Tracing the Use of Research Resources Using Persistent Citable Identifiers

Matthew Mayernik[1], Keith Maull[1], David Hart[2]

1. NCAR Library, 2. NCAR Computational & Information Systems Laboratory

National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR)

Background & Problem

- Many organizations and repositories are assigning persistent web-accessible identifiers to research resources, e.g. data, software, services, grey literature
- The use of persistent identifiers can increase the traceability and discoverability of scientific resources [3]
- Without consistent assignment and use of persistent identifiers, it is difficult to assess and understand the impact of individual resources [1]

Research Questions

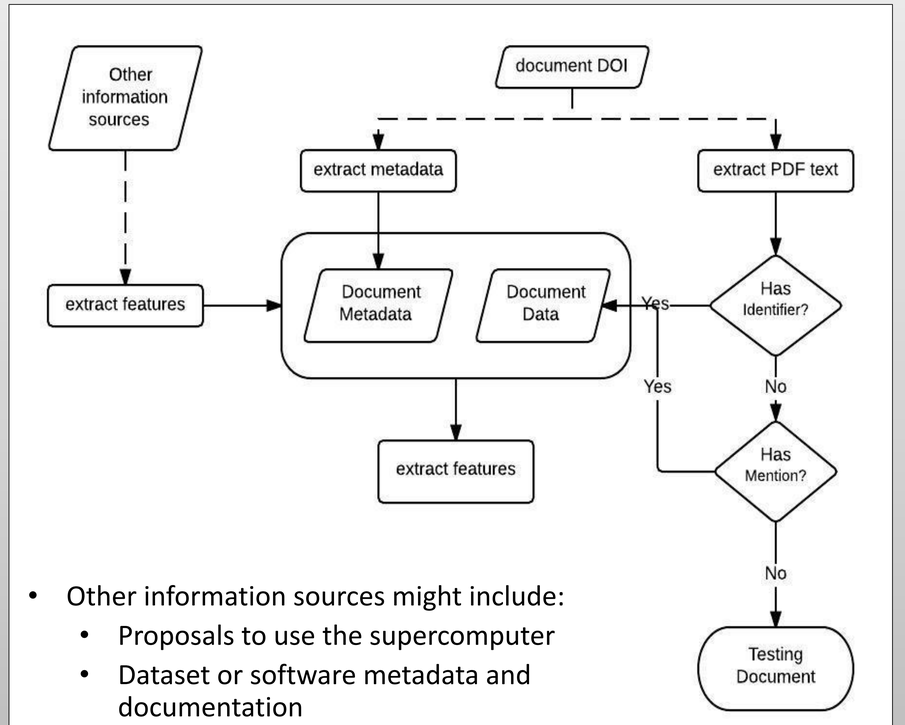
1. Have the overall citations of research resources increased in relation to the assignment of persistent IDs to those resources?
2. What are the characteristics of such citations? Do users create more consistent citations when persistent IDs are assigned to resources?
3. What can we learn from tools already developed that would help us build the necessary computational platforms and algorithms to expand the capacity and success of research to understand the scientific impact of resources? [2, 4]
4. How can open platforms and tools facilitate the tracing of research resources?

Baseline data and corpus gathering – Manual searches

- Manual Google Scholar searches (Nov.-Dec. 2014)
- Three initial assessment cases:
  - Data – North American Regional Climate Change Assessment Program (NARCCAP) data set, DOI assigned 2012-05-03
  - Software – NCAR Command Language, DOI assigned 2012-04-10
  - Facility – Yellowstone Supercomputer, ARK assigned 2012-05-21

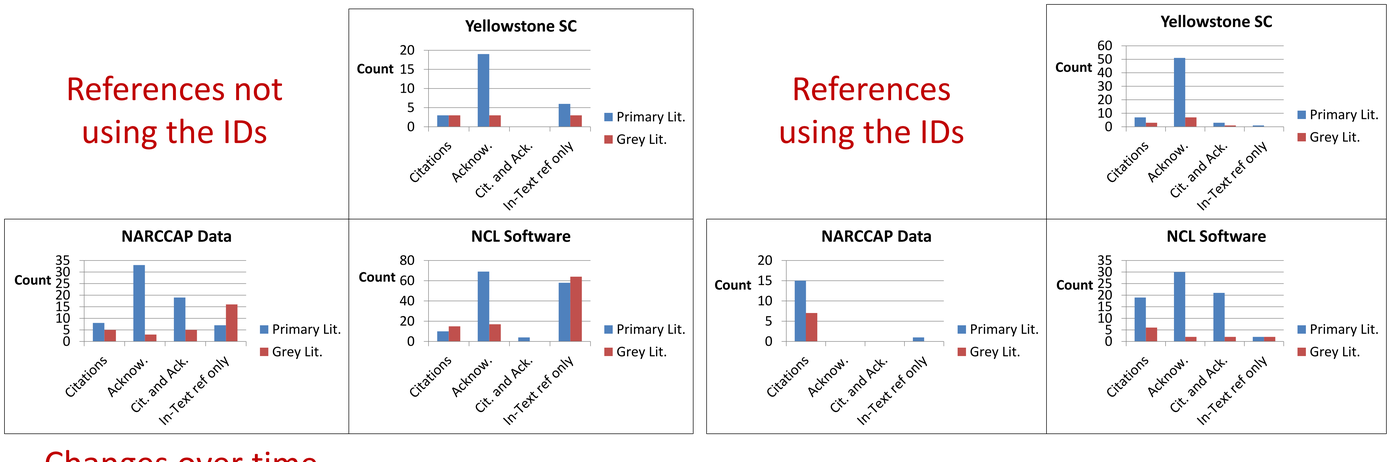
Automating the Tracing Process

- Need to identify explicitly or potentially relevant documents
- When looking for references to particular research resources, documents can be classified into three sets:
  - Set 1: Have explicit formal citation to the resource
  - Set 2: Have mention of the resources without citation
  - Set 3: Have no mention or citation of the resource
- Features of Sets 1 and 2 should be useful in assigning a probability function to the documents in Set 3
- Use probabilistic approach to help focus human effort on cases where human review is required
  - Simple relevant/not relevant cases can be addressed by computer
  - Complex cases can be analyzed in detail by human experts



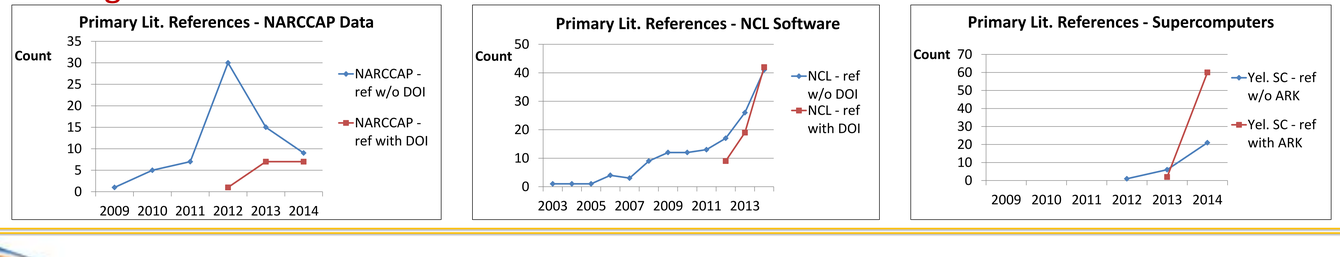
- Other information sources might include:
  - Proposals to use the supercomputer
  - Dataset or software metadata and documentation

References not using the IDs



References using the IDs

Changes over time



Discussion

- Manual search process is very useful but limited in scope
  - Papers with IDs are much easier to find
  - Identifiers are being used, but inconsistently spread throughout articles
  - Shifts in practices over time vary from resource to resource
- Automation is hindered by variability in search tools, journal websites & policies, and individual citation/reference practices
  - Subscription firewalls restrict accessibility of potentially relevant articles
  - Maintaining reliable parsing is challenging due to mutability of web environments
- Outreach, advocacy, recommendations to users are critical once persistent IDs are assigned
- Role of journal editorial policies unclear

References

1. Mayernik, M.S. (2013). Bridging data lifecycles: Tracking data use via data citations workshop report. NCAR Technical Note, NCAR/TN-494+PROC, Boulder, CO: National Center for Atmospheric Research (NCAR). http://doi.org/10.5065/D6PZ56TX
2. Patton, R., Stahl, C. G., Potok, T. E., & Wells, J. C. (2012). Identification of user facility related publications. D-Lib Magazine, 18(7/8). http://doi.org/10.1045/july2012-patton
3. Socha, Y. M., and Coauthors. 2013. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. Data Science Journal, 12, CIDCR1-CIDCR75. http://doi.org/10.2481/dsj.OSOM13-043
4. Wang, F., et al. (2014). Towards a Scientific Impact Measuring Framework for Large Computing Facilities - a Case Study on XSEDE . In Proceedings of the 2014 Annual Conference on eXtreme Science and Engineering Discovery Environment (pp. 1–8). http://doi.org/10.1145/2616498.2616507

Acknowledgments: NCAR is funded by the U.S. National Science Foundation (NSF). This project is supported by NSF Grant #1448480.