DATA DESCRIPTION

Schema for patent-paper citations

The main output file is called *_pcs_oa.csv* and is comma-separated. Each record contains a patent-to-article citation (or, in the case of a citation appearing both on the front-page and in the body text, two citations).

Contents of *_pcs_oa.csv*.

Variable	Type	Notes
oaid	numeric	Unique identifier for each paper in the Microsoft Academic Graph
patent	string	Only patents for which our algorithm established a PCS linkage are included. The format is as follows. The first two characters represent the country of the patent office, e.g. us for USPTO and ep for EPO. Next is a hyphen (-), followed by the patent number.
wherefound	string	**frontonly**, **bodyonly**, or **both** (i.e., both on the front page of the patent, and also in the body text)
confscore	numeric	Assigned confidence score to the match.
self	binary	Indicates whether at least one author on the paper was also an inventor on the patent.
reftype	string	**App** = from applicant **Exm**=from examiner (Note: non-USPTO refs are examiner unless otherwise indicated in the reference.) **Unk** = if unspecified in the unstructured reference (Note: most pre-2006 USPTO references are unkown.)

Differences from previous pcs_mag_doi_pmid.tsv include a) comma separated not tab-separated b) no ‘uspto’ flag c) reordering of fields d) simplified self-citation flag e) no PMID or DOI, can merge these from OpenAlex files below f) ‘oaid’ for OpenAlex instead of ‘magid’ for the Microsoft Academic Graph.

“Shorthand” files for counts of patent-paper citations

We also include two “shorthand” files that collapse the various types of citations per paper *(_pcs_countsbypaper.csv*, for forward citations from patents) and per patent (*_pcs_countsbypatent.*csv, for backward citations to papers). The variable names in the files are identical and represent various types of cites in the below table:

	Count of all citations	Count of self-citations (at least one author is an inventor)	Count of external citations (no overlapping authors / inventors)	Count of front-page citations	Count of body („in-text“) citations
ncit	•				
ncitself		•			
ncitext			•		
ncitfront				•	
ncitbody					•
ncitboth				•	•
ncitfrontself		•		•	
ncitbodyself		•			•
ncitbothself		•		•	•
ncitfrontext			•	•	
ncitbodyext			•		•
ncitbothext			•	•	•

Schema for patent-paper pairs

The main output file is called *_patent_paper_pairs.tsv* and is tab-separated. Each record contains a patent-to-article citation established by our algorithm.

Contents of *_patent_paper_pairs.tsv*.

Variable	Type	Notes
ppp_score	numeric	Assigned confidence score to the patent paper pair, 1-4 where 4 is highest.
paperid	numeric	Unique identifier for each paper in the Microsoft Academic Graph
patent	string	Only patents for which our algorithm established patent paper pair are included. The format is as follows. The first two characters represent the country of the patent office, e.g. US for USPTO. Note that *all* patents in this file are currently USPTO. Next is a hyphen (-), followed by the patent number.
daysdiffcont	numeric	Number of days between the application date of the oldest parent of the patent (found in the continuity_parents file published by PatEx) and the publication date of the paper.
all_patents_for_the_same_paper	string	If a paper is mapped with multiple patents, it indicates whether all the patents share the same parents, titles, abstracts, application and/or grant dates. Each criterion is represented as a string. A blank value for this variable indicates that not all of the patents to which this paper is mapped can be labeled identical.

Validation file for patent-paper citations

The set of known-good patent-to-article citations is called *bodytextknowngood.tsv* and is tab-separated. Each of these patent-to-paper citations was checked by two research assistants. Note that this work was conducted using the Microsoft Academic Graph, so this file uses MAG IDs. Add a 'W' to MAGID to find the corresponding OpenAlex ID.

Contents of *bodytextknowngood.tsv*.

Variable	Type	Notes
patent	string	Patent in which the in-text reference was found. Each reference
magid	numeric	Unique identifier for the paper cited in the Microsoft Academic Graph
doi	string	Digital Object Identifier as provided by Microsoft, if available
pmid	numeric	PubMed ID as provided by Microsoft, if available

Files for certain fields from OpenAlex metadata

Also available is a series of files with metadata regarding not just the references but *all* papers in the end-2022 release of OpenAlex (OA). They are compressed using the 'zip' utility under Unix CentOS5. Those who prefer to download the original OpenAlex data directly can do so from openalex.org. Note however that some of the original OA files are several dozen gigabytes in size, whereas we have partitioned the files into smaller pieces for convenience. The first set of files contain direct metadata for papers in OA.

Filename	Variables	Notes
paperyear	oaid, paperyear	
papervolisspages	oaid, papervolume, paperissue, paper1stpage, paperlastpage	Issue and pages are sometimes blank. First page is available more often than last page.
paperdoi	oaid, doi	DOI is not available for every paper in OA
paperpmid	oaid, pmid	PMID is not available for every paper in OA
paperauthoridorderaffiliation	oaid, firstmidlast, authorid, affiliation	Authororder is not numerical but first,mid,last Affiliation may either be an Affiliation ID, which starts with I followed by a series of numbers, or an textual affiliation *without* an ID
paperjournalid	oaid, journalid	
papercitesfrompapers	oaid, numcitesfrompapers	Count of forward citations from other scientific articles in OpenAlex

The second set of files contains the string values for indirect metadata identifiers:

Filename	Variables	Notes
authoridname	authorid, authorname	
journalidname	journalid, journalname	
affiliationidnametype	affiliationid, name, type	Type is assigned by OpenAlex.