Capstone Project Charter

Project Name	Improving Trust and Interoperability: Metadata for Data Refuge's Open Data Catalog
Students Name(s)	Sam Buechler & Joan Hua
Project Description	Working with a select group of datasets, we will perform crosswalking measures to improve descriptive metadata within the Data Refuge Data Catalog. Following metadata enhancement, we will create workflow documentation for future use by Data Refuge.
Document Date	November 25, 2019
Sponsor Organization	Data Refuge (University of Pennsylvania) & University of Washington iSchool
Sponsor Name	Margaret Janz, Scholarly Communication & Data Curation Librarian & Carole Palmer, Professor and Associate Dean for Research
Sponsor Email	mjanz@upenn.edu & clpalmer@uw.edu

The Project Charter is a living document, and any substantial changes to the scope, work plan, deliverables, or stakeholders should be documented below, and an updated charter re-submitted to Canvas.

Revision History		
Date	**Author**	**Description of change**
December 10, 2019	Joan Hua	Changed estimated time commitment from sponsor, added alternative contact

PART ONE: Project Overview

The Project Overview should very clearly set boundaries for the Capstone Project. Early conversations with the project sponsor should cover all of the below elements of the project, clearly setting expectations between the sponsor and the project team.

Project Overview	
Project background	With our project, we will work to create sustainable metadata practices and standards for Data Refuge's Data Catalog. This will include an investigation of descriptive metadata needs for Data Refuge, the crosswalking of metadata from data.gov, and the creation of workflows for those who work on the catalog in the future. Up to this point, Margaret Janz and others at Data Refuge have worked to develop a workflow and mature documentation for average and advanced volunteers to seed the Internet Archive's End of Term (EOT) Harvest project with thousands of URLs at DataRescue events. These workflows include multiple steps of quality checks (Janz, 2018, p. 31; DataRescue Workflow). Having consulted previously with the EOT Harvest project—which archives government websites in anticipation of presidential administration turnovers—Data Refuge identified that they needed deeper-level URLs to extend the reach of their web crawlers. The DataRescue crowdsourced efforts had worked to address this specific limitation and it's time to move into the next phase: collating information and metadata beyond simply the original source and the URLs. *Start with a brief, high-level description of what the project is trying to accomplish, solve, or investigate. Then, in a few paragraphs, describe project context and background. What is the organizational need or gap in research that gave rise to this project? Describe the organization or intellectual drivers that created the problem, opportunity or requirement. Who are the stakeholders that will benefit?*
Objectives	● To establish a sustainable workflow to crosswalk and maintain descriptive metadata for Data Refuge's Data Catalog based on existing data.gov metadata. ● To enhance descriptive metadata of a select group of datasets to make them more trustworthy and reusable.

	● To assess the relationship between metadata that are captured in the Data Catalog and those that are published on data.gov. ● To determine next steps forward based on a clearer understanding of this relationship: e.g., metadata export/ingest schedule, partially automated steps, etc. *Describe what will be achieved and what will be delivered to the sponsor. What are the benefits that the project provides the organization or to scholarly work? What value will this project deliver to stakeholders? What is the vision - what will change or improve after the project is done? What will be in place or what could be implemented?*
Key deliverables & delivery dates	● Improved metadata of a chosen set of datasets (about 60), 5/7/2020 ● Brief writeup assessing options for crosswalking, discussing the relationship between metadata in the Data Catalog as they currently stand and those on data.gov, 5/7/2020 ● Documentation to allow Data Refuge to apply sustainable metadata workflows to the catalog, promoting discovery, preservation, and reusability of the datasets, 5/7/2020 ● Periodic progress reports ○ 2/2/2020 ○ 2/23/2020 ○ 3/15/2020 ○ 4/14/2020 ○ 4/28/2020 ○ 5/12/2020 ● Final poster, 5/7/2020 ● Presentation recording, 5/7/2020 ● Final report on what we have done, 6/5/2020 *List the actual documents, systems, or artifacts that will be produced by this project. What will you be providing to the sponsor, and when? Briefly describe the approach - how the project will accomplish its goals. For example, a large project may have a feasibility study phase followed by an implementation phase.*
Impacts & critical success factors	● The documentation is written at a level of granularity that is appropriate for the actionable next steps.

	● The written documentation is clear and appropriate for readers with LIS training and those at Penn Libraries and PPEH who would likely work on the Data Catalog. ● The metadata enhanced/added follow existing schemas, are consistently applied, and contain or reference appropriate supporting information (e.g., readme files, data dictionaries). ● The final poster and presentation recording are professional in nature and represent the range of work completed during the process and possible next steps for continued enhancement. *Describe any known or anticipated impact(s) for the organization, to the users, to society or the research community, to infrastructure, etc. How can the sponsor determine if the final are satisfactory? What indicators or measures will be used?*
In scope	● Research to gain familiarity of metadata on data.gov, and—as necessary—the CKAN platform. ● Investigate a bound set of data (e.g., NOAA datasets exclusively) that has already been bagged and uploaded to determine the level of descriptive metadata present. ● Research best practices—manual and automated—for crosswalking metadata and applying those practices to the bound set of metadata. ● Fully evaluate and describe up to 75 datasets (quantity in the NOAA collection). ● Create a sustainable workflow for existing student employees and/or volunteers to apply for the remaining datasets. *Briefly describe the boundaries of the project. What activities and tasks are in scope i.e. what are you willing and able to do to meet the objectives? What is possible given the time frame and resources available to you?*
Out of scope	● Full automation to systematically update the metadata in the entire Data Refuge Data Catalog ● Item-level metadata enhancement for datasets beyond the bound set chosen for this project ● Digital preservation plan of the datasets ● Ingesting new datasets

	What activities and tasks are out of scope? These are typically activities and tasks related to project objectives, but not needed or desired as part of the current project. For example, your project may deliver a prototype but the actual building of the production system is out of scope.

PART TWO: Project Resources

The project resources outlined below should be considered in concert with the expectations for the Project Overview above. Consider the number of team members and the approximate hours each member will have to allocate to the project, as well as access needed to other resources, such as hardware, software, or people (e.g., your sponsor).

Project Resources		
Project team & email addresses	Sam Buechler	buechs@uw.edu
	Joan Hua	joanhua@uw.edu
Hours and any cost estimates	It is estimated that the project will only cost hours of labor which are being provided to Data Refuge on a volunteer basis. The estimated number of hours expected for this project is based on the UW Registrar's expectation of credit-to-total-student-time. This comes to six hours per week for two credits each 10-week quarter, beginning January 6th (the start of Winter Quarter 2020) and ending on June 5th (due date of final Capstone deliverable). This will total to 120 hours committed, per team member, between the dates mentioned above. This estimation includes time needed to complete all progress and final reports required by the UW iSchool. *State the rough estimate of hours for the project, both for development and on-going support costs and time commitment. Note any costs that the sponsor or team may need to plan to incur (such as a license for a technology tool, hardware for digitizing photos, fees for transcribing interviews).* ***Note:*** *Start with a rough estimate, but as the project moves through each phase of the lifecycle, this estimate will become more detailed and may be very different than your original.*	
Other resources: software, hardware, other equipment, or workspace	Joan and Sam will be expected to have consistent computer access as all aspects of the project require technology to complete. *List all materials or outside expertise required to complete the project.*	
Sponsor role	Margaret Janz, co-sponsor, will be asked to: be available for periodic meetings to discuss progress and ask questions; review drafts and final	

	deliverables; give guidance as it relates to the design of metadata structure as needed; and ensure access to materials needed for completion of the project. Carole Palmer, co-sponsor, will be asked to: consult on best practices for metadata applications in a data repository, provide information on applied skills for crosswalking and preserving metadata, *What is the sponsor responsible for? Examples are feedback, testing, business decisions, participation in design and review phase, signing acceptance document.*
Sponsor time commitment	We predict the time commitment from our sponsors would amount to no more than ~15–30 hours total between January and early June. This estimate includes time spent on answering questions, giving advice, commenting on documents/workflow, and attending Zoom conference meetings as needed. At this time, Sam and Joan have opened this discussion with sponsors but they have not been able to commit to a firm number of hours. We expect to know more as the project unfolds. *Estimate the number of hours needed to maintain effective communication and keep the project moving forward. (This may easier to estimate after completing the workplan and the sponsor communication plan in part 5)*

PART THREE: Factors Affecting Project Work

Project team members may have little control over other factors that affect project work. Think carefully about assumptions the team has made (e.g., “the sponsor’s materials will be available to us offsite”), the constraints placed on the team (“the sponsor is not always online to reply to email over the weekend, when we are working”), and dependencies the project has on other events (“our sponsor uses older versions of Windows, and therefore our solution must be backwards compatible” or “the tasks defined will be completed by hourly temp workers, so the procedures we develop must be well documented and easy to understand”).

Factors Affecting Project Work	
Assumptions	● Margaret Janz, as part of Data Refuge, will provide access to workflow documentation and specific datasets. ● Data Refuge’s relevant electronic materials, Data Catalog’s publishing platform(s), and project management tools can be accessed offsite and without particular institution affiliation as necessary. ● Sam and Joan will have the necessary technology to complete metadata cross-walking & communicate remotely with each other/sponsors. *What are the assumptions being made? Use true statements but ones that contain a measure of risk. For example, “the sponsoring organization will provide access to the artifacts” or “the organization will purchase the necessary license for content.”*
Constraints	● Participants in this project are geographically diverse and have other jobs/responsibilities outside the project which can affect the ability to meet and communicate in a timely fashion. ● Outside of this project, both Joan and Sam have full-time job responsibilities during the week: Joan expects to work on this project in the evenings and on weekends Pacific Time, and Sam expects to work during the same hours but can be available for communication purposes throughout the day. Both parties respect the fact that Margaret will primarily be available during business hours M–F Eastern Time and will therefore plan accordingly to make sure there is adequate turnaround time allowed for questions/answers and deadlines. *Describe any constraints on this project that will affect budget, quality, schedule, resources or scope. For example, “the team is geographically*

	dispersed, so time zones will affect the ability to all meet" or "the project requires travel to Portland to visit the site, so access to the participants is constrained by schedule and costs".
Dependencies (if any)	● Select datasets on data.gov will remain accessible in the duration of this project from January to May so that Sam and Joan can compare their metadata to that Data Refuge's. *List any dependencies this project may have on other projects or other infrastructure requirements – for example, "construction of the thesaurus is dependent upon the SharePoint implementation" or "the final program must be launched in concert with the opening of the new building."*

PART FOUR: Communication Plan

In order to ensure sponsor satisfaction and fulfillment, it is crucial to communicate effectively and regularly so that problems can be addressed early by the group. Each of the team members, and the sponsor, should understand the communication frequency (e.g., weekly on a certain day, after deliverables are complete, etc.) and agree upon its terms outlined below.

Communication Plan					
Team meeting frequency	Student team members will meet biweekly as needed; particulars on meeting schedules, needs, and communication are in the Team Contract document.				
	Email	**Phone call**	**In-person meeting**	**Formal Report**	**Other (specify)**
Sponsor-preferred update format (select 1-2)	Priority		X		Second priority: Zoom web conference The link will be https://zoom.us/j/8295560657 (Joan Hua's personal Zoom room) for <40 minutes at a time, unless otherwise specified.
	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Reporting frequency (indicate day of week and how often –	Biweekly email updates will be sent on Mondays				

weekly, biweekly, etc.)	starting Jan. 6, 2020. We will schedule a midway video meeting in February.				
Alternate point of contact at sponsor organization	For now, the alternative point of contact at Penn Libraries is Will Noel, wgnoel@pobox.upenn.edu. *In case the primary sponsor becomes unavailable due to illness or other emergency, who should be the secondary point of contact?* *First and last name. Title. Email and phone number.*				
Issues management plan	When unexpected and/or critical issues arise, the project team will communicate them ASAP to sponsors and propose solutions or backups if possible. The team will discuss the steps forward. Per Margaret's request, when the situation warrants it, we will email Margaret with URGENT in the subject line. To be safe, we will cc her personal email margaret.janz@gmail.com. We can determine a course of action depending on the issue. *When unexpected and/or critical issues arise, what steps do the sponsor (and/or customer) want the project team to take?*				
Change management plan	During the project's active period, the project team will iterate and make changes to project implementation procedures. After the Capstone project duration ends, Margaret will have the full power in making decisions regarding how the project would be implemented. *How will changes required by project implementation be handled within the organization? Make suggestions for managing the changes. These may occur after the team's work has concluded.*				