COLLEGE OF
UNIVERSITY LIBRARIES
& LEARNING SCIENCES

Organization: University of New Mexico

Primary mentor: Jonathan Wheeler, Data Curation Librarian, University of New Mexico

Project Title	Search Stories: Developing methodologies to interpret search behavior of users of institutional repository content.
Description	RAMP, the Repository Analytics & Metrics Service, was built to provide institutional repository (IR) managers with data about the search engine performance of IR content in Google properties including web search and Google Scholar. Since 2017, continuation of the service has allowed RAMP administrators to accumulate and publish a large, longitudinal dataset enabling novel comparisons of the search engine performance of IR content across institutions and platforms. The data are highly structured and support a broad range of quantitative analyses, but taking analyses beyond search engine performance in order to better understand more clearly when, how, and why users access IR content requires combining RAMP data with data from other sources including descriptive metadata, citation information, institutional data (Carnegie classification, enrollment, etc.), and global economic indicators.
Problems/ Research Questions	Scalable, reproducible methods for web scraping, text extraction, and computation are needed in order to maximize the research and analytic potential to be gained from combining RAMP data with data from openly available data sources including the World Bank, IPEDS, Crossref, and others. This project will explore and develop strategies to address challenges including: • Variation and inconsistencies in the use and application of different IR metadata schema • Extraction and validation of DOIs from descriptive metadata • Methodological approaches for analyzing uneven distributions in large, quantitative datasets
Techniques	• Web scraping • Big data storage, indexing, and retrieval • Natural language processing and text mining • Data visualization • Data storytelling

Tools/ Languages used	RAMP data are serialized in JSON format. Item level descriptive metadata for a large subset of items held by RAMP repositories are available in a SQL database. Prior analyses have been completed using Python, R, and Tableau. The project team possess expertise in Python, R, Jupyter Notebooks, SQL and Excel. Some capacity exists to support mentees interested in network analysis and NoSQL/graph databases.
Data	Description: Search engine performance data harvested from Google Search Console daily via API. RAMP data extend from January 2017 – now. For each item hosted by a RAMP participating IR, the data include details about the the item's position in a search result page, how many times it appears throughout a search result page, and whether the item URL received any clicks. Full documentation about RAMP data and data processing methods is available from the references below. We also have complete descriptive metadata in simple Dublin Core format for all items from 57 repositories that participate in RAMP. The data are currently stored in a SQL database that is not publicly accessible, but we have shared and rebuilt the database by sharing schema and CSV data dumps. Data Type: JSON, tabular, SQL database Data Size: 100+ GB, 400 million + rows in tabular format.
Outcome	Published papers, presentations, datasets, software code, and electronic lab notebooks
Milestone Timeline	Six months per selected project
References	Project website: https://rampanalytics.org Wheeler, J., Pham, NM., Arlitsch, K. *et al.* Impact factions: assessing the citation impact of different types of open access repositories. *Scientometrics* **127**, 4977–5003 (2022). https://doi.org/10.1007/s11192-022-04467-7 Arlitsch, Kenning, Jonathan Wheeler, Minh Thi Ngoc Pham, and Nikolaus Nova Parulian. "An analysis of use and performance data aggregated from 35 institutional repositories." *Online Information Review* (2020): https://www.emerald.com/insight/content/doi/10.1108/OIR-08-2020-0328/full/html Wheeler, Jon and Kenning Arlitsch. "Repository Analytics and Metrics Portal (RAMP) Workflow Documentation and Data Definition." (2020).

	https://digitalrepository.unm.edu/ulls_fsp/141 Wheeler, Jonathan et al. (2020), RAMP data subset, January 1 through May 31, 2019, Dryad, Dataset, https://doi.org/10.5061/dryad.fbg79cnr0 Complete data have been released per year, from 2017 through March 2021: • Wheeler, Jonathan; Arlitsch, Kenning (2021), Repository Analytics and Metrics Portal (RAMP) 2017 data , Dryad, Dataset, https://doi.org/10.5061/dryad.r7sqv9scf • Wheeler, Jonathan; Arlitsch, Kenning (2021), Repository Analytics and Metrics Portal (RAMP) 2018 data, Dryad, Dataset, https://doi.org/10.5061/dryad.ffbg79cvp • Wheeler, Jonathan; Arlitsch, Kenning (2021), Repository Analytics and Metrics Portal (RAMP) 2019 data, Dryad, Dataset, https://doi.org/10.5061/dryad.crjdfn342 • Wheeler, Jonathan; Arlitsch, Kenning (2021), Repository Analytics and Metrics Portal (RAMP) 2020 data, Dryad, Dataset, https://doi.org/10.5061/dryad.dv41ns1z4 • Wheeler, Jonathan; Arlitsch, Kenning (2021), Repository Analytics and Metrics Portal (RAMP) 2021 data, Dryad, Dataset, https://doi.org/10.5061/dryad.1rn8pk0tz