A semi-automated method to track dataset reuse in biomedicine

Heather Piwowar
National Evolutionary Synthesis Center

Why track data reuse?

- Understanding how research datasets are reused after their original collection facilitates **rewarding** the data collectors, **measuring** benefits, and **monitoring** policy impact

Why is tracking reuse difficult?

- There are no standard ways to cite datasets
- Sometimes data is acknowledged through a citation to the original data collection paper, but often data is just acknowledged through its database unique identifier, called an *accession number*
- Because accession numbers are embedded within full text, querying and disambiguating the context across a discipline is complex and time consuming

What is unique about this approach?

- The proposed method takes advantage of the NCBI PubMed® and Entrez tools to query full text, exclude known data creation papers, annotate findings and extrapolate the results. This facilitates **tracking thousands of datasets**.

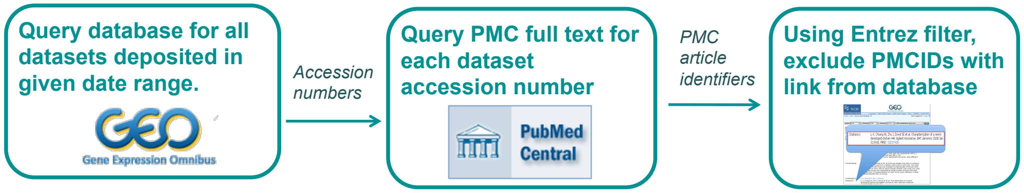
Does the method find all reuses?
Does it mistakenly identify mentions as reuses?

- Preliminary validation compared the results of this method to the GEO third-party usage page
- Found 256 of the 618 reuse articles listed by GEO staff (41%) and 802 articles not on the GEO list.
- Comprehensive evaluation underway

Is it generalizable?

This method is most applicable for tracking the reuse of datatypes that:

- are well represented in the NCBI Entrez databases
- have a well-defined primary data repository
- have a unique accession number format
- have a community norm of citing accession numbers upon reuse



Query database for all datasets deposited in given date range.

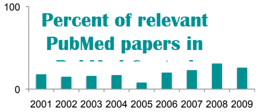
Accession numbers

Query PMC full text for each dataset accession number

PMC article identifiers

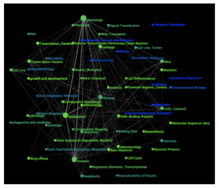
Using Entrez filter, exclude PMCIDs with link from database

Citation(s) Li X, Chiang HI, Zhu J, Dowd SE et al. Characterization of a newly developed chicken 44K Agilent microarray. BMC Genomics 2008 Jan 31;9:60. PMID: 18237426

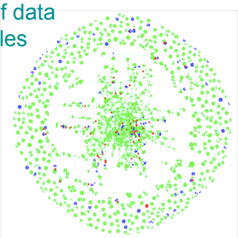


Extrapolate the amount of reuse to all of PubMed

MeSH analysis of reuse patterns



Authorship analysis of data creation and reuse roles



Future work

- Author disambiguation with Torvik and Smalheiser's Author-ity service
- Semi-automated identification of rogue data-creation articles using full-text query

Acknowledgements

This research was funded conducted as a DataONE postdoc.
DataONE is funded by a Cooperative Agreement through the NSF DataNET program (OCI-0830944).

hpiwowar@nescent.org @researchremix github

NSF DataONE DRYAD NESCent National Evolutionary Synthesis Center the Biodiversity Research Centre UBC The University of British Columbia