NDSR Boston 2014/15 Project Description

Host and Project Title	 Tufts University - **Institutional Knowledge of Research Data at Tufts University**
Project Summary	Tufts University proposes a National Digital Stewardship Residency project that would focus on exploring strategies for Tufts to gain a more complete understanding of the research data produced by its faculty, research staff, post docs, and graduate students. In particular, this project would focus on investigating and testing strategies for producing metadata objects that represent Tufts-created research datasets and managing those representative objects in Tufts' Fedora-based institutional repository. These digital objects could reference datasets that are managed externally in environments such as subject-based repositories, or managed internally within Tufts' institutional repository. These digital objects may also reference datasets described in data management/sharing plans that are yet to be created. Understanding the scope of present and future datasets would enable Tufts to better understand its data management and stewardship obligations. The digital objects in the repository would provide baseline stewardship metadata that would enable Tufts to understand the location, structure, subject matter, retention periods, access requirements, and associated rights and responsibilities of these datasets. The goal of this exploratory project is to create a process for Tufts to have a better understanding of the research data it produces, regardless of where those datasets are managed. The project would also give Tufts a framework for understanding what information it needs about its datasets to manage their governance, use, and preservation, and support data management education for the Tufts community. The resident would collaborate with archives, library, IT, and research administration staff to identify environments and documents that contain information about research datasets and develop methods to extract and transform this information into metadata objects that can be ingested into the institutional repository. Likely sources for information about datasets include a research data management system currently being piloted at Tufts, data management plans, domain-specific repositories that contain Tufts-created datasets, and a small number of datasets held by the Digital Collections and Archives (DCA) and the Tisch Library at Tufts. This is a highly collaborative project that will require the resident to work with archives, library, IT, and research administration staff, and researchers. The primary mentor for this project would be Eliot Wilczek, Acting Director and University Archivist, DCA. Secondary advisors would be Alicia Morris, Head of Technical Services, and Regina Raboin, Data Management Services Coordinator/Science Research Librarian, both of Tisch Library. The DCA is a central administration office reporting to the Office of the Provost that serves as the university archives of Tufts and the

	Tisch Library serves as the Arts & Sciences and Engineering library of the university reporting to the Office of the Dean of Arts & Sciences.
Goals	1. Identify dataset metadata sources a. Undertake a survey and assessment of the systems, environments, and documents that contain significant bodies of information about research datasets produced by Tufts researchers. This work would identify existing information about datasets that can be used for this project; it would not be an exhaustive inventory of dataset information. Likely sources include a research data management system currently being piloted at Tufts, data management plans, domain repositories that contain Tufts-created datasets, and a small number of datasets held by the DCA and Tisch Library at Tufts. b. Document what type of metadata this information contains, such as rights, technical, or descriptive metadata, and how this metadata is structured. c. This will build on previous survey and investigative work done at Tufts that explored how its faculty conduct their research and create and manage their research data. 2. Model metadata objects representing datasets a. Determine the descriptive, technical, rights, and other types of metadata that Tufts needs to properly manage, preserve, and share research datasets over time. b. Model a structure to encode this required metadata in an object that can be ingested into the Fedora-based institutional repository. c. Model relationship statements that link these representative metadata objects with datasets. d. The modeled metadata objects and relationship statements should have the flexibility to describe datasets from a variety of disciplines; various states of encoding and structure; and reference datasets that are either in the institutional repository, in an external resource, or do not yet exist. e. This work will include examining best practices and standards in the data management field and local needs at Tufts. 3. Model workflow a. Model an overall workflow that describes processes that move from the original source of information about research datasets, to creating representative metadata objects, to ingesting those objects into the institutional repository. b. This work builds on existing processes and tools for metadata and digital object creation and repository ingest at Tufts. 4. Create and ingest proof-of-concept objects a. Create and ingest a small number of representative metadata objects that reference research datasets produced at Tufts. The metadata objects should represent datasets from a range of disciplines, include metadata originally captured from an array of resources, represent datasets existing in a range of environments, and reference datasets in a various states of encoding. b. Example metadata objects include: objects drawing metadata from data management plans representing datasets that have yet to be created, objects

	representing datasets in domain repositories, objects describing datasets in the institutional repository. 5. Write policies and procedures a. Based on the work of the first four Goals/Objectives, write policies and procedures that document sources of information about research data at Tufts and methods for extracting that information, creating metadata objects that represent datasets, and ingesting those objects into the institutional repository. b. This work builds on policy and procedure management frameworks and policies and procedures for the institutional repositories already in place at Tufts. 6. Contribute to data management curriculum and data management/sharing plans a. Observe data management classes provided by Tisch Library for Tufts faculty, research staff, post-docs, and graduate students. b. Assist in updating data management curriculum as needed based on the lessons learned, best practices, and requirements that emerge from this project and institutional needs. c. Help assist researchers with developing data management/sharing plans. d. Help modify data management/sharing plan templates based on the lessons learned, best practices, and requirements that emerge from this project and institutional needs. 7. Write project report a. Write a project report that analyzes strengths and weaknesses of the processes developed in the project. The report will include an evaluation of the resources required to scale this proof-of-concept project to a production-level process. It will also include a discussion of the viability of the Fedora-based institutional repository as a system of record for Tufts-produced research datasets.
Timeframe & Deliverables	Months 1 to 2 Observe and learn the various processes that concern this project. These include, but are not limited to, writing data management plans, creating metadata, and preparing digital objects for ingest into the institutional repository. Months 1 to 2 Dataset metadata sources (Goal/Objective 1) Months 2 to 7 Model metadata objects representing datasets (Goal/Objective 2) and model ingest workflow (Goal/Objective 3). Months 4 to 8 Create and ingest proof-of-concept objects (Goal/Objective 4). Much of this work will be done iteratively with Goal/Objective 2 and 3.

	Months 5 to 8 Write policies and procedures (Goal/Objective 5). Much of this work will be done iteratively with Goal/Objective 2, 3, and 4. Months 5 to 9 Contribute to data management curriculum and data management/sharing plans service. (Goal/Objective 6). Much of this work will be done iteratively with Goal/Objective 2, 3, 4, and 5. Months 7 to 9 Write project report (Goal/Objective 7). Deliverables 1. Description of the systems, environments, and documents that contain significant bodies of information about research datasets produced by Tufts researchers. 2. Metadata element set or data dictionary for objects representing research datasets and associated relationship metadata. 3. Policies and procedures for creating and managing metadata objects representing research datasets. 4. Updated data management curriculum and data management/sharing plan templates that incorporate lessons learned, best practices, and requirements emerging from the project. 5. Project report.
Required Resources	• Workstation and cubicle. • Access to standard systems and tools available to Tufts staff. • Access to the systems, environments, and documents that contain significant bodies of information about research datasets produced by Tufts researchers. • Access to staff in the archives, libraries, central IT division, research administration office. • As-needed access to faculty, staff researchers, post-docs, or graduate students who created research datasets that are represented by metadata objects created during this project.
Context	Research universities are facing a growing number of challenges in managing, preserving, and providing access to the research data produced by its faculty, research staff, post-docs, and graduate students. Funders are increasingly demanding that universities and researchers make their research data broadly available to the public. Most notably, the National Science Foundation now requires applicants to submit data management plans articulating how they plan to manage and provide access to the data they produce from their NSF-funded projects. Researchers are creating increasingly large and complex datasets that present significant resource and preservation challenges to research universities and institutions. Emerging data research techniques in the humanities, social sciences, and natural sciences often rely on pulling large, disparate sets of data for machine-based analysis, placing an increased importance on ensuring that research data are discoverable and well-structured to enable reuse. While domain-specific data repositories and metadata schemas have played an important role in

	managing, preserving, and providing access to research data within academic fields, research universities and institutions still face the challenge of understanding and documenting the research data its own members produce across a wide spectrum of disciplines. As a student-centered research university, Tufts researchers create a wide range of research datasets in the natural and health sciences, social sciences, and the arts and humanities. Tufts University is undertaking several initiatives to strengthen its infrastructure in order to properly support its research. This work has included expanding it research data storage capacity, implementing a new research administration system, and starting a pilot project to explore research data management solutions. The Tisch Library has been actively engaged in assisting faculty create NFS-mandated data management plans and constructing and delivering data management course material to faculty, post-docs, and graduate students. Despite these advances, Tufts continues to struggle to gain a holistic understanding of the research data that its faculty, post-docs, research staff, and graduate students produce. Researchers store datasets in a variety of environments, including disciplined-based repositories hosted at other institutions, Tufts network storage, and local storage environments. In addition, records that provide evidence of datasets are found in data management plans. Tufts does not have a baseline metadata set for documenting stewardship responsibilities, rights, and requirements for these research dataset. This makes long-term management of these datasets difficult as these responsibilities, rights, and requirements are not clearly delineated.
Required Knowledge and Skills for Resident	• Communication skills. This includes talking with people with diverse professional backgrounds and experience levels in order to collect information and translate among disciplines and areas of expertise. • Workflow development • Project management • Documenting processes and procedures • XML and XSLT • Building metadata schemas
Preferred Knowledge or Experience	• Data modeling • Familiarity with repository systems, particularly Fedora and how Fedora objects are structured • Managing datasets • Research data management systems • Curriculum development or other teaching experience