NARRATIVE

1. STATEMENT OF NEED

Library and information science (LIS) professionals have focused on two issues regarding scientific publications: one is to study scholarly communication and the scientific workforce using large publication data (i.e., ***knowledge discovery***) and the other is to build useful information retrieval systems to satisfy users' information needs (i.e., ***knowledge delivery***). Both threads of research have experienced significant transformation due to the variety, velocity, and volume of scientific data available. While there were efforts to leverage the opportunities presented by publication-based Big Data in multiple spheres, a clear gap exists in how to utilize unstructured or semi-structured data to identify context-rich entities — expressions in the text that convey some discriminatory information about the research-relevant aspects of the publication, such as research methods, theories, and concepts. Thus, we have limited understanding of how to use these entities to better satisfy users' information needs and to conduct richer analyses of the scientific system. Furthermore, the LIS community lacks integrated, unstructured-data compatible methods and tools to conduct their own analyses and learning. We propose building an entity-based research framework for these investigations and building bridges between the current and future communities of LIS. Specifically, this Early Career Development project will address the following needs.

The need to extend beyond metadata and provide context-driven data analytics for scientific publications. Publication data embody the very essence of humans' scientific and technological advances. These data have been continuously examined through multidisciplinary efforts. Citation-based indictors have traditionally been employed to assess research impact (e.g., Hirsch, 2005; Hirst, 1978), portray intellectual landscapes (e.g., White & Griffith, 1981; White & McCain, 1998), or map the evolution of science (e.g., Börner et al., 2010; Boyack, Klavans, & Börner, 2005; Chen, 2006). Modern statistical methods have employed various publication-based networks to cluster research specialties (e.g., Janssens, Glänzel, & De Moor, 2008; Liu et al. 2010; Waltman & Van Eck, 2012), detect author communities (e.g., Ahn, Bagrow, & Lehmann, 2010; Leskovec, Lang, Dasgupta, & Mahoney, 2008; Newman & Girvan, 2004; Wang, Lou, Tang, & Hopcroft, 2011), identify research topics (e.g., Blei, Ng, & Jordan, 2003; Blei & Lafferty, 2007; Ramage, Hall, Nallapati, & Manning, 2009), and discover key venues and contributors (e.g., Sun, Yu, & Han, 2009; Sun et al., 2011; Yan, Ding, & Sugimoto, 2011; Zhou, Orshanskiy, Zha, & Giles, 2007). These studies, however, were largely driven by the analysis of existing publication metadata (e.g., authors, journals, and references). Consequently, we have limited understanding of the ways to analyze the context of individual papers and how to use this context to better organize information resources and serve users' information needs. Without understanding the context, publication data analyses are largely confined to descriptive statistics or interpretations of small samples. As the variety of sources available for textual analyses is increasing, there is the need to incorporate context-driven data analytics into knowledge discovery and delivery. The enriched data sources will not only provide more efficient means of analyses, but also entail a paradigm shift in modes of inquiry. ***To this end, this project will design new methodologies to complement the metadata-driven mode of inquiry with a context-driven one for library research and services.***

The need to examine context-rich entities for scholarly communication and knowledge production. There is a rich body of literature that uses bibliographic data to discover knowledge (e.g., Cronin & Sugimoto, 2014; Borgman & Furner, 2002). While most of these studies employed research papers as the unit of analysis, knowledge is more effectively expressed through unstructured or semi-structured publication fields, such as titles, abstracts, keywords, or full-text. For instance, co-word analysis (e.g., Callon, Courtial, & Laville, 1991; Courtial, 1994; Ding, Chowdhury, & Foo, 2001) and the recent development of topic models (e.g., Blei, Ng, & Jordan, 2003; Blei & Lafferty, 2007; Ramage, Hall, Nallapati, & Manning, 2009) are representative text-based methods. In the same vein of research, the current project is motivated to tackle the complexity of unstructured data, with a particular focus on detecting context-rich entities from text. By achieving these, the project will reveal scholarly communication at a new contextualized level and will address questions on the provenance, diffusion, and

adoption of knowledge at a heretofore unexplored extent and depth. Conducting these entity-based knowledge discovery activities has the readily apparent advantage of gaining a concrete and fine-grained perception of how different scientific entities are embedded and relate to each other. It also helps us obtain an in-depth understanding of production and dissemination patterns of scientific knowledge, innovations, and influences. ***This project will therefore extend the unit of analysis to the entity level, examine the creation, dissemination, and provenance patterns of context-rich entities, and identify ways to enhance library services through these entities.***

The need to conduct context-aware retrieval and fine-grained analysis to enhance digital services in the linked data era. The composition of the research landscape is evolving—we have witnessed a growing trend in information retrieval and analysis from keyword-based match, to ranking enabled search, to classification enabled search, to personalized recommendations, and to semantics enabled search (e.g., Baeza-Yates & Ribeiro-Neto, 1999; Manning, Raghavan, & Schütze, 2008). The success of this transformation will require the use of more advanced tools to extract more granular and context-aware information from large linked data. In particular, entity-level analysis will enable us to merge heterogeneous data sources as well as to disambiguate and integrate context-rich entities across different corpora through linked entities. While modern natural language processing (NLP) techniques are available, systematic approaches are lacking on how to utilize these methods to understand the latent meanings of the text of scientific publications and how to use them to address users' information needs on digital services. It is imperative that analyses and tools are adapted and created to meet the needs of this evolving landscape. It is our goal to focus on these domains so the dynamics, interactivity, and innovation in these areas are highlighted. ***Thus, this project will help enhance users' information seeking experiences by delivering new practices of research that fit their information needs in the big data era.***

The need to design effective, unstructured-data compatible methods and tools for LIS in knowledge discovery and delivery. The biggest challenge of the Big Data era will be figuring out how to make sense of data and how to translate it to a broader audience (Horowitz, 2008). Accordingly, one key component of the project is the development of an entity-based research framework with analytical tools for data-driven communities in LIS. Given that knowledge creation is dynamic, dynamic methods are required to analyze and visualize the products of this process. These methods will take into account the complexity and scale of contemporary data and will be grounded in the epistemological assumptions of the data from which they are drawn. To this end, we will design, prototype, and build a variety of text-based analysis methods (e.g., entity recognition, content analysis, taxonomy induction, and relation extraction) and apply these to data diachronically in order to describe the trajectory of concepts and identify emergent themes. Moreover, the framework's focus on analyzing context-rich entities will complement the state-of-the-art cyberinfrastructure that focuses on network analysis of publications (Batagelj & Mrvar, 2013; Chen, 2006; Sci2 Team, 2009; Van Eck & Waltman, 2010). To sustain the project beyond the parameters of the grant, integrated tools will be designed to afford easy access and modularity. Researchers and practitioners in LIS can adopt these tools for their individual knowledge discovery and delivery tasks and in turn better satisfy users' information needs. ***By incorporating these dynamic analysis methods and tools into the entity-based research framework, the project will contribute to knowledge discovery and delivery and build the capacity to sustain data-driven communities in LIS.***

Audience and audience needs

Two groups of audiences can be identified for this project: the first involves users who have the need to ***retrieve scientific publications***. This includes students, scientists, scholars, practitioners, and reference librarians. This information seeking need can be personal: students and scientists need to find relevant literature to satisfy their own information needs. In the meantime, the need can also be professional: librarians need to find useful literature to satisfy their patrons' information needs. Currently, the way to satisfy both personal and professional information seeking needs is carried out by a keyword or controlled vocabulary search in academic databases or search engines. Such searches tend to bring generic results; this limitation has become aggravated in the big data era as the size of accessible scientific publications has

increased significantly (e.g., Fernández et al., 2011). Furthermore, the existing search configurations, for instance, are incapable of finding articles that applied research A to data set B and addressed the issue of C—because these context-rich entities A, B, and C are treated simply as text strings without considering their semantic types and relations. The proposed project is motivated to address this hindering limit by the design and triangulation of several text-based methods to extract, categorize, and characterize context-rich entities from the text. ***Findings of this project will therefore provide audiences with innovative ways to retrieve scientific publications and more effectively satisfy their personal or professional information seeking needs, particularly for STEM publication retrieval and learning.***

The second group involves users who have the need to ***analyze scientific publications***. This includes researchers and practitioners working on different spectra of science studies, such as science policies, scholarly communication, the scientific workforce, and knowledge production and innovation. The typical workflow for this group of audiences is to obtain publication data on certain research areas and then apply a variety of bibliometric, network, statistical, or qualitative methods to examine different aspects of science. Traditionally, a published paper has been employed as the unit of analysis. While this unit of analysis has delivered rich analyses of science, researchers and practitioners are demanding more fine-grained methods and tools to contextualize findings and make sense of bibliometric indicators and numbers. It is argued that science studies should not be self-serving or retrospective (Yan & Guns, 2014); instead, these studies should be about using data to create new knowledge and giving insights on the future trajectory of science. To accomplish this, we will conduct entity-level science studies and use context-rich entities to reveal the mechanisms of science at a new contextualized level. ***That is, the project will transform science studies to the entity-level and contribute methods and tools to conduct context-aware science studies.***

2. IMPACT

The project will contribute to LIS by addressing two fundamental questions inherent in this research area: a) how to discover latent knowledge from large linked data; and b) how to deliver knowledge more effectively to satisfy users' information needs. Indeed, there is a concerted effort to examine these two questions from multiple spheres (e.g., Cronin & Sugimoto, 2014; Sun & Han, 2013); the proposed entity-based research framework, however, will tackle them through a unique context-aware perspective. This distinctive perspective will be particularly useful to probe into the underlying meanings of various archival objects (e.g., journal publications, patents, and books) and will greatly enhance the rendering of these objects in different library services (e.g., references, cataloging, preservation, and collection management). New knowledge and expertise gained from this project will be transferrable to these library services and provide tailored context-driven analytics and services.

The project will also contribute to LIS research by introducing context-rich entities as a unit of analysis. The proposed entity-based research framework will provide a comprehensive solution to entity-based research. Additionally, these methods will be applied to a large publication corpus as a proof-of-concept of the framework design (see *Section 3 Project Design*). This corpus contains the state-of-the art research in science, technology, engineering, and mathematics (STEM) fields as well as in social sciences and humanities. Methods and tools to be developed as part of the framework will create a unique layer by connecting various scientific publications through contained concepts, theories, and methods. The framework will therefore make novel contributions by informing the understanding of the life-cycle of these entities (i.e., provenance, projectile, and popularity) and also provide opportunities to reexamine knowledge production and innovations as codified by the entities.

Finally, the project will spur interdisciplinary research on knowledge production and innovation. The proposed methods and tools will build capacity to engage a community of researchers and practitioners on studies of knowledge production and innovation. This research area is highly interdisciplinary and has a tradition of using individualized tools to conduct research. The proposed methods and tools will bring these scholars together and form collaborative workforces and interdisciplinary science teams to address questions in ways that were not attainable before. Furthermore, the entity-based research framework will

invite a community of researchers and practitioners with diverse backgrounds, computational skills, and research interests and will help build a strong and diverse community for the future. These efforts will **"[invest] in the nation's information infrastructure…to address the education and training needs of the professionals**" (IMLS, 2014).

3. PROJECT DESIGN

Project objectives. The proposed project addresses the identified needs in *Section 1.0* by the development of an entity-based research framework. The framework with analytical tools and methods will allow for the identification, categorization, disambiguation, and representation of context-rich entities from scientific publications. The design of the framework and methods will facilitate more granular examinations of information analysis and retrieval and engage communities in LIS through the use of entity-based methods and tools. Specifically, the objectives of this study are:

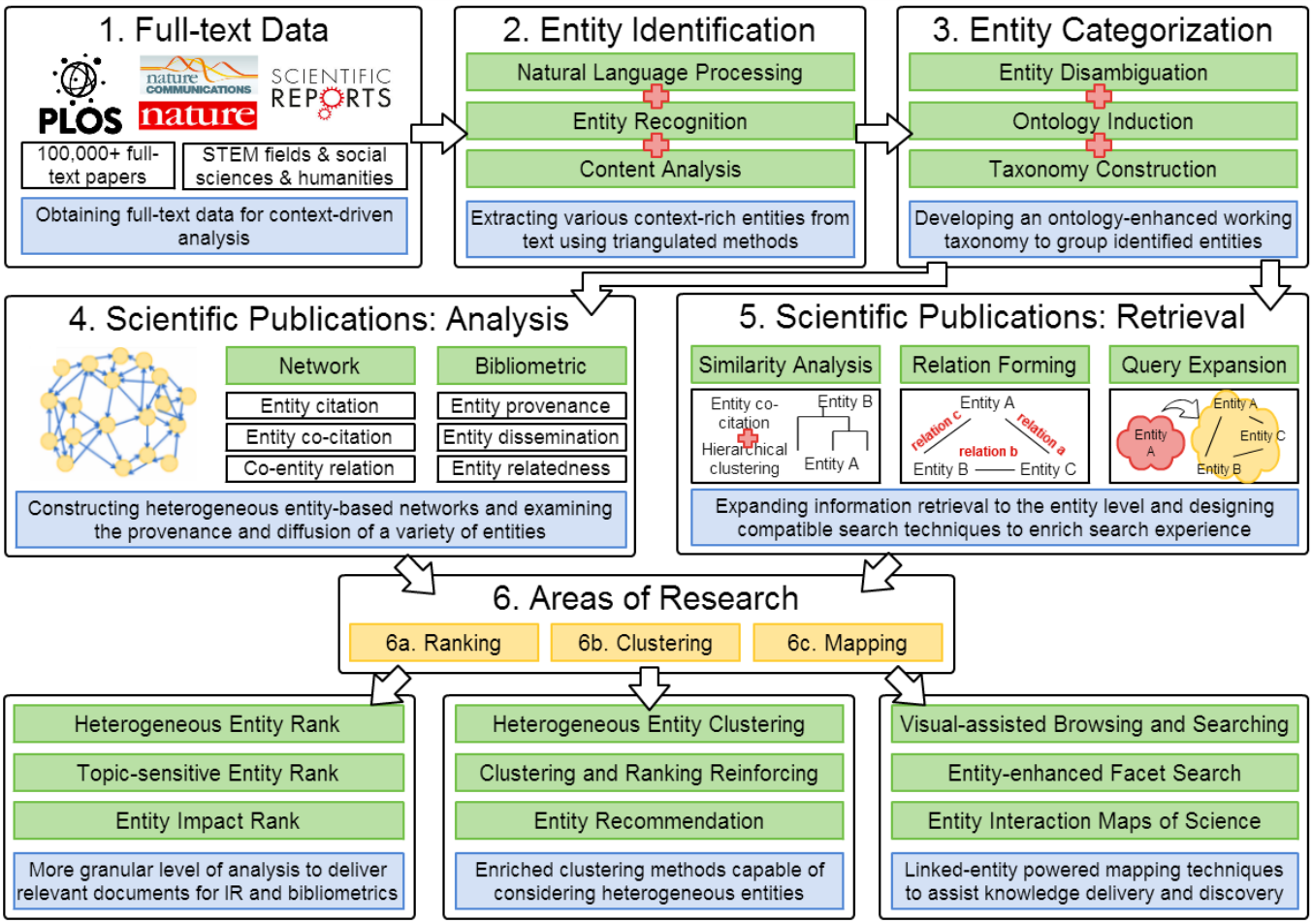
1. To develop context-driven data analytics for scientific publications and to complement the metadata-driven mode of inquiry with a context-driven one for library research and services;
2. To identify and organize context-rich entities in large scholarly corpora for more fine-grained examinations of scholarly communication and knowledge production;
3. To design methods that enable context-aware retrieval and analysis to better satisfy users' information and learning needs through linked entities; and
4. To sustain the research by building an entity-based research framework with analytical tools to facilitate knowledge discovery and delivery and by developing multimodal courses (i.e., face-to-face and online) on scholarly data analysis.

This project will allow us to further develop librarianship and digital services through the use of context-rich entities. The design of the framework and methods will demonstrate the needs for, offer analytical tools on, prove the concept of, and provide best practices for entity-based research. By achieving these, the project will build an open and systematic knowledge ecosystem and foster more effective use of linked publication data. The entity-based methods will provide a solid foundation for succeeding empirical information retrieval and analysis by researchers and practitioners in LIS.

Project activities. This project involves six main components (Figure 1): 1) full-text data preparation, 2) entity identification, 3) entity categorization, 4) scientific publications analysis, 5) scientific publications retrieval, and 6) three areas of research (6a ranking, 6b clustering, and 6c mapping). The analysis and retrieval of scientific publications are the two types of needs identified in the preceding paragraphs. They share similar applications in principal and can be synthesized into three major areas of research: the analysis and retrieval of scientific publications focus on the development of advanced methods to rank, cluster, and map documents to enhance users' information needs. The project will allow these components to work seamlessly and will also support the design of new modules.

3.1 Data. The dataset for this Early Career Development project contains more than 100,000 open access full-text papers published in four leading multidisciplinary journals: *PLOS ONE*, *Nature*, *Nature Communications*, and *Scientific Reports*. These papers represent the state-of-the-art research in STEM fields as well as in social sciences and humanities. This dataset follows the Journal Publishing Tag Set (JATS: http://jats.nlm.nih.gov/publishing). Thus, tools designed by this project can be extended to other JATS powered datasets. The access point for this data set is provided by the PubMed Central Open Access Subset (http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/) which is freely accessible to the public. The XML structured data will be segmented, annotated, and stored in a database. Digital data will be curated at the Drexel University E-repository and Archives (iDEA; http://idea.library.drexel.edu/), a distributed storage service that provides free access, browsing, searching, and downloading to the Drexel community as well as the general public. Because full-text data contain the highest degree of contexts, ***methods developed for the full-text data analysis will be transferrable to bibliographic data (e.g., title and abstract), citation context data, or segmented data (e.g., introduction, methods, and discussion sections)***.

Figure 1. Project design diagram



3.2 Entity identification. **Term identification** is the basis for the following natural language processing (NLP) procedures. It will be used to identify the most relevant terms in a given corpus. It can be achieved by first using a part-of-speech tagger to identify noun phrases and then determining the termhood (i.e., likelihood of being a relevant term) of each noun phrase (Kageura & Umino, 1996). The output of term identification will be fed into **named entity recognition (NER)**. NER provides a valuable solution to extract entities from large unstructured full-text data. Traditional NER methods utilized a variety of dictionaries to search for key terms (Levy & Andrew, 2006). This look-up method, however, may not be effective to extract newly named entities. Instead, we will employ the conditional random field (CRF; Lafferty, McCallum, & Pereira, 2001) to learn new named entities from text. We will utilize the gazette function in the CRF by adding Wikipedia terms to the gazette. This will help identify domain specific terms and phrases. Our preliminary results have shown that the inclusion yielded the best performance among all other information extraction methods. The CRF will be used jointly with ontology induction and taxonomy construction approaches (Poon & Domingos, 2010; Velardi, Faralli, & Navigli, 2013) to not just discern entities but also classify them into appropriate classes (see *Section 3.3 Entity categorization*). Lastly, **Content analysis** will be used to characterize genres and features within text. It will use features such as research fields and year of publication to cross-reference entities detected by entity recognition methods. Features will be delineated in codebooks in content analysis. The codebook will first be applied to a training dataset; the accuracy of the coding results will be assessed. When a high-level of reliability is achieved, the enhanced codebook will be applied to the whole dataset in an automated way. The triangulation of these three methods will provide a high-level of validity and allow us to identify context-rich entities from text.

3.3 Entity categorization. Once entities are identified from the above methods, the next step will involve **entity disambiguation**. Lexical variations of the same concept are not uncommon in natural languages. Thus, entity disambiguation will be necessary to integrate these lexical variations to produce reliable results. Previous efforts in this direction have largely used manual construction of parsing rules or labor-intensive annotations of exemplar sentences. These efforts, however, do not scale with the volume and the variability of scientific publications. In lieu of this, we will adopt an unsupervised semantic parsing technique (Poon & Domingos, 2010) to disambiguate and merge entities that represent the same concept. Additionally, we will also compose relations to place entities into a hierarchical structure; that is, we will **construct taxonomies** to represent entities in hierarchical structures. We will consider not only the relatedness of entities but also tree-structured clusters of them (Kozareva & Hovy, 2010). Entities taxonomies will be built through Probase (http://research.microsoft.com/en-us/projects/probase/), an evidence-predicated knowledge base that contains more than 2.6 million concepts and their associated evidences, predicated on both syntactic (e.g., the "such as" syntax) and semantic relations on the Web. These concepts will then be used to **induct ontologies** through Markov logic (Kiddon & Domingos, 2010; Poon & Domingos, 2010). This unsupervised approach will guarantee the efficiency of the inference and the learning of the hierarchical structure of entities. These processes (entity disambiguation, taxonomy construction, and ontology induction) are essentially equivalent to the construction of controlled vocabularies, but the difference is that we are creating a robust vocabulary from the target documents. The derived vocabularies will provide an accurate and extensive coverage to highly-focused domains, new domains, or fast growing domains where existing controlled vocabularies may lack the granularity or coverage.

3.4 Scientific publications: analytics. Using the identified and categorized entities, we will construct several heterogeneous entity networks, including **entity citation networks**, **entity co-citation networks**, and **co-entity networks**. **Bipartite networks** will also be formed when aggregating entities with authors and institutions to investigate invisible colleges and schools of thoughts (Yan & Sugimoto, 2010). Previous effort on scholarly network analysis largely focused on paper-, author-, and journal-level networks (Yan & Ding, 2012); these studies, however, did not provide enough granularities on the interactions of various context-rich entities. The construction and analysis of the novel entity-level networks will enhance our understanding of the **provenance**, **dissemination**, and **adoption** of research concepts, methods, and theories as signified by entities. For instance, not only will we examine how knowledge diffuses through papers, we will be able to examine how research concepts, such as digital curation, have emerged, grown, and been reshaped over time through entity-level network analysis.

3.5 Scientific publications: retrieval. Results from entity identification and categorization will also guide the retrieval of scientific publications. Previous retrieval techniques were mainly driven by keyword matches; recent techniques have employed linked entities to expand the search to terms that are semantically related to the query terms (Fernández et al., 2011; Manning, Raghavan, & Schütze, 2008). However, these advances are largely confined to biomedical areas where semantic relations have been formally expressed (i.e., through Unified Medical Language System, UMLS), but not readily to other areas where formal language systems are lacking. The taxonomy construction and ontology induction introduced in *Section 3.3* will allow us to create a working ontology to represent entities and their relations for any domain. Through hierarchical clustering (e.g., Steinbach, Karypis, & Kumar, 2000; Yan, Ding, & Jacob, 2011), we will measure the **similarity** of entities belonging to the same class or possessing upper- or lower-level relations (i.e., related terms, broader terms, and narrower terms) in entity co-citation networks. Results will be used in **query expansion** to recommend terms that are most similar to the queried ones—to enrich users' information seeking experience. In addition to query expansion, we will also **form relations** using bibliographic information, for instance, to link entities with associated papers, journals, authors, or institutions through adjacency matrices, so that when users search for these entities, recommendations will be made on which bibliographic sources to further search.

3.6 Areas of research. Recent studies have demonstrated the benefit of applying bibliometric analysis methods to enhance retrieval results (e.g., Mayr et al., 2014) as well as applying information retrieval

techniques to inform the understanding of bibliometric outputs (e.g., Yan & Guns, 2014). In this project, we will further identify ways to integrate the two primary specialties in LIS using tools from **ranking**, **clustering**, and **mapping** research.

3.6a Ranking. Ranking is a key task in both bibliometrics and information retrieval: in bibliometrics, scholars and science policy makers have used a variety of indicators (Waltman, Yan, & Van Eck, 2011) to evaluate the impact of authors, journals, and institutions; in information retrieval, because users primarily focus on top retrieved items, scholars have used normalized discounted cumulative gain (nDCG) to give higher weight to top ranked items (e.g., Yan & Guns, 2014). Only one type of objects is considered (e.g., paper, author, and journal) when ranking these objects. Thus, little attention was given to the context of such rankings. The context, however, is particularly valuable to inform the ranking results: for instance, we are more interested in knowing the most relevant papers on certain venues or the most representative authors on certain research topics. In light of this, we will utilize bibliographic data to form a variety of meta paths (Sun et al., 2011), such as entity-author, author-journal, author-paper, and entity-author relations. These meta paths will then be employed to conduct **heterogeneous entity ranking**. Additionally, entities will be assigned into topics based on the results from topic modeling techniques to produce more granular evaluation of entities through **topic-sensitive rank** (e.g., Yan, 2014; Yan, in press). Finally, temporal analysis will be used to highlight the **impact of entities over time** through accumulation of the number of citations in entity citation networks.

3.6b Clustering. Clustering is another fundamental area of study in bibliometrics and information retrieval. In bibliometrics, a variety of clustering techniques have been applied to portray the academic landscape and examine it's disciplinarity and interdisciplinary (e.g., Leydesdorff, Carley, & Rafols, 2013; Yan, Ding, & Jacob, 2011); in information retrieval, clustering techniques have been used to recommend related query terms in the same word clusters (e.g., Jain, 2010). These studies typically rely on co-occurrence relations, without considering the heterogeneity of the clustering objects or their diverse relations. To address these issues, we will utilize bipartite clustering (Dhillon, 2001) and star-relation clustering (Sun et al., 2013) to develop **clustering techniques compatible with heterogeneous entity types and relations**. Combining with heterogeneous entity ranking methods, **clustering and ranking will be consolidated** to bring more accurate results for both clustering and ranking. Results will also be used for **entity recommendation**, for instance, to recommend books or documents not just of the same authors but also authors who may be specialized in the same genre based on the heterogeneous clustering.

3.6c Mapping. Clustering results can also be presented in visualizations to communicate results with broader audiences. For bibliometrics, this involves the visualization of the backbone of science through paper or journal co-occurrence networks (e.g., Leydesdorff, Carley, & Rafols, 2013), knowledge flow maps through citation networks (e.g., Yan, Ding, Cronin, & Leydesdorff, 2013; Yan, in press), or author communities through coauthorship networks (e.g., Yan, Ding, & Zhu, 2010). For information retrieval, this involves the design of visual-assisted browsing and searching interfaces. Previous efforts on the design and implementation of these visualizations have communicated the results to diverse stakeholders (e.g., Boyack, Klavans, & Börner, 2005). The use of the document as the unit of analysis, however, was not able to illustrate the dynamics and interactivity of the embedded context. In this project, we will design visualization techniques that allow for the **entity-enhanced facet search** to expand the facet search beyond metadata and provide context-rich, ontology-driven entities for objects under search. This will complement the state-of-the-art facet search for biomedical entities (e.g., gene, compound, drug, and chemical) and promote the **visual-assisted browsing and searching techniques** to other domains. It is our goal to design streamlined methods to detect, disambiguate, categorize, and visualize field-specific entities for different research domains. Moreover, the entity-based framework will also deliver **entity interaction maps of science** to illustrate the dissemination and adoption of entities across disciplines.

3.7 Evaluation

The project will be evaluated in five ways. First, the intellectual results will be assessed by peer reviews via submission of articles to high-impact journals, conferences, and panels. Second, retrieval-based results will be evaluated through appropriate measures, such as precision, recall, F1 score, Receiver operating

characteristic (ROC) curve, and normalized discounted cumulative gain (nDCG) (e.g., Yan & Guns, 2014). Third, bibliometric-based results will be assessed through comparisons with established science indicators, including the *h*-index (Hirsch, 2005), Crown Indicator (Waltman et al., 2011), Eigenfactor (Bergstrom, West, & Wiseman, 2008), SCImago Journal Ranks (Gómez-Núñez et al., 2014), and Source Normalized Impact per Paper (SNIP; Waltman & Van Eck, 2013). Fourth, senior field members (i.e., journal editorial members and conference program committee members) will be invited by email to provide feedback on the development of the methods. This feedback will be iteratively integrated into the design process. Finally, semi-annual formative-assessment reports will be included to document: (a) project progress, (b) improvements in project outcomes, and (c) plans for the next half year. These semi-annual reports will be formally reviewed by the project consultants Dr. Sugimoto of Indiana University and Dr. Zhai of University of Illinois (see commitment letters: *Letter_Sugimoto.pdf* and *Letter_Zhai.pdf*).

4. PROJECT RESOURCES: PERSONNEL, TIME, BUDGET

Personnel. The PI will lead the team as an assistant professor at the College of Computing and Informatics (CCI). Dr. Yan will oversee, participate, and coordinate all project-related activities. He will bring direct expertise on impact assessment, network analysis, and knowledge diffusion. He has contributed more than a dozen publications in these areas which will lay a sound groundwork for the proposed project (see *Resumes.pdf*). Two undergraduates (15 hours/week, academic year) and one doctoral student (20 hours/week, academic year) will also join the team as research assistants—each with varied backgrounds and skills. The project will therefore engage an integrative group of researchers. To ensure smooth project execution, two domain experts Drs. Sugimoto and Zhai will join the project as consultants. Both consultants have rich experience in scientific data analysis and management and have successfully directed funded projects. Their involvement will support the delivery and validity of the project outcomes.

Work plan. This Early Career Development project is scheduled to take place over three years starting on 5/1/2015 (see the Gantt chart in *Scheduleofcompletion.pdf*). The plan below illustrates the implementation for each year of the project with respect to project goals.

- **Year 1**. **Data**: download data from PubMed Central through their publicly available portal; process the XML-formatted data for use in identification (component 1 in Figure 1). **Methods**: design, prototype, and build entity identification methods for scientific publications; triangulate and evaluate natural language processing, entity recognition, and content analysis methods (components 2 and 3 in Figure 1). **Deliverables**: methods in identifying entities in scientific publications; identified context-rich entities in STEM fields, socials sciences, and humanities. **Communication and dissemination**: entity identification methods will be disseminated during workshops and conferences; diachronical and disciplinary patterns will be revealed through journals publications. [**Objectives 1 and 2**]
- **Year 2**. **Data**: connect with the external knowledge base Probase to extract class information for entities; curate the obtained taxonomy in Protégé. **Methods**: design, prototype, and build entity categorization methods; develop taxonomy and ontology induction methods through identified entities in Year 1 and external data sources such as Probase (components 4 and 5 in Figure 1). **Deliverables**: taxonomy and ontology induction methods for scientific publications; a working taxonomy and ontology to organize the identified entities. **Communication and dissemination**: the taxonomy and ontology and their interactive visualizations will be presented in workshops and conferences; they will also be presented in the project website. [**Objectives 2 and 3**]
- **Year 3**. **Data**: apply and evaluate the methods in the three areas to the XML-formatted data as a proof-of-concept for the entity-driven research framework. **Methods**: examine the three areas of research using identified and categorized entities from Year 1 and Year 2; design, prototype, and build ranking, clustering, and mapping applications (component 6 in Figure 1). **Deliverables**: A context-driven research framework for publication analysis and retrieval with novel methods in ranking, clustering, and mapping. Design of undergraduate- and graduate-level courses on scholarly data analysis with a focus on information extraction and network analysis in both face-to-face and

online learning environments (see *Section 7 Sustainability*). **Communication and dissemination**: findings in the three areas of research will be disseminated in conferences and journal publications; reports and tutorials will be included in the project website. [**Objective 4**]

Budget. Because the project will produce large quantities of data materials, two computers—one for the PI and one for the students to share—are requested to store and process data. Travel funds are requested to support the PI and students to participate and disseminate project outcomes at core LIS conferences listed in *Section 6 Communication Plan*. In total, an amount of $247,713 in IMLS funding is requested to support the successful implementation of the project. Detailed budgets and a budget justification are listed in the attached documents (*Budget.pdf* and *Budgetjustification.pdf*). Drexel University and its College of Computing and Informatics will provide cost-sharing contributions in the amount of $147,585. The contributions include the PI's salary ($30,050), associated benefits ($9,917), indirect cost ($22,581), and RA's tuition ($85,037).

5. DIVERSITY PLAN

Products of this project will serve diverse communities of researchers and practitioners dedicated to promoting LIS research and services—this expansive cast would include librarians, information scientists and professionals, science policy makers, educators at various levels of student development, and system evaluators. These diverse communities are increasingly demanding methods and tools to reveal latent meaning from unstructured textual data and to enhance digital services. The project will address these needs by the development of methods and tools compatible with fine-grained analyses of scientific publications at the entity-level. Furthermore, the application of the methods and tools by these diverse communities of researchers and practitioners will benefit the digital services and learning of library patrons and students by effectively improving access to and retrieval of scientific publications. For those without advanced information literacy or prior knowledge of information retrieval, the entity-driven design will have a catalyst effect—reducing the knowledge barriers, spurring them on to greater explorations of scientific literature, and thus satisfying their information needs. For instance, for library patrons and students who previously had been overwhelmed by the complex retrieval interface or the sheer volume of unorganized retrieval results (Battleson, Booth, & Weintrop, 2001; Mahmood & Richardson, 2011), the intuitive design and interactive implementation of these approaches (e.g., through visual-assisted deliveries) will help knowledge reach these underserved groups. Thus, the project will serve the needs of the diverse communities of "**shared experiences, world views, and ways of learning**" (IMLS, 2014).

6. COMMUNICATION PLAN

Dissemination and communication. The deliverables of the project will include methods in identifying and categorizing entities in scientific publications, a working taxonomy and ontology in organizing the identified entities, and a context-driven research framework for publication analysis and retrieval. They will be disseminated throughout the project lifecycle in multiple forms, the most dominant of which will be journal articles; additionally, conference papers, panels, and posters will be produced. We will seek generalist journals (such as *PLOS ONE*) as well as those in our respective disciplines (e.g., *Journal of the Association for Information Science and Technology*, *Journal of Informetrics*, and *Library and Information Science Research*). We will also seek a range of conference venues for feedback of the project outcomes, including, but not limited to *Association for Information Science and Technology (ASIST) meetings, Joint Conference on Digital Library (JCDL)*, the *Association for Library and Information Science Education (ALISE)* conference, the *International Society for Scientometrics and Informetrics (ISSI)* meetings, and *Science, Technology, and Indicators (STI)* conference (please see *Scheduleofcompletion.pdf*). The development of methods and tools is central to this proposal; therefore, the designs will be advertised and demonstrated at conference workshops and webinars, acting as another form of dissemination to reach the target audiences.

Interdisciplinarity. This team (PI, students, and consultants) fulfills the IMLS Laura Bush 21st Century Librarian Program to "**[develop] information professionals who can help manage the burgeoning data**

generated by the nation's researchers" (IMLS, 2014) with training and expertise across multiple disciplines including information science, library science, computational linguistics, statistics, science policy, and scholarly communication. Training across the disciplinary spectrum is necessary for designing methods of analyzing and understanding the project's heterogeneous scholarly corpora.

Facilitating students' involvement and collaboration. Frequent communication is critical for this project and will be achieved through in-person meetings and mentoring. The funding of this proposal will strengthen and extend the mentor-student collaboration beyond the grant into further dimensions of inquiry and exploration. The PI has a history of distributing dominant authorship positions to collaborators and students; therefore, we anticipate easy allocation of credit for authorship and other contributions.

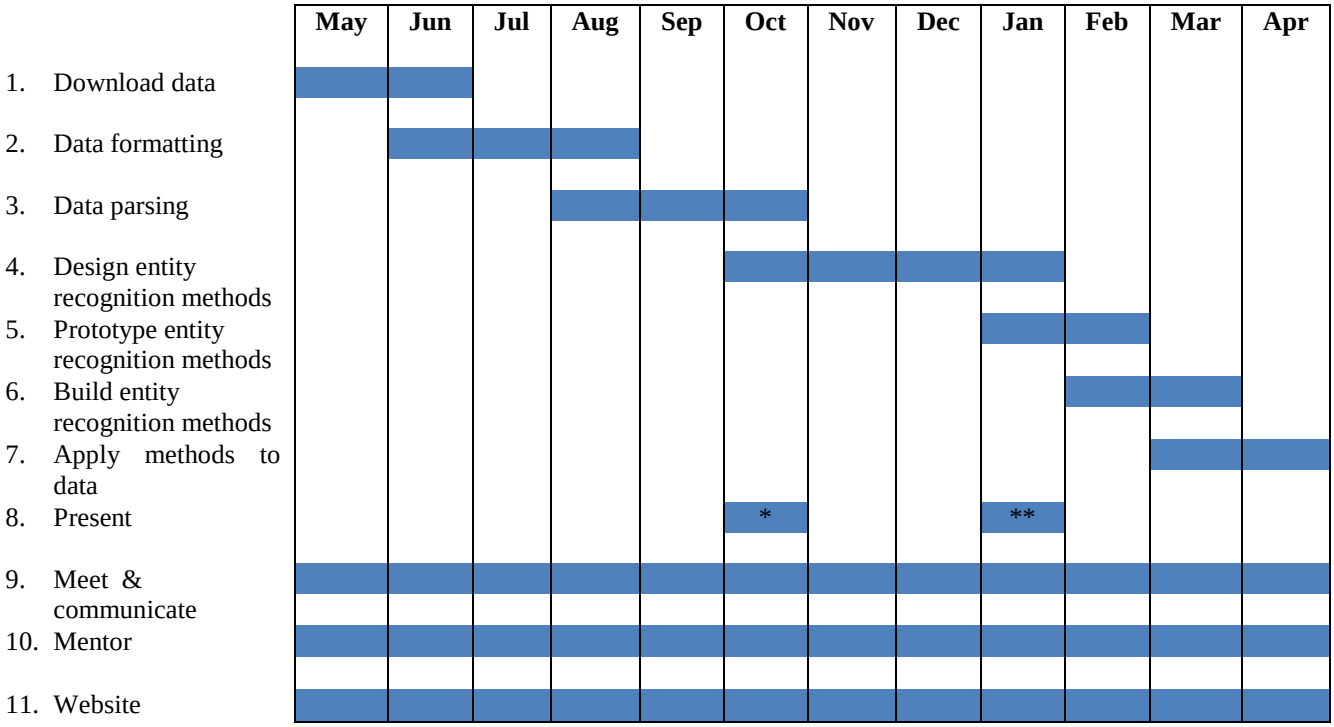
7. SUSTAINABILITY PLAN

Framework and methods building. To ensure the sustainability of the project beyond the performance of the grant, we will introduce our community to the methods, tools, visualizations, and practices of the research. From a cyberinfrastructure standpoint, the proposed framework will afford all essential components of entity-based research, including the identification and categorization of entities, entity taxonomy induction, and applications of methods to publication analysis and retrieval. Different components will be designed as open source tools under the guideline of OSGi Alliance (also known as the Open Services Gateway initiative) to support the easy plug-in of each component and smooth coordination among different ones. Such a design will also guarantee the extensibility of the framework to connect with new components and modules. Content-wise, the project will serve to prove the need of conducting entity-based research in LIS, enhance our understanding on the use of entity-based methods for knowledge discovery and delivery, and contribute to ongoing endeavors on science studies and information retrieval. Findings and products from this project will lay a solid foundation for empirical science studies and implementation of retrieval systems by researchers and practitioners in advancing a sustainable future.

Education and learning. The educational component of this proposal will be achieved through a combination of curriculum design, workshops, undergraduate and graduate student mentoring, and students' independent studies. By involving undergraduate and graduate students in grant-funded publications, the project will also train and educate the next generation of LIS professionals. In addition, undergraduate- and graduate-level courses will be designed on scholarly data analysis. These two courses will be delivered in multimodal learning settings (i.e., face-to-face and online) and will target students with little or no background in computing and statistics, including those from LIS, sociology, and arts & humanities. The course experience will equip them with the necessary skills to conduct data analytics in their future studies and career in the era of the STEM learning and the Big Data. Courses will be initially delivered to Drexel students and will then be delivered and shared in Drexel sponsored free Massive Open Online Courses (MOOCs: https://drexel.coursesites.com/). This platform enables course developers to publish and share courses as Open Educational Resources (OER) under a Creative Commons Attribution license (CC BY) (CourseSites, n.d.). The free MOOCs will be actively maintained during and after the project; licenses can be transferred to other instructors under CC BY when necessary for maximum dissemination and sustainability. Additionally, workshops and webinars will be planned at major LIS conferences (as described in *Section 7.0 Communication Plan*). These efforts will help engage an interdisciplinary community of librarians, information scientists, linguists, network scientists, and scientometricians and help form science teams to tackle some of the fundamental societal issues on knowledge discovery and delivery.

Sharing. The investment in human, teaching, and technical capacities will sustain the research and ethos of the proposed project beyond the confines of the funding period. In this spirit of openness, products from the grant will be made freely available to the public throughout the duration of the grant via a project website. In furthering transparency, results will be promoted using established scholarly communication channels and popular social networking tools to allow for the widest dissemination.

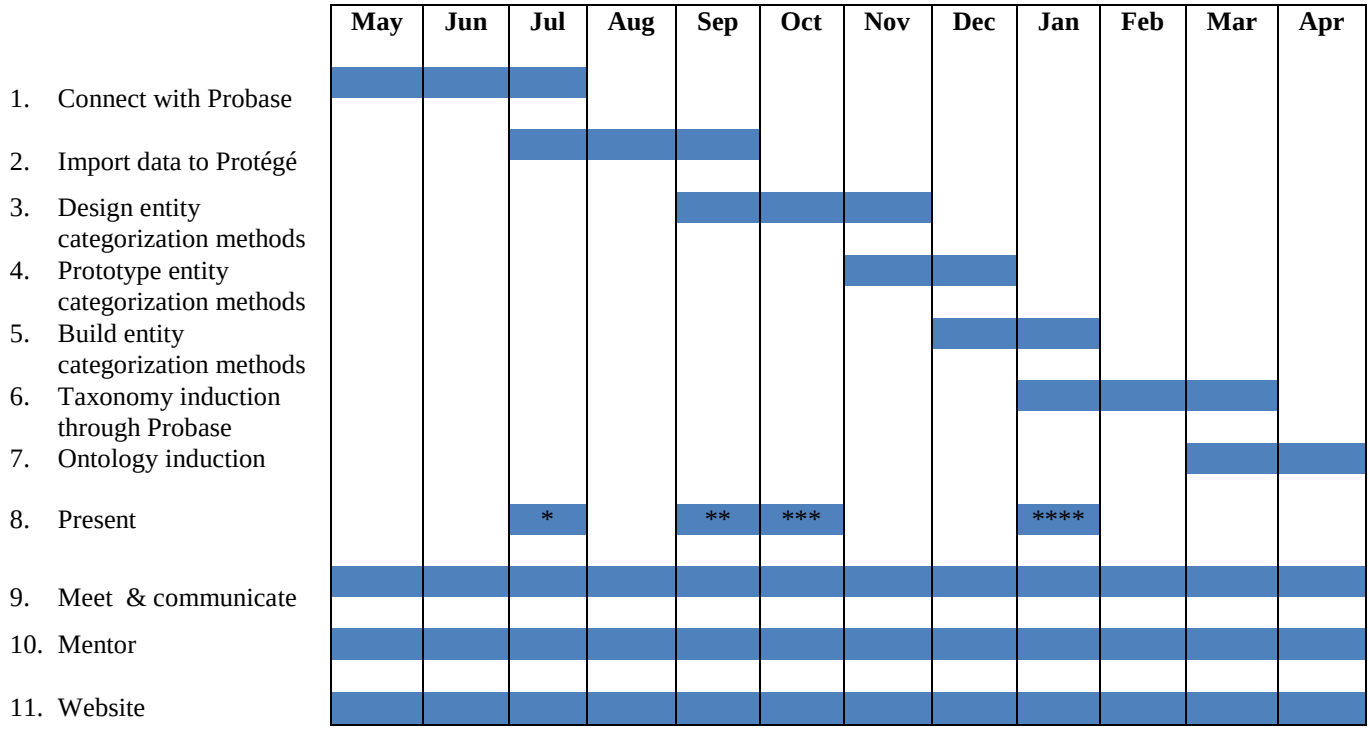
2015-2016 Work Schedule



* Association for Information Science and Technology (ASIST) Meeting 2015

** Association for Library and Information Science Education (ALISE) conference 2016

2016-2017 Work Schedule



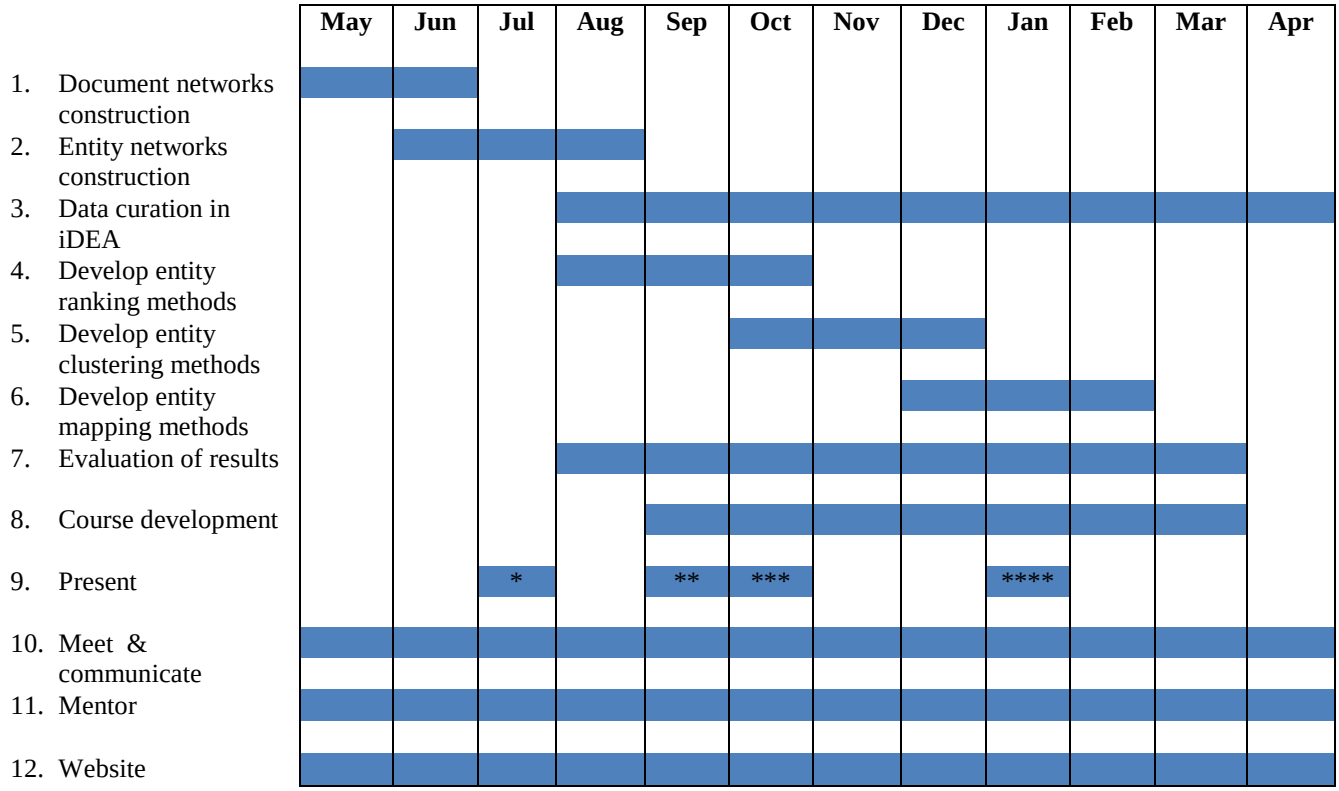
* Joint Conference on Digital Library (JCDL) 2016

** Science, Technology, and Indicators (STI) conference 2016

*** Association for Information Science and Technology (ASIST) Meeting 2016

*** Association for Library and Information Science Education (ALISE) conference 2017

2017-2018 Work Schedule



* International Society for Scientometrics and Informetrics (ISSI) Meeting 2017

** Science, Technology, and Indicators (STI) conference 2017

*** Association for Information Science and Technology (ASIST) Meeting 2017

*** Association for Library and Information Science Education (ALISE) conference 2018