The Road From DSpace 6 to DSpace 7 and Beyond
- Building (and Building on) Two Modern Digital Repositories at Rice University

Challenges & Opportunities
This presented interesting challenges and opportunities. The challenges were that we had to preserve handles, make hard decisions regarding legacy customization, and update our preservation strategies across DSpace, DuraCloud, Quartz, and Archivematica. The opportunities were that these modern platforms offered affordances that allowed us to give stakeholders and staff the ability to creatively and sustainably enrich their collections' metadata and use APIs to build external UIs in a modular, iterative, and hopefully sustainable manner.

Background
In 2021, Rice University’s Fondren Library formed a task force to research new platforms for its Institutional Repository. After 16 months of research among several candidates, the committee selected two platforms, DSpace 7 and Quartz, to meet the diverse needs of its stakeholders.

As a result, Fondren Library undertook the unique project of simultaneously launching two modern digital repositories: transitioning from DSpace 6 to 7 and migrating 27% of cultural heritage-related content to the Quartz service, with 11% of content retired.

Two Examples
PCAST Project:
The White House Scientist and Science Policy Dynamic Digital Archive collects and preserves materials related to the contemporary history of US Presidential Council of Advisors on Science and Technology (PCAST), from 1988 through the present day.

The collection, which is managed collaboratively by the Woodson Research Center and the Baker Institute Science and Technology Policy Program, has acquired roughly 10,000 digital assets, totaling over 30,000 pages, which need to processed for inclusion in Quartz.

Subway Ad. Project:
The Chao Center for Asian Studies Ephemera Archive houses over six thousand images of commercial subway advertisements collected and photographed by Baker Institute Fellow, Steven Lewis, between 1998 and 2003.

The Django Module (Central Hub)
This Django module works as a central hub for our extended services workflow and is a pilot project for the PCAST team, which aims to help users study the links between people, institutions, and policy documents, and how those relationships have changed across time, presidential administrations, and different areas of science, technology, and innovation policy. The module pulls the metadata from Quartz using API, and feeds the local SQLite database. This allows the team to leverage their existing metadata, connect their database to external services for OCR and named entity recognition like Apache Tika and ChatGPT, and do so without hosting any images but rather simply by maintaining pointers to Quartz IIF endpoints. Ultimately, the goal is to construct a knowledge graph by building this data out in a controllably flexible manner, allowing much of the metadata gathered to be mapped back into a long-term digital repository (DSpace or Quartz) using linked open data standards.

With the success of the pilot project, we want to generalize the module and provide the same extension to the Subway Ad. and other projects in the future.

Automation with AI
Both collections present significant challenges for traditional OCR methods. In the first, many PDFs arrive in large, low-quality PDFs generated from Freedom of Information Act requests. In the second, digital photographs of subway ads contain grainy or out-of-focus text in Mandarin. OCR is critical for the wide accessibility and searchability of these items by users. Additionally, the scholars plan to conduct text analysis of the records to study their contents, which will require appropriate itemization, accurate named entity recognition, and automated metadata creation at scale. The Microsoft OCR can recognize 164 languages with high accuracy. Yet the low cost makes it ideal for large OCR projects. Our prototype can batch process multi-page PDFs’ images and generate searchable PDFs. The future work includes streamlining the process, make it run on docker, and incorporate Tika, ChatGPT or other AI tools to extract metadata from the OCR text.