CS598 Course Project Pilot: End-to-End Data Curation Workflow

Note: For Fall 2025, this is an optional pilot available to a limited number of students as an alternative to the midterm and final exams. Your project must be approved by course staff to proceed.

Overview

The purpose of this project is to demonstrate your understanding of key concepts and techniques addressed in this course through the planning and implementation of an end-to-end data curation workflow. It is designed to give you hands-on experience applying course concepts to use cases or research questions of your choosing.

Your project should include, as appropriate:

- Ethical and legal data handling
- Data collection and/or acquisition
- Data modeling
- Quality assessment
- Transformation, cleaning, and/or integration
- Creation of an automated workflow including provenance capture
- Metadata and data documentation
- Packaging your work so that it is understandable, reproducible, and reusable

Exams or Project?

Your choice should reflect your learning style, time availability, and interest in hands-on application. Both options are graded with equal weight.

- **Exams**: Structured, with a fixed timeframe and a defined set of questions. Best for students who prefer a focused assessment of conceptual knowledge within clear boundaries.
- **Project**: More open-ended, allowing you to design and carry out a self-directed curation project. This option requires initiative and time management, and may appeal to students who want deeper engagement with the material.

Teams

You are permitted (but not required) to work in teams of 2-3. Team projects are expected to be more complex than individual projects. Additionally, team member contributions must be evident through, for example, Git commits, log files, or contribution statements in the status and final reports.

Milestones

The project has three milestones:

Milestone	Due	Weight	Description
Proposal	9/15	10%	A proposal describing your project idea, data sources, and planned curation activities.
Progress Report	10/27	10%	A status update describing what you've accomplished so far, any challenges you've faced, along with next steps
Final Submission	12/10	80%	A complete project report and all supporting artifacts.

Outside of these milestones, throughout the semester course staff will be available to provide support and guidance on your project. For more details, see the Project Milestones question.

How to apply

- Read these **Course Project Instructions**
- Complete the **Course Project Interest Form** by 9/5.
- Submit your **Project Proposal** via Coursera by the deadline. *Important: Late submissions will not be accepted*

Selection process

The goal is to accommodate as many students as possible while ensuring sufficient resources for support.

- If submissions are limited, all qualifying proposals will be accepted.
- If we receive more proposals than we can accommodate, we will prioritize high-quality submissions as assessed by course staff.
- If there are still more proposals than slots, we may use a lottery.

Project Milestones

Project Proposal

Instructions

Create and submit a project proposal in PDF format that includes the following:

- **Overview**: Describe the overall goal of your project including use case or research question.
- **Plan:** Describe your project stages, related back to a data lifecycle model.
- **Data sources**: Identify and describe the dataset(s) that you intend to use.
- **Team**: If working in a team, clearly define roles and responsibilities.
- **Timeline**: A timeline for implementing your project. If working in a team, specify who will complete each task.
- **Constraints**: Describe any known constraints
- **Gaps**: Identify any known gaps or areas where you need additional input or information.

Your plan should anticipate later course topics even if you don't yet know all the details. It is expected that your plan will evolve over time.

Deliverable

- Submit a PDF of your plan
    - ~750-1000 words (3-5 pages double spaced). (Note: These are intended only as rough guidelines. Use the space you need to address the plan requirements)
    - Include references to any sources or documentation reviewed in APA citation style.

Progress Report

Submit a report on the status of your project that includes the following:

- Status of deliverables outlined in your preliminary plan.
- Describe and justify any changes in scope or deliverables
- Evidence of progress through corresponding artifacts.

Deliverable

- Submit a PDF of your progress report.
    - ~500-750 words (2-3 pages double spaced). (Note: These are intended only as rough guidelines. Use the space you need to address the plan requirements)

    - Include references to any sources or documentation reviewed in APA citation style.
- Zip file, link to GitHub or other repository that contains any artifacts of your project (data, code, workflows, documentation, etc) demonstrating progress.

Final Submission

Submit a single PDF containing your final narrative report along with any supplementary materials as a link (e.g., to a Github repository or Box folder) or Zip archive. The purpose of the report is to summarize your data curation project.

The report should be concise, summarizing your workflow and situating it in the course concepts. The primary evidence of your work will be in your supplementary materials (repository/artifacts).

Your report must include the following:

- A narrative summary of your project motivation and context including use case or research question.
- A brief profile of the datasets used.
- Description of the actual data curation workflow as it was performed (see list of possible activities below). Your description should reference any supplementary materials (e.g., data, scripts, etc) and can point to documentation in an external repository.
- A brief analysis of your workflow as it relates to one or more of the lifecycle models discussed in class.
- Summary of any findings, problems encountered, and lessons learned (including possible next steps).
- Connection to course concepts and readings (including references).

Supplementary materials:

- Zip file, link to GitHub or other repository that contains any artifacts of the curation process (data, code, workflows, documentation, etc).
- Minimum required items: scripts, workflow, documentation, and environment specification.

Your final report should describe relevant actions taken based on concepts introduced in class.

Required for all projects:

- **Data Lifecycle (cf. M1):** Relate your project to one or more of the lifecycle models discussed in class.
- **Ethical, legal, or policy constraints (cf. M2):** Describe any ethical, legal, or policy constraints and how they were addressed.

- **Data models and abstractions (cf. M3-5):** Identify and describe data models and abstractions used.
- **Metadata and data documentation (cf. M8)**: Provide detailed metadata describing your dataset and project following a standard such as DataCite, schema.org, or similar. Include a data dictionary or codebook where appropriate.
- **Workflow automation, provenance, and reproducibility (cf. M12):** Document or automate your workflow in a way that is transparent and reproducible.
- **Dissemination and communication (cf. M15):** Package your project in a self-contained structure that is understandable, reproducible, and reusable. For example, a GitHub repository or Zip archiving containing all artifacts and complete documentation.

If relevant to your project:

- **Data integration and cleaning (cf. M6):** Describe integration issues (e.g., heterogeneities) and how they were addressed. Describe quality assessment and cleaning processes.
- **Identity and identifier systems (cf. M9)**: Identify and describe how any identifiers and systems (e.g., UUIDs, hashes, URIs, DOIs) are used in your project. Justify identifiers that you selected.
- **Standards and standardization (cf. M11)**: Identify and describe standards used in your project. Justify any standards that you selected.
- **Data concepts (cf. M7) and Data practices (cf. M13):** Relate your project and workflow to the concepts discussed in class related to the ontology of data (e.g., BRM model) or data practices research.

Deliverables:

- Final project report as PDF. Your report should be ~1500-2500 words. (Note: Word counts are intended only as rough guidelines. Use the space you need to describe your project).
- Include references to any sources or documentation reviewed in APA citation style.
- Supplementary materials as Zip file, GitHub repository, or archived package.
- Submit a PDF of your plan.

Grading Rubric

Your project will be evaluated both on your written reports (plan, progress, and final) as well as artifacts (data, code, workflows, documentation, metadata, etc).

Below is a breakdown of the contribution of each element to your final grade. The project is worth a total of 150 points.

Element	Points	Percent
Project Plan	**15**	**10**
Progress Report	**15**	**10**
Final Submission	**120**	**80**
Artifacts & Workflow	30	20
Metadata & Documentation	15	10
Reproducibility & Transparency	15	10
Application of Course Concepts	30	30
Final Report	30	30
Total	**150**	**100**

Detailed Rubric

(Note: This rubric is subject to change to improve clarity)

Element	Full Credit	Partial Credit	Low/No Credit
Project Plan	Clear, feasible, well-structured plan; data sources identified; realistic timeline; constraints and gaps thoughtfully addressed.	Plan missing some details, timeline vague, or feasibility unclear. *Only plans that receive full credit will be permitted to proceed.*	
Progress Report	Provides clear status updates with progress measured against plan; challenges identified; scope adjustments justified; evidence of progress through submitted artifacts.	Artifacts are included but incomplete or disorganized. Workflow only partially documented. Protected/restricted artifacts identified but not described.	Minimal update, no evidence of progress or reflection.
Final Submission			
Artifacts & Workflow	Complete set of well-organized and working artifacts (scripts, workflow, documentation, environment specification). Workflow is automated or clearly documented. Access to protected/restricted artifacts is clearly documented.	Artifacts are included but incomplete or disorganized. Workflow only partially documented. Protected/restricted artifacts identified but not described.	Few or missing artifacts; workflow not documented. Little or no information about protected/restricted artifacts.
Metadata & Data Documentation	Detailed descriptive metadata provided using a relevant standard (e.g., DataCite, schema.org) or similar to support discovery. Clearly documented data	Some metadata or documentation provided by incomplete. Missing or incomplete descriptive	Little or no metadata or data documentation provided.

	dictionary/codebook provided describing variables, units, etc. to support independent understanding and reuse.	metadata. Missing or incomplete data dictionary/codebook	
Reproducibility & Transparency	Clear, step-by-step instructions to reproduce results or processing workflow; results/outputs can be reproduced using provided artifacts including a detailed description of the computational environment (or container image). If using protected/restricted artifacts, a transparent record of processing is provided.	Reproduction is possible but requires guesswork or manual fixes.	Reproduction not possible; documentation insufficient
Application of Course Concepts	Demonstrates strong understanding of required elements (data lifecycle, ethics/legal constraints, metadata, workflow automation, provenance, reproducibility, dissemination). Concepts applied appropriately and explicitly tied to course content. Any missing elements are clearly justified.	Some course concepts addressed but superficial or incomplete application without justification (~2% per topic/requirement).	Minimal or no evidence of applying course concepts.
Final Report	Concise, clear narrative linking project to course concepts; narrative explanation of the curation workflow; addresses all required elements or justifies missing elements; discussion of findings, challenges and lessons learned.	Report submitted but lacking clarity, organization, or explicit course connections. No discussions of findings, challenges, or lessons learned.	Missing or very incomplete report.

Example Project Ideas

The following summaries illustrate the kinds of projects that would be considered in scope. When selecting your own project, keep these points in mind:

- **Choose something meaningful**: You are encouraged to select a project that connects to your personal interests, academic work, or workplace context.
- **Keep the scope manageable**: This is a semester project, not a full research study. Aim to demonstrate the core requirements—lifecycle, acquisition, cleaning, metadata, workflow automation, reproducibility, documentation—without needing to solve a complex real-world problem.
- **Address access and transparency**: You may work with private or protected data, but you will need to address issues of access and transparency. For example, if using workplace data that cannot be shared, you must document your process clearly so that course staff can evaluate your work.

Predicting Food Inspections

This project will create a curated dataset and reproducible workflow to integrate data from the City of Chicago data portal to explore patterns of food inspection violations in city restaurants. Inspection, business license, and other data (e.g., weather, rodent complaints) will be acquired, quality checked, cleaned and integrated to analyze factors that may contribute to inspection failure. A clearly documented manual workflow will be provided along with all scripts to reproduce analytical results. The curated dataset will include comprehensive data documentation and metadata in schema.org format to support discovery.

Curating Bike-Share and Weather Data

This project will create a curated dataset and automated workflow to integrate Divvy (and other city-operated bike-sharing) trip data with public weather data sources in order to explore patterns in bike-sharing usage under different weather conditions. The workflow will include data acquisition, quality assessment, cleaning, and integration across sources. All processing will be conducted in compliance with data provider terms of use. The resulting analysis-ready dataset, automated workflow, and supporting documentation will be published in a GitHub repository and archived with a persistent identifier. Detailed dataset documentation will be provided and metadata in DataCite JSON format to facilitate discovery.

Curating Riot's TFT Match Data

This project will develop a curated dataset and reproducible workflow for analyzing match data from Riot Games' Teamfight Tactics (TFT) public API. The end-to-end workflow will include automated acquisition, quality assessment (e.g., detecting formatting changes, handling missing

or inconsistent fields), and organization of data into a structured, analysis-ready format. All steps will be carried out in compliance with Riot's terms of use. The resulting workflow (implemented using Snakemake), dataset, and documentation will be published via a GitHub repository. Detailed dataset documentation will be provided and metadata in schema.org format to facilitate discovery.

Curating a Privacy-Protected Health Activity Dataset

This project will demonstrate how to create a privacy-protected dataset from wearable device data. The workflow will include acquiring raw data exports, performing quality assessment, and applying privacy-preserving methods (e.g., de-identification, aggregation, pseudonymization). The curated dataset and workflow will be documented with a clear data dictionary and metadata in schema.org format to facilitate discovery. The final package will include original workflow scripts, documentation of privacy steps taken, and an analysis-ready dataset suitable for sharing without exposing personal information.

FAQ

This section will be revised as needed.

Q. Does the project replace both exams?
Yes. The project is worth 30% of your grade, the same as the combined midterm + final exams.

Q. What if I change my mind?
If you are selected and decide not to complete the project before the midterm, notify course staff via Campuswire and just complete the exams. After the midterm, you must complete the project.

Q. Why is the proposal due so early?
This will give course staff time to assess submissions and provide feedback in advance of the midterm exam.

Q. How will my grade be reported in Coursera?
Good question. I am not aware of a way to have conditional assessments in Coursera. For now, since they are optional, the Course Project deliverables will have 0% weight in the gradebook while the midterm/final will have weights. Your grades for the project will be visible, but not reflected in the exam scores. If we can't find another option we will either (1) override the exam scores or (2) apply your project score during final grade submission.

Q. I registered late, can I still apply?
Yes, if you are a late registrant reach out to **Instructors & TAs** via Campuswire.

Q. Can I base my project on one of the examples?
Yes, you can base your proposal and project on one of the examples. Obviously they are just summaries, so there are many questions that you'll need to address.

Q. Can I use an existing dataset or do I need to collect my own?
You can use an existing dataset or collect your own.

Q. Can I use GenAI?
Use of GenAI is acceptable. You should make clear where it is used and possibly cite it (e.g. if using as a source of information). If part of a pipeline, you'll need to think about what this means for transparency/reproducibility.