9) Data Management Plan

Data to be Generated:

Preliminary data will be produced directly from the images of the playbills:

- **JSON data** produced from transcriptions produced by both public crowdsourcing and dedicated transcription by the project assistant using the Ensemble software
- **Text files** produced by OCR, which will be converted, using text mining approaches and manual work, into structured **JSON** files

Note that OCR work will be performed on all the same data as the transcription data as a comparison of the data generated by the two approaches. The OCR work will be much more experimental in terms of how accurately the OCR technology will be able to read the varying fonts of the playbills and how well the resulting text files are able to be converted into structured data. The transcription methods will present the challenge of cleaning up potential human errors and inconsistencies produced when entering the transcriptions.

- This preliminary data as will then be cleaned using Open Refine and converted to **RDF triples**. Portions of the existing metadata contained in **Marc** records for images in the Furness Theaterical Image Collection will also be converted to RDF.

Data Formats and Dissemination:

The final format for the project data-set will be RDF. The final RDF data, as well as the clean versions of the preliminary data (both JSON and the best text files produced by the OCR work) will be made openly available on GitHub. The project will also produce a website where the final data set will be made publicly available. All images used in the project are or will be publicly available on Penn Libraries' website, and digitized images of the playbills will be available on Penn Libraries OPenn repository project.

Data Management and Maintenance:

Upon completion, all data for the project, along with a white paper documenting the projects process and outcomes, will be stored in the Penn Libraries' institutional repository. Long term maintenance of the project website will be transferred to Penn Libraries' IT staff. Penn Libraries provides a $20 million dollar facility, and both Special Collections and Digital Humanities are at the heart of its strategic plan. The Kislak Center staff work between and across the worlds of traditional special collections work and digital innovation, with a team that includes staff with blended expertise in data curation and book and manuscript history, including a curatorial position exclusively dedicated to Digital and Research Services. Both the Penn Libraries Digital Scholarship Center, and the recently founded Price Lab for the Digital

Humanities at Penn also offer technical expertise and resources for ensuring a secure and sustainable environment for Digital Humanities projects at the University of Pennsylvania.