Project Management Plan: Digital Curation of the British Library Drama Dataset

Team Composition: Chahna Paresh Ahuja, Liangyu Gan, Xinran Liu
Project Duration: 13 October 2025 – 7 December 2025

1. Project Overview and Context

Our team acts as digital intermediaries to process the British Library's *MicrosoftBooks_Drama.xlsx* dataset. To ensure a coherent analysis of theatrical works, the project strategically targets three specific spreadsheets within the source file: Tragedy, Comedy, and Play, while excluding the General Drama and Recitation sheets.

The primary objective is to transform raw, inconsistent metadata into a structured and clean dataset suitable for computational analysis. To address the unique challenges of the source file's fragmentation, we developed a custom "Divide, Standardize, and Integrate" strategy. Instead of relying on generic cleaning pipelines, we implemented a specific workflow that allows for parallel processing of the Tragedy, Comedy, and Play sheets while ensuring seamless final integration.

To ensure data interoperability during the eventual merger, we established a centralized "Cleaning Protocol" prior to execution. The process involves exporting the sheets for cleaning in OpenRefine, and finally, merging the three harmonized files into a single, consolidated dataset by using the Pandas library in Python. This unified output, alongside comprehensive documentation of the cleaning logic (GREL scripts), will serve as the foundation for the subsequent analytical phase.

2. Workflow and Work Breakdown Structure (WBS)

The project lifecycle is divided into four sequential phases to ensure a systematic and feasible workflow.

Phase 1: Initiation and Protocol Definition (13 Oct – 26 Oct): This phase focuses on preparing the ground for distributed work. We begin by assessing the selected spreadsheets to identify recurring inconsistencies in key bibliographic fields. Based on this assessment, a shared cleaning and standardization protocol is defined, covering column selection, author name normalization, date extraction, handling of missing values, and character encoding. In parallel, we set up the collaborative working environment, including shared repositories and agreed communication channels.

Phase 2: Distributed Data Cleaning (27 Oct – 23 Nov): During this phase, each team member cleans one spreadsheet using OpenRefine, strictly following the shared protocol. Clustering functions and GREL transformations are applied to address inconsistencies in author names, titles, imprint information, etc. Regular coordination ensures that emerging issues are discussed and resolved collectively.

Phase 3: Quality Assurance and Integration (24 Nov – 1 Dec): Quality assurance is carried out through peer review and semi-automated sampling of approximately 10% of the records in each dataset. This step verifies compliance with the agreed protocol. Once validated, the three datasets are aligned at the schema level and merged into a single consolidated dataset.

Phase 4: Documentation and Closure (1 Dec – 7 Dec): The final phase focuses on documentation and delivery. OpenRefine operation histories (JSON) and GREL scripts are compiled, and the final cleaned datasets and documentation package are prepared for submission.

3. Team Roles and Responsibilities

To ensure balanced workload and accountability, roles are distributed according to both functional responsibilities and data ownership.

Team Member	**Dataset Responsibility**	**Functional Role**	**Key Responsibilities**
Liangyu Gan	Play	Project Coordinator	Oversees timeline and Gantt chart; coordinates meetings; manages final submission.
Chahna Paresh Ahuja	Comedy	Technical Lead	Develops and validates GREL transformations; ensures schema consistency.
Xinran Liu	Tragedy	QA & Documentation Lead	Manages version control; conducts peer review; compiles final documentation.

4. Timeline and Milestones

The project timeline is structured around clearly defined milestones to ensure feasibility and timely delivery. A detailed Gantt chart accompanies this plan and aligns with the milestones.

- **Milestone 1 (26 Oct): Protocol Freeze.** Data standards and column structure are finalized.
- **Milestone 2 (23 Nov): Completion of distributed data cleaning**.
- **Milestone 3 (1 Dec): Quality assurance sign-off and dataset integration**.
- **Milestone 4 (7 Dec): Final delivery of datasets and documentation**.

5. Coordination, Tools, and Resource Management

As the project relies exclusively on open-access data and institutional tools, resource management focuses on coordinating time, labor, and technical infrastructure rather than financial planning. We utilize OpenRefine for cleaning, OneDrive/GitHub for version control, and Notion for task tracking. Weekly synchronization meetings complement real-time messaging.

6. Quality Assurance and Risk Mitigation

Potential risks include inconsistent interpretation of metadata standards across team members and delays in the distributed cleaning process. These risks are mitigated through a shared cleaning protocol, regular coordination meetings, peer review sampling, and strict milestone enforcement.

7. Key Deliverables

Three cleaned sub-datasets (CSV): Tragedy, Comedy, and Play; Consolidated master dataset; Workflow documentation, including cleaning protocol, OpenRefine JSON histories and GREL scripts; GitHub Repository: https://github.com/chahna-ahuja/dh-g7-project.git.

Conclusion: Overall, this project management plan ensures a transparent, feasible, and methodologically robust approach to bibliographic data curation, establishing a reliable foundation for subsequent analytical research.