Texas Digital Library Dataverse R-Integration Project

Anna Pechenina, University of North Texas & Reid Boehm, University of Houston

Introduction

About this project As a part of the Open Research, Open Data, and other Open initiatives, the Texas Data Repository (TDR) is exploring the integration and use of R client for Dataverse repositories by way of the R dataverse package.

This integration supports Open Science and reproducible research by enabling researchers to discover, retrieve, and archive data and metadata using RStudio interface.

Poster Summary In this poster we share the preliminary learnings from the process of testing R-integration for the TDR Dataverse instance as it relates to research data:

- search,
- retreival, and
- archival.

We also share takeaways that we've made.

Dataverse Repository Defined

Dataverse is "... an open-source application for publishing, referencing, extracting and analyzing research data." [1]

One of the main goals of a Dataverse Repository is to incentivize data-sharing for social science researchers. It does so by enabling:

- author recognition
- persistent citation
- data discovery
- data preservation

The software is free, easy to use and maintain, and adheres to interoperability standards that are used by many digital libraries.

References

[1] D-Lib Magazine. The dataverse network®: an open-source application for sharing, discovering and preserving data. *D-lib Magazine*, 17(1/2), 2011.

Example code for Dataverse demo instance & TDR instance

Install Dataverse Package

Figure 1: Installation from CRAN fails because to package was pulled due to "policy violation"

Figure 2: Installation from GitHub succeeds

Server & API Key Set Up

Note: users must register with Dataverse's demo instance to get API token to make demo code work.

Figure 3: Use your unique Dataverse demo instance API Key in quotes

Figure 4: Use your unique TDR instance API Key in quotes. It will be different from Dataverse Demo site.
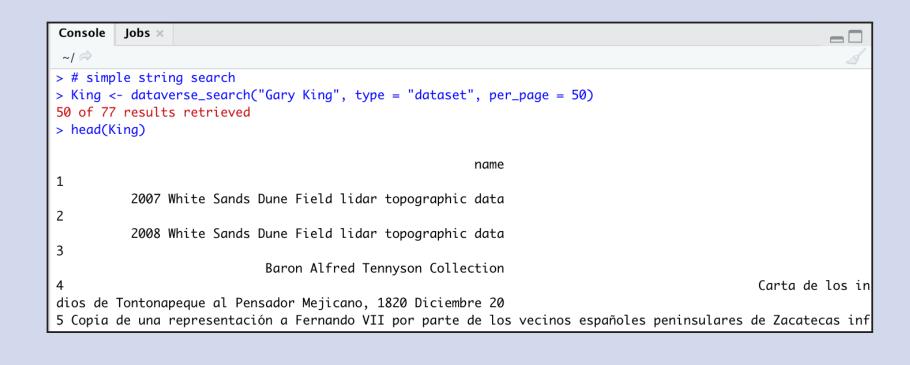
Search

Figure 5: Dataverse's demo site search for "Gary King" returns thousands of results and can be subset using per_page statement

Figure 6: TDR search for "Ju Dong Lee" also returns results and can be subset using per_page statement

Retreieve

Figure 7: Dataverse's demo site allows to download .tab files as Stata files using haven::read_dta.
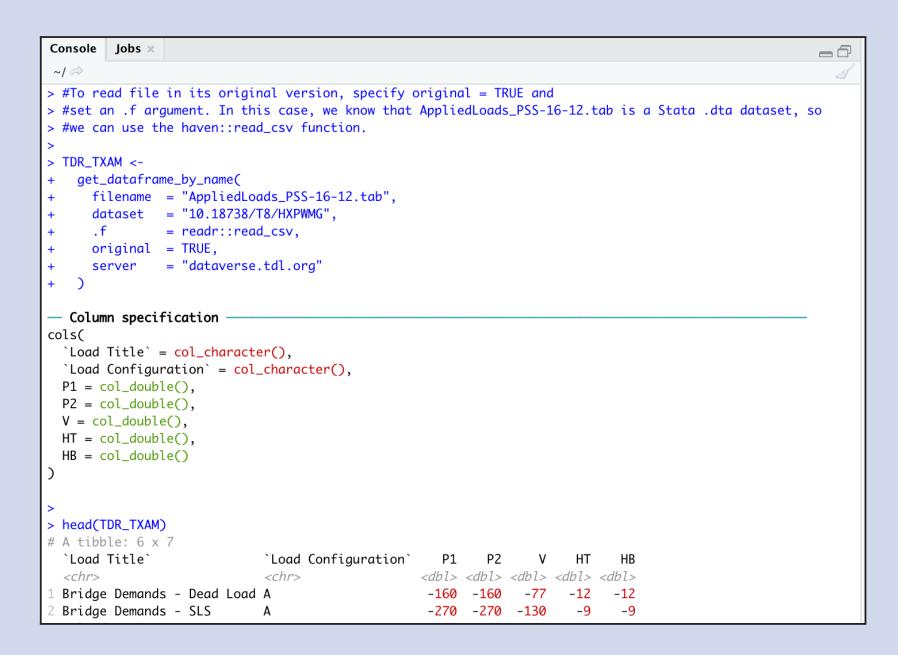

Figure 8: Note we used readr::read_csv to parse .dta file from TDR's demo site.

Archive

We encountered challenges working with the "Archive" function. We get the following error: "Error in service_document() : Forbidden (HTTP 403)." when we try to use the service_document command for Dataverse's demo site. We are still testing this functionality.

Figure 9

Takeaways

By engaging in this work learned the following things:

- Dataverse package availability through GitHub only (not on CRAN) makes it less reproducibility-friendly.
- Clear and complete documentation is critical for user success, especially beginners (e.g. knowing to use API token when using demo site).
- Absence of demos, specifically of Archival functionality, makes it harder to adapt the technology.

Conclusion

Dataverse R-integration is a promising solution for reproducible research that aligns directly with Open Research, Open Data, and other Open initiatives.

TDR will continue exploring Dataverse integrations to enable researchers perform data search, retrieval, and archival tasks.

Contact Information

Email annapechenina@my.unt.edu