altmetric

The study then will use the interview findings to expand and refine the set of interview questions and develop a survey instrument. A link to a finalized survey instrument will be emailed to a unique set of 100 researchers each week until 418 valid data cases are collected. Personalized email reminders will be used to increase the response rate of the survey. Based on the experience of a study of the data practices of a similar population that the PIs carried out in the past, the project expects a 20-25% response rate (Stvilia et al., 2015). Hence, the project expects to sample 1,600-2,000 potential participants. Before participating in an interview or completing an online survey, participants will be given a consent form approved by the Human Subjects Committee of FSU (HSC Number: 2015.16120). The form contains information about the project, including information about potential risks associated with participation in the data collection. Participants who complete an interview or a survey will receive a $30 Amazon gift card.

4. Diversity Plan

The project will ensure the broad diversity of participant samples within the limits of the target population definition specified in Section 3 and the amount of information available about the demographic characteristics of potential participants. The samples will be stratified by gender, race, seniority, institutional and disciplinary affiliations. The project staff itself is diverse by gender, ethnicity, institutional affiliation and seniority. In addition, when recruiting an undergraduate research assistant to work on the project, special consideration will be given to students from underrepresented groups. The School of Information at FSU prides itself with the large, highly diverse undergraduate and graduate programs in LIS, IT, and ICT. By working on this project the undergraduate student will gain invaluable practical experience of conducting research as a member of a diverse distributed research team, using different research methods and tools.

5. Project Resources: Personnel, Time, Budget

This one year planning grant project is a collaboration among the School of Information at FSU and the Graduate School of Library and Information Studies at Queens College at City University of New York (CUNY). The project will have three principal investigators (PIs). Besiki Stvilia, an Associate Professor in the School of Information at FSU, will serve as a Project PI and Director. He brings project leadership expertise and published research in the areas of online peer production communities, data curation, and data quality assurance. Dr. Stvilia will lead the overall effort to conduct the proposed research, and design and administer the survey. Shuheng Wu, an Assistant Professor at CUNY, will serve as a Co-PI and contribute expertise in data curation, research communities, and qualitative research methods. Dr. Wu will lead on conducting interviews and analyzing interview data. Dr. Dong Joon Lee, an Adjunct Instructor and Research Data Management Associate at FSU, will serve as a Co-PI. He brings expertise in data identifier schemas and data curation. Dr. Lee will lead the analysis of research identity data and service models. To foster the close collaboration needed, the project staff will hold online Skype meetings once every two weeks to report on and coordinate the research agenda and work.

The success of the project will be ensured by the staff's qualification, funding provided by this planning grant, cost share from FSU, and an OCLC/ALISE award the PIs received on January 7, 2016, and which will cover a part of the data collection and travel costs. The FSU share of the proposed budget for the planning project includes salaries, health insurance and fringe for Stvilia and Lee for 0.8 summer months (the total of [redacted]); travel money for two two-person trips and one one-person trip to three conferences in 2017 – 2018 (Open

Repositories, Research Data Access & Preservation Summit, and ALISE) for them to make it possible to disseminate the project results at conferences. We have budgeted for $1,800 per person per trip, with the total of $9,000 for three trips including $5,400 cost share from OCLC. A trip budget includes roundtrip airfare, hotel, meals, conference registration, and local transportation. One hourly undergraduate research assistant (RA) is budgeted for 7 hours a week, [redacted] per hour for fall and spring semesters to assist the PIs with data collection and documentation with the total of $2,744. One of the largest items of the budget is reimbursement of interview and survey participants. We have budgeted reimbursement of 436 participants at $30/participant rate with the total of $13,080, including $6,540 cost share from OCLC. In addition, the FSU share of the project budget includes the cost of two licenses of NVivo content analysis software ($1,300). The indirect costs at FSU are assessed at 52% and equal to $17,088.

The proposed budget also includes a subcontract to CUNY. The subcontract comprises 0.6 summer month salary, fringe, and one conference trip for Wu. The total of the subcontract budget, including the indirect costs assessed at 39%, is [redacted]. The total requested budget from IMLS for this planning project is $49,950. FSU provides cost share of the PI's 11% academic year. The cost share from OCLC is $14,984. The total of cost share is $40,119.

6. Communication Plan

Findings of the project will be distributed at three LIS conferences (Open Repositories 2017, Research Data Access & Preservation Summit 2017, and ALISE 2017) through poster and panel presentations, and open discussions. In addition, findings of the project and design recommendations based on those findings will be published in open access peer-reviewed journals (e.g., PLoS One, First Monday, Information Research), practitioner blogs and social media groups. The generated datasets will be anonymized and distributed freely together with the copies of related presentations and publications from the project's website, FSU's IR, and Dryad Digital Repository so that interested researchers and practitioners can replicate the study, evaluate the project's outcomes, and/or use them in the design of best practice guides and CRIS features.

7. Sustainability (Not Required for Planning Grants)

8. Future Research – Followup Proposal

This planning project will establish a ground for the second phase of the research which will include the design and testing of a best practice guide for jumpstarting and managing online research identity data curation communities. A follow-up, two year, full project proposal will be developed and submitted to IMLS in the 2016 – 2017 grant cycle. The follow-up project will take the outcomes of this planning project and translate them into design claims. The PIs will collaborate with the expertize management system Expertnet.org at FSU to implement and test those claims through controlled experiments and use. Results of those experiments will be encoded as a best practice guide, policy templates, and a training module. These reusable knowledge resources then will be widely distributed to librarians and IR managers across the country through a training workshop organized by the project, conference presentations, panels, open discussions, peer-reviewed publications, and the project's website, and help them to develop cost-effective solutions for research identity management in their libraries and IRs.

Schedule of Completion

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
A1	Analyze the literature and research identity models used by ORCID, ResearchGate, and Google Scholar											
A2	Design interview protocol											
A3		Recruit and interview 18 participants. Transcribe interview recordings										
A4				Analyze interview data, design and test survey protocol								
A5					Recruit and survey 418 participants							
A6								Analyze survey data				
A7											Write and submit a final project report	
A8				Disseminate project findings through conference presentations, journal papers, blogs and social media groups								
A9											Document and deposit data into Dryad	

Table Legend: M=Month; A=Activity

References

1. Ames, M., & Naaman, M. (2007). Why we tag: Motivations for annotation in mobile and online media. In B. Begole & S. Payne (Eds.), *Proceedings of the SIGCHI.* (pp. 971-980). New York, NY: ACM.
2. Arazy, O., Ortega, F., Nov, O., Yeo, L., & Balila, A. (2015). Functional roles and career paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15, pp. 1092-1105). New York, NY: ACM.
3. Atkins, D., Droegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., Messerschmitt, D., & Wright, M. H. (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Arlington, VA: National Science Foundation. Retrieved from http://www.nsf.gov/od/oci/reports/atkins.pdf
4. Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059-1078.
5. Choi, W., & Stvilia, B. (2015). Web credibility assessment: conceptualization, operationalization, variability, and models. *Journal of the Association for Information Science and Technology, 66*(12), *2399-2414.*
6. Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2006, April). Using intelligent task routing and contribution review to help communities build artifacts of lasting value. *In Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 1037-1046). New York, NY: ACM.
7. Crowston, K., Jullien, N., & Ortega, F. (2013). Sustainability of open collaborative communities: Analyzing recruitment efficiency. *Technology Innovation Management Review*, 20-26.
8. Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP-CoNLL* (Vol. 7, pp. 708-716).
9. Dempsey, L. (2014, October 26). Research information management systems - a new service category? [Web log post]. Retrieved from http://orweblog.oclc.org/archives/002218.html
10. Elmasri, R., & Navathe, S. (2000). *Fundamentals of database systems* (3rd ed.). Reading, MA: Addison-Wesley.
11. Gagné, M., & Deci, E. (2005). Self-determination theory and work motivation. *Journal of Organizational behavior*, 26(4), 331-362.
12. Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438*(7070), 900–901.
13. Hinnant, C., Stvilia, B., Wu, S., Worrall, A., Burnett, G., Burnett, K., Kazmer, M. M., & Marty, P. F. (2012). Author team diversity and the impact of scientific publications: Evidence from physics research at a national science lab. *Library & Information Science Research, 34(*4), 249-257.
14. Institute for Museum and Library Services (2015). *Requirements for projects that develop digital products*. Retrieved from http://www.imls.gov/applicants/projects_that_develop_digital_content.aspx
15. Lakhani, K., &Wolf, R. (2005). Why hackers do what they do. In J. Feller, B. Fitzgerald, S. Hissam, & K. Lakhani (Eds.), *Perspectives in Free and Open-Source Software* (pp. 3–22). Cambridge, MA: MIT Press.
16. Lee, D. J. (2015). *Research data curation practices in institutional repositories and data identifiers.* Thesis (Ph. D.) – Florida State University - Tallahassee, Florida.
17. Lee, D. J., & Stvilia, B. (2014). Data curation practices in institutional repositories: An exploratory study. *Proceedings of the American Society for Information Science and Technology, 51*(1), 1-4.

18. Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, *58*(13), 2047-2054.
19. National Endowment for the Humanities (2015). *Data management plans for NEH Office of Digital Humanities proposals and awards*. Retrieved from http://www.neh.gov/files/grants/data_management_plans_2015.pdf
20. National Institutes of Health. (2015). *NIH data sharing policy and implementation guidance (NIH Publication No. 03-05-2003).* Bethesda, MD: Author. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#goals
21. National Science Foundation. (2015). *Grant proposal guide (NSF Publication No. gpg11001).* Arlington, VA: Author. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpgprint.pdf
22. Niu, J. (2013). Evolving landscape in name authority control. *Cataloging & Classification Quarterly, 51*(4), 404-419.
23. Nov, O. (2007). What motivates Wikipedians. *Communications of the ACM, 50*(11), 60–64.
24. Nov, O., Naaman, M., & Ye, C. (2010). Analysis of participation in an online photo-sharing community: A multidimensional perspective. *Journal of American Society for Information Science and Technology, 61*(3), 555-566.
25. OCLC Research, Task Force on the Registering Researchers. (2014). *Report of the OCLC Research Task Force on the Registering Researchers.* Retrieved from http://www.oclc.org/research/themes/research-collections/registering-researchers.html
26. Oreg, S., & Nov, O. (2008). Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values. *Computers in Human Behavior, 24*, 2055–2073.
27. Palmer, D. (2013). The HKU Scholars Hub: Reputation, identity & impact management. How librarians are raising researchers' reputations (Asia-Pacific focus): an exploration of academic networks, profiles and analysis, Library Connect Webinar, 5 December 2013. http://hub.hku.hk/bitstream/10722/192927/1/Reputation.pdf.
28. Resnick, P., Konstan, J., Chen, Y., & Kraut, R. E. (2012). Starting new online communities. In *Building Successful Online Communities: Evidence-based Social Design*, 231-280.
29. Rogers, E. M. (2002). Diffusion of preventive innovations. *Addictive behaviors*, *27*(6), 989-993.
30. Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, *55*(1), 68.
31. Salo, D. (2009). Name authority control in institutional repositories. *Cataloging & Classification Quarterly*, *47*(3-4), 249-261.
32. Shneiderman, B. (2000). Designing trust into online experiences. *Commun. ACM*, *43*(12), 57-59.
33. Smalheiser, N., & Torvik, V. (2009). Author name disambiguation. *Annual Review of Information Science and Technology (ARIST), 43*, 1-43.
34. Steiner, T. (2014). Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A global study of edit activity on Wikipedia and Wikidata. In *Proceedings of The International Symposium on Open Collaboration* (OpenSym '14). New York, NY: ACM.
35. Stvilia, B. (2006). *Measuring information quality*. Thesis (Ph. D.) - University of Illinois at Urbana - Champaign, Urbana.

36. Stvilia, B., & Jörgensen, C. (2009). User-generated collection level metadata in an online photo-sharing system. *Library & Information Science Research, 31*(1), 54-65.
37. Stvilia, B., Gasser, L., Twidale M., B., & Smith L. C. (2007). A framework for information quality Assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720-1733.
38. Stvilia, B., Hinnant, C., Schindler, K., Worrall, A., Burnett, G., Burnett, K., Kazmer, M. M., & Marty, P. F. (2011). Composition of scientific teams and publication productivity at a national science lab. *Journal of the American Society for Information Science and Technology, 62*(2), 270-283.
39. Stvilia, B., Hinnant, C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M., & Marty, P. F. (2015). Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *Journal of the Association for Information Science and Technology*, *66*(2), 246-263.
40. Stvilia, B., Jörgensen, C., & Wu, S. (2012). Establishing the value of socially created metadata to image indexing. *Library & Information Science Research, 34*(2), 99-109.
41. Stvilia, B., Twidale, M., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983–1001.
42. Svenonius, E. (1989). Design of controlled vocabularies. *Encyclopedia of Library and Information Science*, *45*(suppl 10), 82-109.
43. Tan, C., Agichtein, E., Ipeirotis, P., & Gabrilovich, E. (2014). Trust, but verify: Predicting contribution quality for knowledge base construction and curation. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (WSDM '14, pp. 553-562). New York, NY: ACM.
44. Tate, D. (2012). Implementing a CRIS with PURE [PowerPoint slides]. Presented at *Institutional Repository Managers' Workshop (IRMW12)*, 15 Jun 2012, University of London. Retrieved from http://www.slideshare.net/ULCCEvents/implementing-a-cris-with-pure/11
45. Tenopir, C., Birch, B., & Allard, S. (2012). Academic libraries and research data services. *Association of College and Research Libraries*.
46. Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research, 11*, 342–365.
47. Wasko, M. M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, 35-57.
48. Wu, S., Stvilia, B., & Lee, D. J. (2012). Authority control for scientific data: The case of molecular biology. *Journal of Library Metadata*, *12*(2-3), 61-82.

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. Your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions
If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project
(Check all that apply):

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
☐	Digital content	Part II
☐	Software (systems, tools, apps, etc.)	Part III
X	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections

The study will generate interview and survey data, as well as publications. As dictated by Florida State University's Data Management Policy, the intellectual property to the data will remain with Florida State University, stewarded by the PIs. The datasets will be anonymized and available under a Creative Commons Attribution (BY) licenses.

.

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

Florida State University will own datasets collected by this study. The PIs will own stewardship on the datasets. The datasets will be anonymized and distributed openly with an "attribution only" license from the project's website and the Dryad Digital Repository. The datasets will be supplemented with rights metadata using Creative Commons Rights Expression Language (REL). The REL metadata will inform both human and automated agents (e.g., search engines and automated aggregators) about the copyright status and use conditions of the data.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

As with any study involving human subjects, this study too will have a risk of a possible inadvertent disclosure of private identifiable information that may damage participant's reputation. The study will employ thorough procedures to minimize this risk and protect the participant' confidentiality and anonymity at the extent allowed by law.
Publications about the findings from the study will mask the identity of the individual. Interviews will be tape recorded; transcripts will be prepared with names and any personal identifiers changed. Participants will have the right to have the tape turned off at any time during the interview. All intermediary data files will remain in the possession of the primary investigators and stored on a password protected server system run by the Academic and Research Technologies Office of the College of Communication and Information.

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and/or collect, the quantities of each type, and format you will use.

A.2 List the equipment, software, and supplies that you will use to create the content or the name of the service provider who will perform the work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant

information on the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

B. **Digital Workflow and Asset Maintenance/Preservation**

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: You may charge the Federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the Federal award. (See 2 CFR 200.461).

C. **Metadata**

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, or preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, or PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

D. **Access and Use**

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name and URL(s) (Uniform Resource Locator) for any examples of previous digital collections or content your organization has created.

Part III. Projects Creating Software (systems, tools, apps, etc.)

A. **General Information**

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) this software will serve.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

B. **Technical Information**

B.1 List the programming languages, platforms, software, or other applications you will use to create your software (systems, tools, apps, etc.) and explain why you chose them.

B.2 Describe how the intended software will extend or interoperate with other existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software you will create.

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software.

B.5 Provide the name and URL(s) for examples of any previous software tools or systems your organization has created.

C. **Access** and **Use**

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will be publicly depositing source code for the software developed:

Name of publicly accessible source code repository:
URL:

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

To identify researchers' motivations to contribute to and participate in research identity data curation in research information management systems the study will interview 18 and survey 418 researchers. Interview data will be collected from July 2016 to August 2016. Survey data will be collected from October 2016 to December 2016.

2. Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

The project has an IRB approval for the proposed data collection activities from the FSU's Human Subjects Committee (FSU; HSC Number: 2015.16120; See Supporting Documents for the IRB approval).

3. Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

The only risk associated with participation in this study is a possible inadvertent disclosure of private identifiable information that may damage participant's reputation. The study will employ thorough procedures to minimize this risk and protect the participant' confidentiality and anonymity at the extent allowed by law.
Publications about the findings from the study will mask the identity of the individual. Interviews will be tape recorded; transcripts will be prepared with names and any personal identifiers changed. Participants will have the right to have the tape turned off at any time during the interview. All intermediary data files will remain in the possession of the primary investigators and stored on a password protected server system run by the Academic and Research Technologies Office of the College of Communication and Information.

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

Each participant will sign a consent form approved by the FSU's Human Subjects Committee. Each participant will be assigned a numeric identifier by the researchers. The identifier then will be used to reference data objects related to that participant. Both the digital copies of signed consent forms, and the name to identifier mappings will be encrypted and stored on a password protected server system of the College of Communication and Information. Only the PIs will have access to those files.

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Individual interview data will consist of audio recordings, interview transcripts as ASCII text files, and coded interview transcripts stored as NVivo files. After the transcription process is completed, the audio recordings of the interviews will be disposed. Only the coded transcripts of interviews will be retained for analysis. Survey data will be stored in the CSV ("Comma Separated Values") file format.
The PIs will clean, anonymize and document the data, and assemble archival information packages (AIPs) for ingestion into the Dryad repository.

6. What documentation (e.g., data documentation, codebooks, etc.) will you capture or create along with the dataset(s)? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

The study will use the Data Documentation Initiative (DDI) and MOD metadata schemas to document data files. Metadata files in the RDF Turtle and XML text formats will be deposited together with data files and linked to those data files through persistent identifiers.

7. What is the plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Data files together with their documentation files will be deposited in the Dryad repository for long term preservation and access. The copies of derived publications will be deposited into FSU's IR and linked to the data files through persistent identifiers.

8. Identify where you will be publicly depositing dataset(s):

Name of repository: Dryad Digital Repository
URL: http://datadryad.org/

9. When and how frequently will you review this data management plan? How will the implementation be monitored?

The proposed data management plan will be reviewed in December 2016 when the proposed data collection activities will be completed, and again in April 2017 when the project team will start documenting datasets for long term preservation and access. The implementation of the data management plan will be monitored by the Project Director, Besiki Stvilia, who will be responsible for this planning grant project as a whole.

Original Preliminary Proposal

Towards engaging researchers in research identity data curation

This collaborative planning project, addressing the IMLS priority of establishing a shared national digital platform, will explore researcher participation in research identity management systems. In particular, it will examine researchers' perceived value of research identity metadata, motivations to participate in and commit to online research identity management systems, and contribute to research identity data curation. Accurate research identity determination and disambiguation are essential for effective grouping, linking, aggregation, and retrieval of digital scholarship; evaluation of the research productivity and impact of individuals, groups, and institutions; and identification of expertise and skills. There are many different research identity management systems from publishers, libraries, universities, search engines and content aggregators with different data models, coverage, and quality. Although knowledge curation by professionals usually produces the highest quality results, it may not be scalable because of its high cost. The online communities' literature shows that successful peer curation communities which are able to attract and retain enough participants can provide scalable knowledge curation solutions of a quality that is comparable to the quality of professionally curated content. Hence, the success of online research identity management systems may depend on the number of contributors and users they are able to recruit, motivate, and engage in research identity data curation.

The National Digital Platform proposed by the IMLS community, in addition to shared content, software, and hardware modules, may need to provide shared research based knowledge for effective design, configuration, and management of those resources. In particular, a shared design knowledge base is necessary to design effective access to library resources, recruit users, build communities around those resources, and engage them in library events and activities, which may include the curation of digital research identity and authority data. Although there is a significant body of literature on authority control in library databases, automated entity extraction, determination and disambiguation on the Web, and the design and management of online peer-production communities, there is still a dearth of research on researcher participation in and commitment to online research identity management information systems and communities. This project will address that need.

Research questions and study design

The proposed research consists of two phases. The scope of this one year planning project proposal is limited to the first, exploratory phase of the research. In particular, the planning project will explore the following research questions: *(a) Why and how do researchers use an online research identity management system(s)? (b)Why do researchers participate or do not participate in online research identity management systems and related communities?*

The planning project will start with an analysis of the data models and services of three research identity management systems (ORCID, ResearchGate and Google Scholar). The lists of research identity profile elements and services identified through this analysis will be used to develop a set of items for interview and survey protocol questions. Next, to gain an initial understanding of researchers' perceptions, participation in and/or avoidance of research identity management systems and related contexts, the project staff will interview 3 researchers with and without an identity profile in each of the three databases (a total of 18 participants). Convenience sampling will be used. The audio recordings of the interviews will be transcribed and content analyzed. The study then will use interview findings to expand and refine the set of interview questions and develop a survey instrument. 200 researchers will be surveyed. Before participating in an interview or completing an online survey, participants will be given a consent form approved by the Human Subjects Committee of Florida State University (FSU; HSC Number: 2015.16120). The form contains information about the project, including information about potential risks associated with participation in the data collection. Participants who complete an interview or a survey will receive a $30 Amazon gift card.

The outcomes of the exploratory phase of the research funded by the planning grant will include but not be limited to *a qualitative theory and quantitative models of researcher motivations and/or amotivations to participate in and commit to online research identity management systems and research identity data curation*.

Project staff

Besiki Stvilia, an Associate Professor in the School of Information at FSU will serve as a Project PI and Director. He brings project leadership expertise and published research in the areas of online peer production communities, data curation, and data quality assurance. Dr. Stvilia will lead the overall effort to conduct the proposed research, and design and administer the survey. Shuheng Wu, an Assistant Professor at the City University of New York (CUNY) will serve as a Co-PI and contribute expertise in data curation, research communities, and qualitative methods. Dr. Wu will lead on conducting interviews and analyzing interview data. FSU Adjunct Instructor Dong Joon Lee, a Co-PI, brings experience and expertise in data identifier schemas and data curation. Dr. Lee will lead on the analysis of research identity data and service models.

Budget

The FSU share of the proposed budget for the planning project includes salaries, health insurance and fringe for Stvilia and Lee for 0.8 summer month (the total of [redacted]); travel money for two one person trips to domestic conferences in 2017 – 2018 (ALISE, Research Data Access & Preservation Summit, and/or iConference) for them to make it possible to present the project results at conferences. We have budgeted $1,800 per trip, with the total of $3,600 for two trips. A trip budget includes roundtrip airfare, hotel, meals, conference registration, and local transportation. One of the largest items of the budget is reimbursement of interview and survey participants. We have budgeted reimbursement of 218 participants at $30/participant rate with the total of $6,540. In addition, the FSU share of the project budget includes the cost of two licenses of NVivo content analysis software ($1,000). The indirect costs at FSU are assessed at 52% and equal to $17,084. The proposed budget also includes a subcontract to CUNY. The subcontract comprises 0.6 summer month salary, fringe, and one conference trip for Wu. The total of the subcontract budget, including the indirect costs assessed at 39%, is [redacted]. The total requested budget from IMLS for this planning project is $49,938. FSU provides cost share of the PI's 11% academic year which is equal to [redacted].

Evaluation and dissemination Plan

The success of this planning project will be defined by researcher and practitioner communities' evaluation and recognition of the importance and value of the results of the proposed research. Ultimately, the success of the project will be evaluated based on the reuse of the project's outcomes (i.e., methodology, data collection instruments, and findings) and measured by the number of peer-reviewed publications produced by the project and their impact. Findings of the project will be distributed at three LIS conferences (ALISE 2017, Research Data Access & Preservation Summit 2017, and iConference 2017) through poster and panel presentations. In addition, findings of the project and design recommendations based on those findings will be published in peer-reviewed journals. The generated datasets will be anonymized and distributed freely together with the preprints of related presentations and publications from the project's website and FSU's institutional repository, so that interested researchers and practitioners could replicate the study, evaluate the validity of the project's outcomes, and/or use them in the development of best practice guides and policies.

Follow-up future research

This planning project will establish a ground for the second phase of the research which will include the design and testing of a best practice guide for jumpstarting and managing online research identity data curation communities. A follow-up, two year, full project proposal will be developed and submitted to IMLS in the 2016 – 2017 grant cycle. The follow-up project will take the outcomes of this planning project and translate them into design claims. The PIs will collaborate with the expertize management system Expertnet.org and FSU's institutional repository (diginole.lib.fsu.edu) to implement and test those claims through controlled experiments and use. Results of those experiments will be encoded as a best practice guide, policy templates, and a training module. These reusable knowledge resources then will be widely distributed to librarians and IR managers across the country through a training workshop organized by the project, conference presentations and panels, peer-reviewed publications, and the project's website, and, help them to develop cost-effective solutions for research identity management in their libraries.