How do levels of description affect discoverability of the Web Archives at the Library of Congress?

SAA Research Forum. August 2, 2019.
Carlyn Osborn, Library of Congress

Background

Since its creation in 2000, the Web Archiving program at the Library of Congress has created varying levels of description for different categories of their collections. These various levels of description have created an opportunity to observe possible associations that may exist between description levels and discoverability and access.

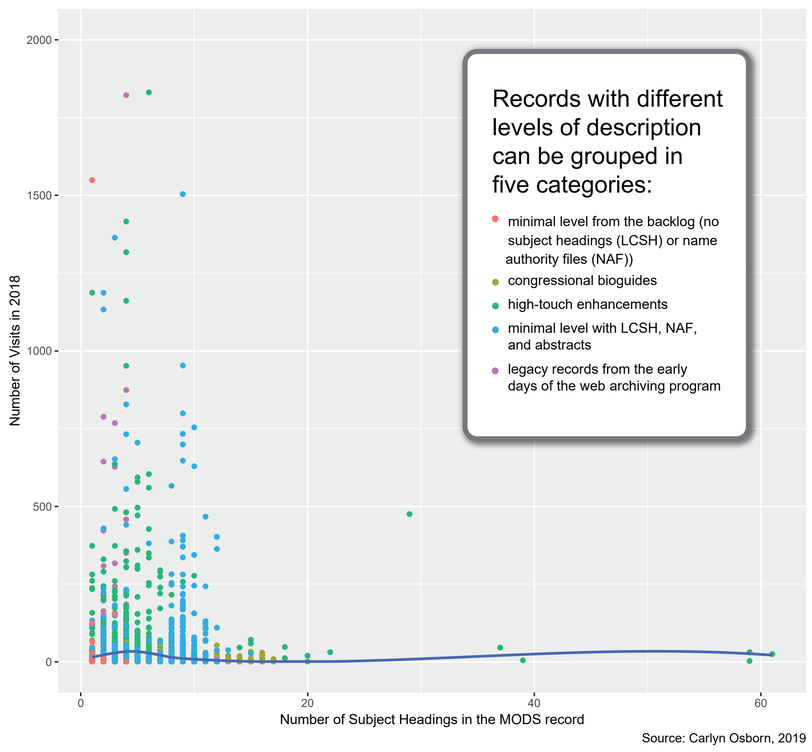
Questions

- Is there a relationship between the web archives' level of description and how many visits it received in 2018?
- Do subject headings make web archives more discoverable?
- Can we determine which parts of the descriptive record have the strongest relationship with annual visits?

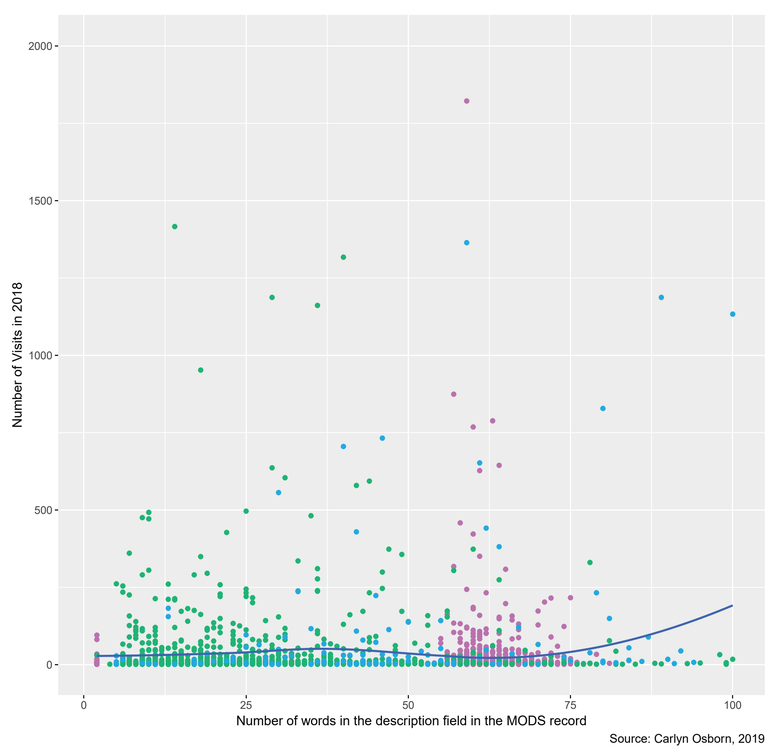
Method

- Use the Library's publicly available JSON API to pull 17,399 MODS records for web archives.
- Match all records to annual visitor data from the Web Analytics team.
- Normalize dataset and identify tagert metadata tags from the MODS.
- Perform exploratory data analysis on target variables to determine possible relationships between data.

Do subject headings relate to visits?



Does description length relate to visits?



Initial conclusions and next steps

- **47% of web archives with subject headings were visited.**
- **21% of web archives without subject headings were visited.**
- **In other words, if a web archive has subject headings, it's twice as likely to be visited at least once.**
- **We now have an independent, sustainable, and reproducible way to bulk access our own records.**
- **From the exploratory data analysis, we have a better overall sense of what our web archives look like.**
- **New variables selected for next research.**

Explore the web archives for yourself at loc.gov/websites!

More about Web Archives

Blog
https://bit.ly/2KwsPPK

Email
webcapture@loc.gov

Q&A
https://bit.ly/315Bguc

Website
https://bit.ly/2OvceDo

LIBRARY OF CONGRESS

Acknowledgments

Many thanks to
- The Digital Content Management Section at the LC (esp. Grace Thomas, Abbie Grotke, & Trevor Owens)
- Leah Ibraheem at the Library of Congress
- The 2019 SAA Research Forum Coordinators