**Fondren Fellows Project Proposal Submission Form**

**Please complete this form by noon on July 22, 2024, in order for your project to be considered for the 2024-2025 Fondren Fellows program.**

**Any questions can be directed to** [**fondrenfellows@rice.edu**](mailto:fondrenfellows@rice.edu)

**Name** Kenneth Evans, Kirstin Matthews

**Department/Organization** Baker Institute, Science and Technology Policy Program

**Email** [ke1@rice.edu](mailto:ke1@rice.edu), [krwm@rice.edu](mailto:krwm@rice.edu)

**Phone Number** (713) 348-3824 (Evans), (713) 348-4784 (Matthews)

**Project Title** Building and Employing New Tools for the Study of US Science Policy

**Proposal Summary (This is the shorter version of your proposal, which will be used to advertise your project to potential fellows and on the Fondren Fellows website. 50 words max.)**

This project will build an open-source tool to automate the processing of complex, born-digital records. The tool will be employed to process PDFs generated in response to Freedom of Information Act Requests to be included in the White House Scientist and Science Policy Dynamic Digital Archive at Woodson Research Center.

**Please provide a brief (up to 500 word) summary of the proposed project, explaining why it is suitable for the Fondren Fellows program. What impact will this project have? How feasible is it to implement?**

*Aim 1: Processing of complex, born-digital records*

Born-digital records present enduring challenge for digital archiving—textual data are unstructured and files are incompliant with ADA regulations. We will leverage AI/ML services to build an open-source tool for batch processing complex textual records. In particular, we plan to automate the processing of large, mixed-media PDFs generated in response to Freedom of Information Act (FOIA) requests related to the role of scientists in US policymaking. The records will be included in the White House Scientist and Science Policy Dynamic Digital Archive (DDA), an online heritage collection housed at Woodson Research Center.

The FOIA PDFs contain an array of scanned and redacted documents, such as emails and other communications; born-digital documents with embedded images, such as policy reports and news media; and handwritten notes and personal correspondence. We have collected roughly 30,000 pages of PDFs, each 200-800 pages in length. These documents need to be OCRed, corrected for read order, itemized, and described with metadata. Jin, Evans, and Mulligan (2024) outlines a workflow that uses Microsoft Azure AI to OCR, Apache Tika for named-entity recognition (NER), and ChatGPT to generate descriptive metadata. This project will develop an open-source tool that integrates these services into a user-friendly interface capable of processing these documents at scale.

*Aim 2: Building a public relational database*

Quartex offers a searchable platform for viewing and studying heritage collections. However, it remains limited for studying the data via computational text analysis. Our 2023-24 Fellow, Devin Von Arx, built a pilot relational database in Django (Von Arx, Traylor, and Evans, 2024). The database provides users with a platform for visualizing and understanding the relations between people and documents through linked metadata.

The Django database is currently local and not yet published online for public use. During this Fellows cycle, the database will be launched on a website and made accessible to scholars. Further, as documents are processed under *Aim 1*, they will be uploaded to the relational database in Django and to the DDA on Quartex. Finally, we will create an API capable of pushing data from Django to external software for research using computational text analysis.

*Impact*

This project addresses two core challenges in digital archiving: processing of complex, born-digital records and data access and preservation. Both the proposed tool and database will enable the study of US science policy by students and scholars beyond Rice University. The project will result in both a generalizable, open-source tool for processing born-digital records, as well as academic works—conference presentations and academic publications—that utilize and study the processed records.

*Feasibility*

The Fellows and Mentors will draw on the expertise of Ying Jin, the project library liaison, the wisdom of John Mulligan, and the experience of Jordan Traylor. Jin is deeply familiar with Microsoft Azure AI and experienced with Django. Mulligan, an accomplished digital humanist and Django evangelist, will continue to advise on the project. Traylor is a highly capable digital librarian and researcher, who has been leading DDA development.

*References*

Jin, Y., Evans, K.M., and J. Mulligan. (2024). “The Road From DSpace 6 to DSpace 7 and Beyond: Building (and Building on) Two Modern Digital Repositories at Rice University” (poster). Open Repositories 2024, Gothenburg, Sweden. doi: [10.25611/b4zd-ht84](https://doi.org/10.25611/b4zd-ht84).

Von Arx, D., Traylor, J., and K.M. Evans. (2024). “Building and Employing New Tools for the Study of US Science Advisors” (poster). Alliance for Digital Humanities Organizations 2024, Arlington, VA. Forthcoming.

**How many fellows are you requesting (up to 3)? How would their work be broken down and managed?** Three. This project will be highly collaborative—the two Aims overlap significantly and will need to be developed in parallel. The group will be managed through weekly coding sessions coordinated by Evans. All code developed during the project will be hosted on GitHub, which enables Mentors, project advisors, and Fellows to monitor changes to scripts and collaborate on building the tool and database together.

**Outline the key tasks that the Fondren Fellow(s) would work on.**

Fellow 1: This fellow will be responsible for collaborating with Jin on optimizing Microsoft Azure AI for use on FOIA PDFs. They will also be responsible for researching how to intelligently itemize PDFs, perform NER, and develop generative AI code to write descriptive metadata of itemized PDFs.

Fellow 2: This Fellow will be responsible for packaging AI/ML services and open-source code into a user-friendly interface.

Fellow 3: This fellow will be responsible be responsible for developing, publishing, and expanding the Django relational database.

**What qualifications would you expect from students working on this project?**

Fellow 1: This Fellow will be expected to be familiar with the current landscape of open-source generative AI models, as well as the programming languages and compilers to run them.

Fellow 2: This Fellow will be expected to be familiar with software development, usability, and accessibility.

Fellow 3: This Fellow will be expected to be fluent in Python to effectively manage, publish, and expand the Django relational database.

**What would students learn through their participation in this project?**

All three students will develop an advanced familiarity with the history of US science policy, experience with database creation and management, and text analysis tools. They will also develop critical communication skills through presentations and writing projects associated with the project.

**If you don’t work in Fondren Library, please note who you would like to serve as the library liaison for your project. (If you don’t know, just write TBD.)**

Ying Jin

**Please attach/provide a link/include in your email a copy of your proposed budget of up to $6000 (if applicable). This budget should include a brief budget narrative with a line-item breakdown of planned costs. In the narrative, explain why each line item is necessary to meet the project’s goals and how the cost was calculated. (Keep in mind: No money may be budgeted for faculty salary. A separate fund will be available for students traveling to present at conferences. Proposals not requesting project funds do not need to include a budget.)**