Scholarly: An open, freely accessible dataset of the academic citation network

Harry Rybacki[1, 3], Joshua Carp[2, 3] and Jeffrey Spies[3]

[1] University of North Carolina at Greensboro [2] University of Michigan [3] Center for Open Science

What is Scholarly?

Scholarly is a open source, collaborative project being developed at the Center for Open Science. The primary purpose of Scholarly is to provide the with a free, open, and comprehensive dataset containing meta-data for academic citations as well as corresponding references. This dataset will allow the public to access, analyze, and distribute academic citation meta-data without restriction.

The citation network

The citation network is a directed acyclic graph. Graph vertices (nodes) represent unique articles. Graph edges (paths) represent one article citing another.

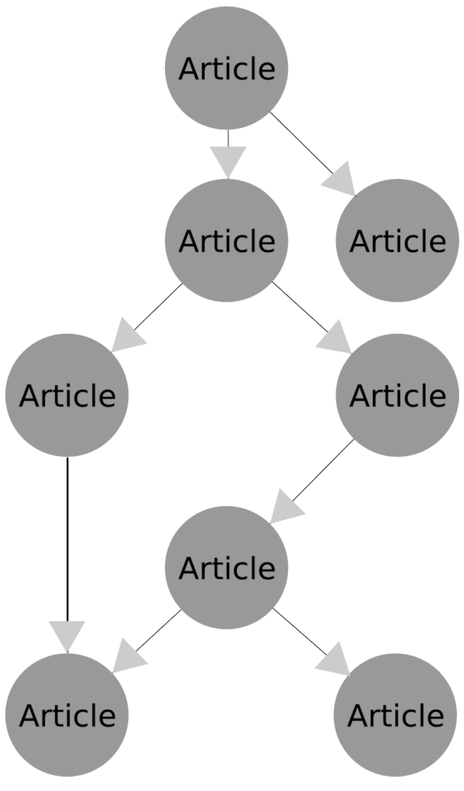

Figure 1: Paths between articles represent citations and articles cite n number of children articles.

Current state of scholarly metadata

Existing tools use similar data:

- Google Scholar
- Microsoft Academic
- Web of Science
- Psycinfo

However, these datasets are closed:

- No programmatic access to data –violation of terms of service
- Only a small subset of data is publicly available

The problem and our approach

Beyond the size of the corpus of academic literature, the data required to build this dataset comes from a wide variety of sources (e.g., APA and ACM) and is delivered in various formats (e.g., XML, HTML, text documents, and PDF files). Project Scholarly is developing a set of tools to collect, manage, and distribute the data.

Front end

1. **Citelet:** A Chrome extension that allows users to effortlessly send citation metadata to our servers as they browse articles at publisher websites.
2. **Citebin**: A clean and easy-to-use web interface that provides users with the ability to either enter metadata manually or simply copy and paste it into a form.
3. **API:** For individuals that would like to programmatically contribute data, we have constructed an application programming interface (API) to which they can either send formatted documents or unparsed citations.

Together, these tools will feed into a set of validation, parsing, and conflict management engines (both automatic and crowd-sourced) before being entered into the production dataset.

Back end

1. **Validators and Parsers**: After arriving from a wide variety of sources, data must be checked for validity and parsed into a consistent standard derived from CSL (Citation Style Language).
2. **Databases**: Raw data enters the staging database. After being validated and parsed it moves into the conflict database. Entries are crosschecked for similarities and placed into conflict groups. Finally, conflicts are resolved with internal algorithms as well as external crowd-sourced tools before moving to the production database.
3. **Crowd-sourcing**: Before entering the production database, conflicts among the data must be resolved. One of the tools in development will allow users to help resolve conflicts via a simple, intuitive online user interface.

The future with Scholarly

An open dataset will allow for innovation:

- New search engines built by developers
- Research tools built to supplement the research workflow
- Recommendation engines
  - Find researches working in similar areas
  - Find articles related to your work
- Citation network analysis
- Next generation alt-metrics

How small is the world?

Citation metadata will allow for meta-science reserach. E.g., Does the scholarly citation network have small world network distributional properties? In other words, does this network have a high clustering coefficient and are most nodes connected to other nodes with a short path cost?

Get involved

1. **J**oin us at the Open Science Framework sprints on June 28-29: We have a wide array of projects for contributors of all skill levels.
2. **C**ontribute to the project:
   - **Citelet:** https://github.com/jmcarp/citelet
   - **Citebin:** https://github.com/hrybacki/crowd-scholar
   - **Conflict Management:** https://github.com/hrybacki/conflict-management-engine

More information

Jeffrey Spies' talk: The Open Science Framework: Improving, by Opening, Science – Thurs., June 27th in the Reproducible Science Room: 106

Links:
http://centerforopenscience.org
http://opensciencefframework.org

Contact Information

- Web: http://rybacki.io
- Email: hrybacki@gmail.com

CENTER FOR OPEN SCIENCE