

Tracking the Evolution of Web Archiving in the United States

Grace Thomas

Library of Congress
grth@loc.gov | @gracehthom

Matthew Farrell

Duke University

Edward McCain

University of Missouri

Maria Praetzellis

Internet Archive
maria@archive.org

Paige Walker

Boston College

Introduction & Methods

The information presented here has been selected from the recently published NDSA survey report: *Web Archiving in the United States: A 2017 Survey*.¹ From October 2 to November 20, 2017, a working group of volunteers representing five NDSA member institutions and interest groups conducted a survey of organizations in the United States actively involved in, or planning to start, programs to archive content from the Web. This effort builds upon and extends a broader effort begun in three earlier surveys, which the working group has conducted since 2011.

The goal of these surveys is to better understand the landscape of Web archiving activities in the United States by investigating the organizations involved, the history and scope of their Web archiving programs, the types of Web content being preserved, the tools and services being used, access and discovery systems being offered, and overall policies related to Web archiving programs.

The working group self-organized and reviewed previous surveys in order to identify necessary changes, introduce additional questions, and ensure consistency for longitudinal analysis. The survey was available October 2 - November 20, 2017 via SurveyMonkey. Respondents were solicited through mailing lists, blogs, social media, and other channels. After the survey closed, the working group reviewed a total of 159 responses and removed 39 responses that were tests, substantially incomplete, or from outside the United States, leaving a total of 119 responses for analysis.

Organizational Representation

The types of organizations represented in all four surveys has been consistent, with one notable increase in 2017. Organizations identifying as public libraries increased to 13% (15 of 119) of respondents from less than 3% of respondents in each of the previous surveys. The increase in public libraries can be traced to the Internet Archive's Community Webs program,² funded by the Institute of Museum and Library Services, the Kahle-Austin Foundation, and Archive-It, which provides Web archives training and technical services to public libraries for local history preservation.

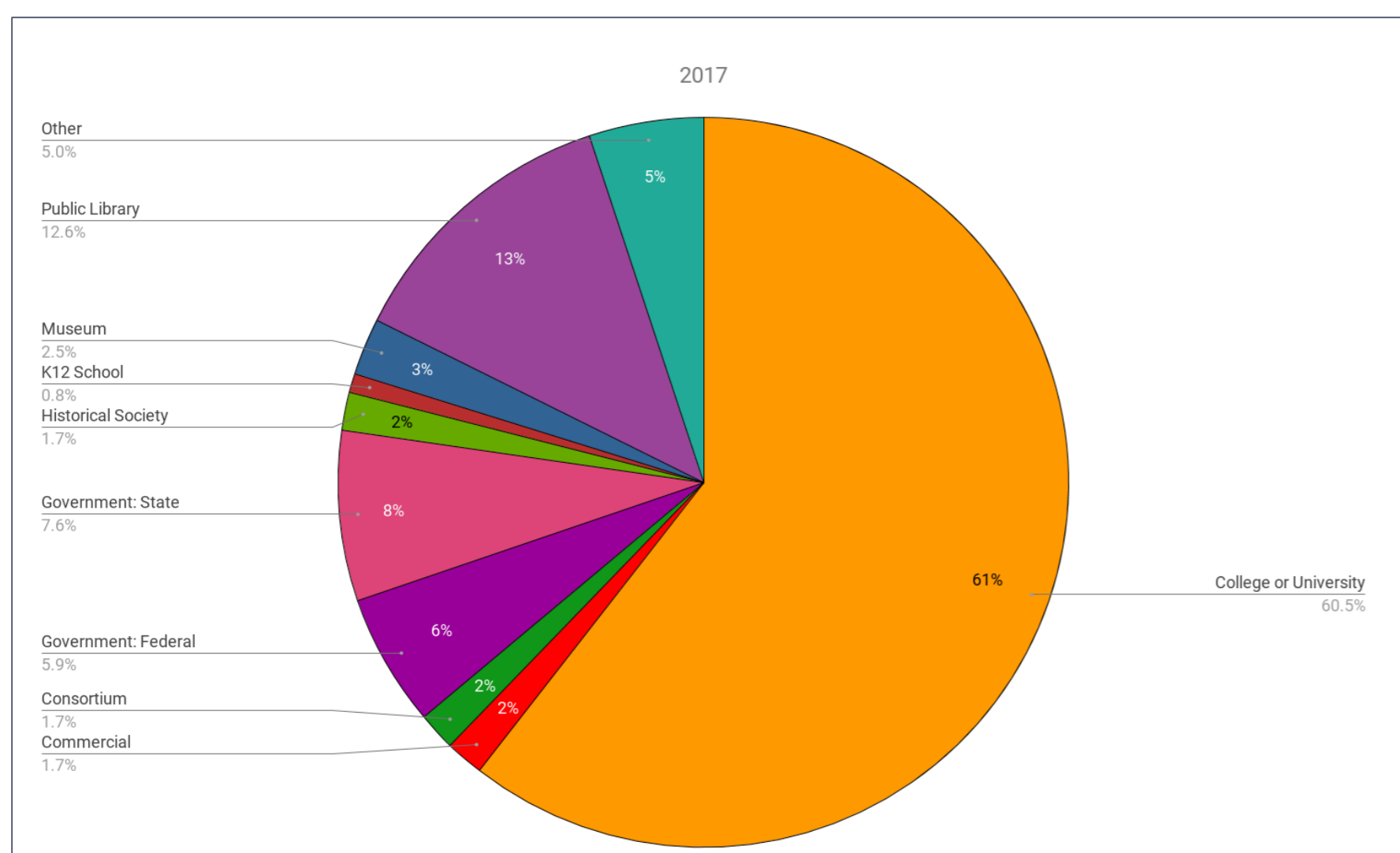


Figure 1. Responding organization type in 2017

Policy

Results from the 2017 survey indicate a growing comfort for archiving without permission or notification. Seventy percent (71 of 101) of institutions in 2017 did not seek permission or attempt to notify the content owner. Also, 91% (106 of 117) of respondents reported never receiving a takedown or stop crawling request. The 2017 survey introduced this question and future surveys will determine if this stark discrepancy is indeed a trend.

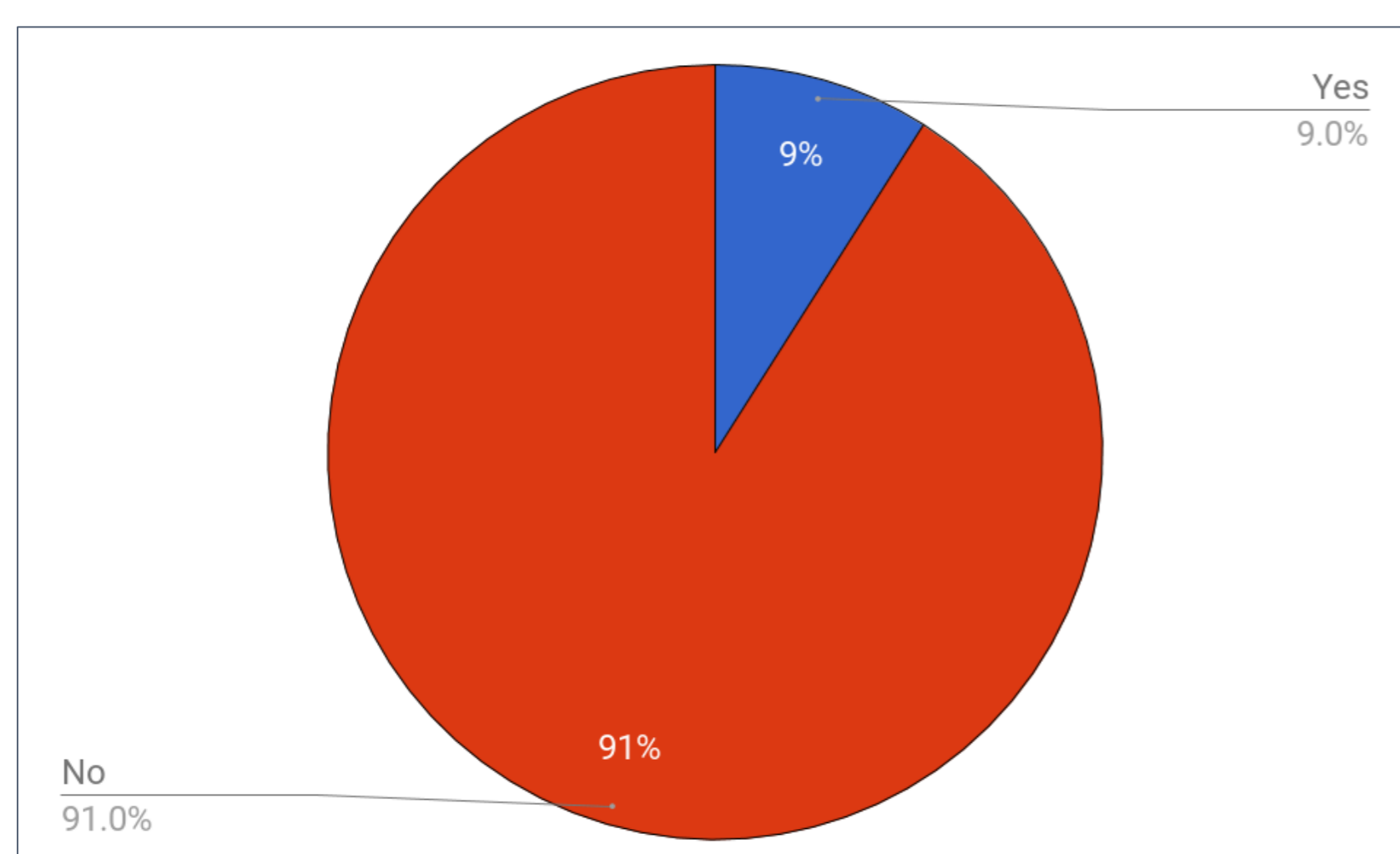


Figure 2. Organizations that received a take-down or stop crawling content request.

Technological Diversification

Few (20 of 108) organizations reported downloading their WARC files for local preservation or access, continuing a trend denoted in previous surveys.

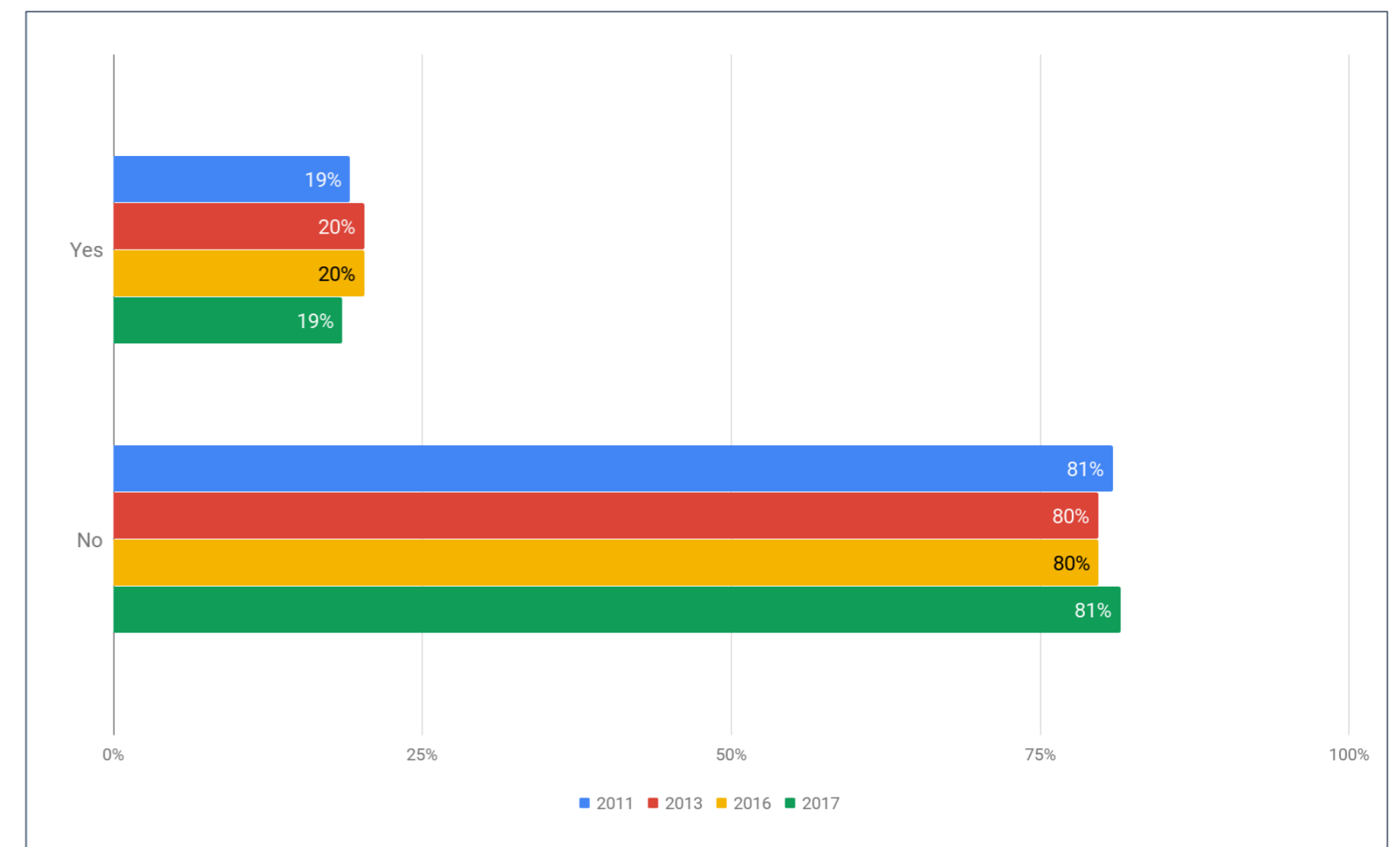


Figure 3. Organizations that chose to replicate data

The majority of responding organizations reported utilizing multiple tools to capture content successfully. The diversification of tools utilized is exemplified by the development and adoption of the Web capture service, Webrecorder. A notable 51% (23 of 45) of organizations reported using Webrecorder, which was publicly launched in 2016 as a browser-based tool to allow for the capture of content difficult to capture via traditional link-based crawling. Archive-It continued to be the preferred external service for harvesting Web content, with 94% (97 of 103) of respondents using this service.

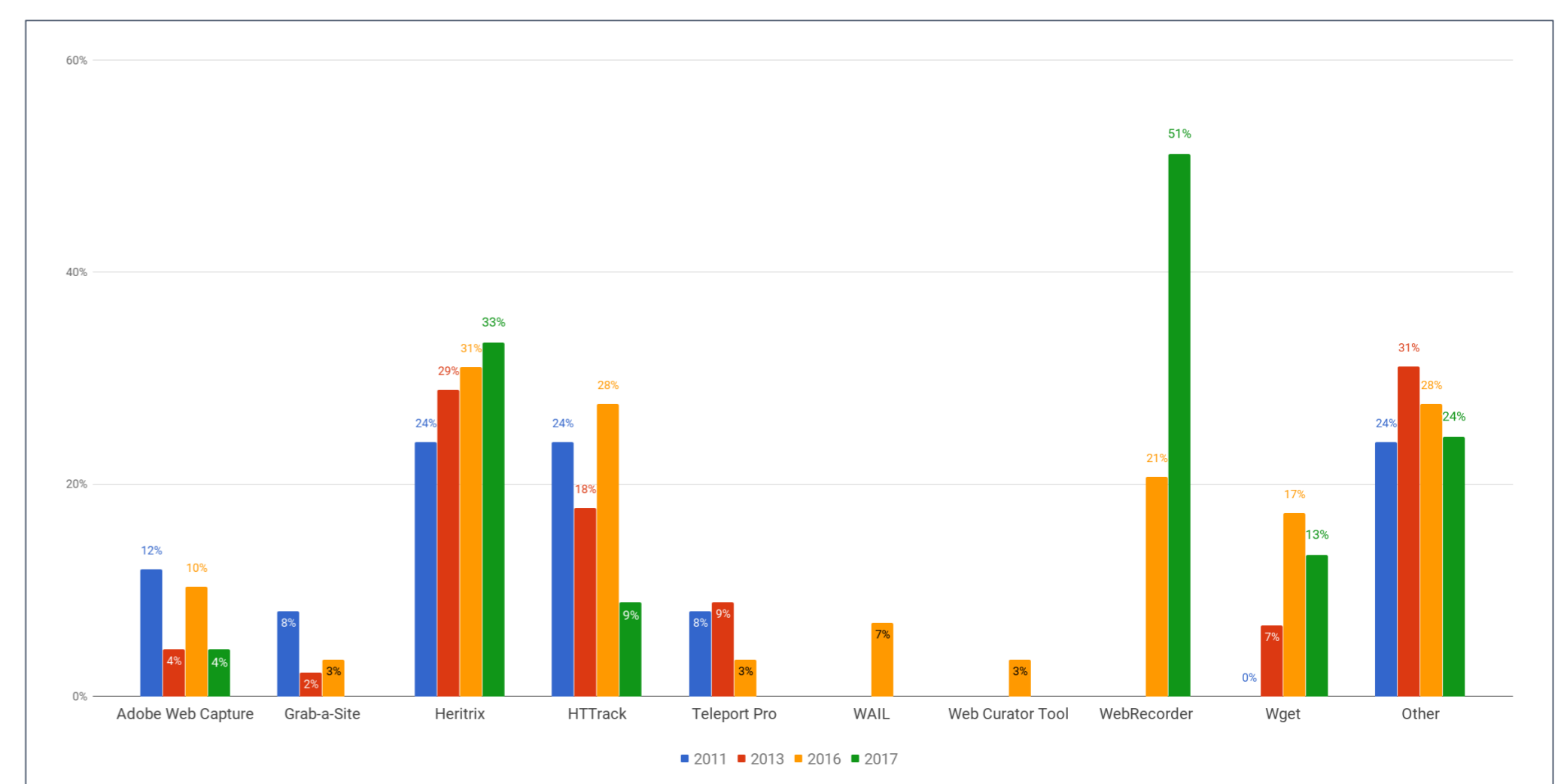


Figure 4. Tools used for local capture.

Reference

¹ Access the report at: <https://osf.io/ht6ay/>

² Jefferson Bailey, "IMLS Grant to Advance Web Archiving in Public Libraries," Internet Archive Blogs, July 18, 2017, accessed August 22, 2018, <https://blog.archive.org/2017/07/18/imls-grant-to-advance-Web-archiving-in-public-libraries/>.

Acknowledgements

We would like to thank all of the participants in the 2017 Web Archiving Survey for their time and willingness to share information about their programs with the broader Web archiving community. Many thanks to NDSA, CLIR, and others for their support and feedback throughout the project, especially the NDSA Coordinating Committee, Jefferson Bailey, Jesse Johnston, Carol Kussmann, Christie Moffatt, and Nicholas Taylor for their editorial assistance.

