Colandr: an open-source, open-access tool using machine learning for evidence synthesis

For more project details contact Caitlin Augustin // caitlin@datakind.org

ABSTRACT

In recent years, there have been many calls for more evidence-informed decision making. Evidence syntheses, particularly systematic mapping and systematic reviews, are rigorous methods used to garner insight from scientific literature while minimizing bias and maximizing transparency, objectivity and comprehensiveness. Collecting and filtering evidence for those tools, however, is time and labor-intensive. The common practice for developing this evidence base currently uses simple (keyword) searches of research papers to shrink their search space, and then manual screens of the remaining papers, often numbering in the thousands, to classify and extract relevant information. While evidence synthesis methods are becoming more prevalent, the high resource cost required has been a major disincentive to producing high quality and updated resources despite their critical value.

Colandr – a tool created and maintained by volunteers for social sector organizations looking to use limited resources to maximize evidence synthesis – is an example of partnerships that can flourish in the data for good space. As one of the few tools that are widely available both cost-free and containing computer-assisted decision-making, it has been cited by users as a key resource in research in the conservation domain as well as domains such as medicine and political science.

Background

The SNAPP research team was focused on three questions
- How do we find evidence?
- How do we communicate evidence?
- How do we use evidence?

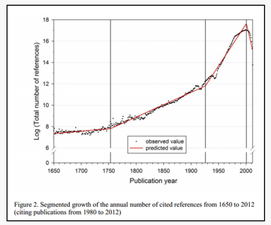
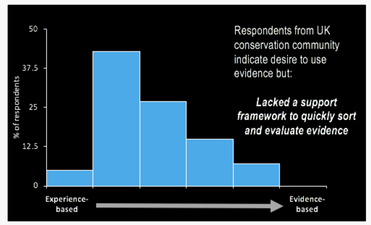
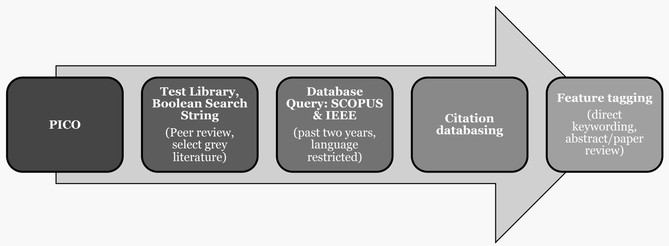


Figure 2. Segmented growth of the annual number of cited references from 1650 to 2012 (citing publications from 1980 to 2012)

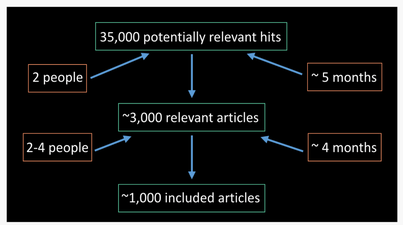


Global scientific output doubles every nine years[1] and practitioners want to use evidence in conservation decision-making but lack access and opportunity.[2]

Practitioners need **access to research insights** from academic and grey literature for evidence-based decision making Researchers need a framework to follow to create these resources



The systematic map process is a framework for formal assessment of current, relevant literature

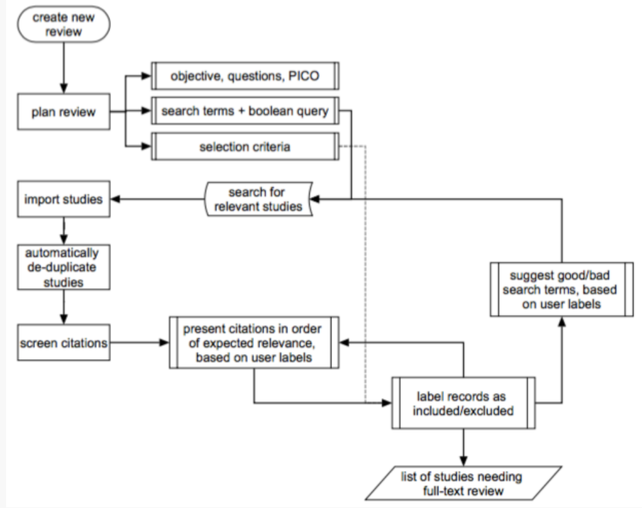


However, the common process is manual and exceptionally labor intensive.[3]

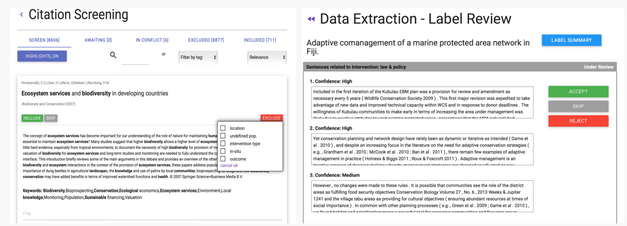
Approach

The DataKind team developed colandrapp.com, an open-source, open-access tool for computer-assisted systematic mapping.

Colandr is built on two systems:
1. Distributional word vectors as features for a support vector classifier that predicts inclusion or exclusion; use confidence of that classification as expected relevance
2. "Named Entity Recognition" system to find mentioned locations in the document and suggest these as metadata labels. It uses global vectors for word representation (GloVe) and logistic regression to train a model of ranker-tags



Colandr's workflow



colandrapp.com screens

Discussion

As only one of only five evidence synthesis tools using ML/AI[4] and the only tool that is open source, open-access[5], in approximately one year of use:
- Over 200 unique registered users, 76 of which are academic users, 30 of which are organizational users
- 274 reviews created spanning topics of conservation, medicine, education, climate change, marine stewardship and community engagement
- Multi-continent users: users from countries in North America, Europe, and Asia
- Over 100 attendees at multiple training events
- Research community at colandrcommunity.com

	Format: Ease of using specific GUI vs. non-specific formats	Error: Catching missed references, mis- assigned tags, duplicates	Efficiency: How many citations screened to find 100 included?
Case 1: Conservation & human well-being (McKinnon et al. 2016)	Version control issues when screening in Microsoft Excel. Oftentimes would crash the program. Multiple columns for exclusion criteria made for lots of unnecessary scrolling back and forth	Many duplicates still cropped up even after data was extracted. The deduplication function in Colandr allowed for us to find duplicates faster than by eye. Colandr also suggested tags for articles that upon closer read, were in fact an appropriate tag for that paper that we had misassigned by hand.	Colandr: 250 Manual: 1436
Case 2: Forests & poverty (Cheng et al. 2017)	Screening in EPPI Reviewer is comparative in format, allowing for multiple users and structured format to standardize criteria. However, costs for EPPI quickly rose as we added members to the review team.	Colandr allowed for quicker identification of key sentences that could lead to insight into document tags. Rather than reading through often dense text, it was very useful and efficient to view suggested sentences. While some of the these sentences were not always helpful, having them collated in one place streamlined the process.	Colandr: 167 Manual: 407
Case 3: Synergies, tradeoffs, equity in marine conservation	The GUI facilitated faster title and abstract screening with: clear text layout, highlighted keywords, radio buttons to select reasons for exclusion, and smooth transitions from one entry to another. Also facilitated screening on mobile devices.	Colandr's deduplication function eliminated the need for the reviewer to do this tedious process manually. In total, the app identified 70 duplicates and only missed 7 (90% success rate).	Colandr: <568 Manual: NA

Early research results

References

1. Van Noorden, Richard. "Global scientific output doubles every nine years." Nature News Blog (2014).
2. Pullin, Andrew S., et al. "Do conservation managers use scientific evidence to support their decision-making?." Biological conservation 119.2 (2004): 245-252.
3. McKinnon, Madeleine C., et al. "Sustainability: Map the evidence." Nature News 528.7581 (2015): 185.
4. Kohl, Christian, et al. "Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools." Environmental Evidence 7.1 (2018): 8.
5. Cheng, S. H., et al. "Using machine learning to advance synthesis and use of conservation and environmental evidence." Conservation Biology (2018).

CONSERVATION INTERNATIONAL + DataKind + SNAPP