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 OpenAlex
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.
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.)
self	string	Indicates whether at least one author on the paper was also an inventor on the patent. Values are „selfcite“ „nonselfcite“ „unkselfcite“ where the third value means it is unknown.

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.

Schema for patent-paper pairs

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

Contents of *_ patent_paper_pairs.csv*.

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 OpenAlex
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.