Documentation

This document describes data collection, processing, and different open access data files related to the text of scientific publications from Microsoft Academic Graph (MAG) (now OpenAlex). If you use the code or data, please cite the following paper:

Arts S, Melluso N, Veugelers R (2023). Beyond citations: Text-based metrics for assessing novelty and its impact in scientific publications. https://doi.org/10.48550/arXiv.2309.16437

Paper, data and code

Paper: https://doi.org/10.48550/arXiv.2309.16437

Data: https://zenodo.org/record/8283353 (DOI: 10.5281/zenodo.8283353)

Code: https://github.com/nicolamelluso/science-novelty

Data processing

We collect all publications from the latest version of Microsoft Academic Graph (MAG), dated 06.12.2021 via Azure Data Share (Sinha et al., 2015)[1]. MAG contains about 270,000,000 records including books, book chapters, conference proceedings, datasets, journal articles, papers, repositories, theses and preprints. We restrict the sample to journal publications and published conference proceedings (103,117,246 records). We further remove 3,571,940 publications with duplicate titles, abstracts or Digital Object Identifiers (DOI), 5,838,367 publications with non-English source[2], 11,881,805 publications with a non-English title, 3,303,938 publications with a non-English abstract[3], and 29,320 records of retracted publications. The final sample consists of 83,490,800 publications from 1800 to 2020.

We identify two issues concerning the publication date in MAG. First, the exact day of publication may be unclear for 16,267,071 publications which report the 1st of January as date of publication. Given that this is a significant share of the sample, we double check these publication dates with both CrossRef and Scopus using the DOI of the publication (Martín- Martín et al., 2021; Visser et al., 2021). 63% of records with 1st January as publication date have no DOI and cannot be matched to either CrossRef or Scopus. Among the matched records, 87% also have the 1st of January as publication date in CrossRef or Scopus while the remaining 13% have a different date. We keep the 1st of January as date of publication in case of agreement between MAG and CrossRef or Scopus. In case of disagreement, we replace the publication date in MAG with the publication date in Scopus and/or CrossRef. In case of disagreement between Scopus and CrossRef, we use the publication date from CrossRef, which turned out to be more accurate after a manual inspection of a randomly selected sample of 100 papers. Finally, we remove the publications that exist only in MAG (and not in

[1] Microsoft (2022), Microsoft Academic Graph, URL: https://www.microsoft.com/en- us/research/project/microsoft-academic-graph/, last accessed 31.10.2022

[2] For non-English sources, we use the language information about the URL, where the record is retrieved, available in MAG's Table 'Paper Urls' from the official data schema.

[3] We use fasttext, a language detection python package to determine the language of the record (Joulin et al., 2016).

CrossRef or Scopus) with the indicative publication date as 1st January, due to the uncertainty in their publication date. The second issue with the recorded dates in MAG is that there is both an 'online' date and a publication in print date that are not necessarily the same. Publications are made available online on the publisher's website before the date of the print issue or after the date of the print issue for older publications. We take the earliest date as the date of publication because this date is the closest to the scientific discovery.

Next, we process the title and abstracts of these publications. The abstract is only available as an inverted index in MAG, which we convert back to the original abstract[4]. Notice there is only a title and no abstract available for 29% of the publications. Next, we concatenate title and abstract. Afterwards, we process the text with the following steps: tokenization, lemmatization, stop word removal, chunking, baseline removal, and vectorization. First, we lowercase the text and tokenize it to words (or unigrams)[5]. Second, we apply lemmatization to each word using the WordNetLemmatizer from the NLTK library in Python so that we are left with a collection of lemmas representing the scientific content of the publication. Third, we remove: natural stop words[6], a manually compiled list of 33,764 very common words unrelated to the scientific content of a publication[7], words composed only by numbers, one character words and punctuations. The final list of removed stop words consists of 67,150 words. In addition, for each publication, we perform chunking by extracting bigrams (two consecutive words) and trigrams (three consecutive words) using the following strategies: for bigrams we keep the ones without any stop word and for trigrams we only keep those without any stop words or with a stop word in the middle of the trigram.

We use all publications published between 1880 and 1900 to construct a baseline dictionary and restrict the analysis to papers published between 1901 and 2020 (n=72,245,396). The entire cleaned vocabulary contains 20,953,136 unique keywords (unigrams), 114,939,991 unique bigrams, and 341,735,389 unique trigrams. The average and median number of unique keywords per publication is 33 and 32, the average and median number of unique bigrams per publication is 15 and 13, and the average and median number of unique trigrams per publication is 14 and 11.

[4] Inverted indexes store information about each word in a body of text, i.e. they map the content of the body of text to their relative location. This includes the number of occurrences a word is found and its position of occurrence.

[5] We use the following regular expression: [a-z0-9][a-z0-9-]*[a-z0-9]+|[a-z0-9] to tokenize the text.

[6] We remove all stop words which we define as: (i) numbers or words referring to numbers written in full (i.e. one, two, etc.); (ii) single character words; (iii) natural stop words collected from the NLTK Python library (127 words), natural stop words collected from Gensim Python library (337 words), natural stop words collected from Spacy Python library (326 words) (iv) words referring to days and months (i.e. Monday, July, etc.); and (vi) words referring to cities (i.e. New York, London, etc.), countries (i.e. USA, Europe, etc.), first names (i.e. John, Marie, etc.) and institutions (i.e. Nasa, Yale, etc.) retrieved from Wikidata (8,570 words).

[7] We compile a list with the most frequently occurring words in publications. Using an iterative approach where three people independently review this list, we identify and exclude (i) non-scientific keywords unrelated to the scientific content of papers (e.g., abstract, university, journal, publication); and (ii) mistakenly combined words resulting from the tokenization process (e.g., improvementthat, includeoper). We maintain the frequently occurring keywords related to the scientific content of the publication (e.g., patient, chemical, virus).

Open access data files

The csv file "**papers**" contains one row and twelve columns for each paper published between 1800 and 2020 (n=83,490,800). Table 1 describes in details each column.

Table 1 – Content of the file "papers.csv"

Variable	Type	Short description
PaperID	numeric	MAG identifier. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
DOI	string	Digital Object Identifier.
PubMedID	numeric	PubMed identifier.
Year	numeric	Publication year.
Date	string	Publication date in the format *yyyy-mm-dd*. Earliest date between the 'online' date and the print date.
JournalID	numeric	Journal MAG identifier.
ConferenceSeriesID	numeric	Conference MAG identifier.
Volume	numeric	Journal volume.
Issue	numeric	Journal issue.
ReferenceCount	numeric	Number of references to prior papers.
CitationCount	numeric	Number of citations received by the paper (up to 2021).
AuthorCount	numeric	Number of authors.

The csv file "**papers_words**" contains one row and three columns for each paper published between 1800 and 2020 (n=83,490,800). We removed stop words.

Table 2 – Content of the file "papers_words.csv"

Variable	Type	Short description
PaperID	numeric	MAG identifier. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
Words_Title	string	Processed words in the title text.
Words_Abstract	string	Processed words in the abstract text.

The csv file "**papers_bigrams**" contains one row and three columns for each paper published between 1800 and 2020 (n=83,490,800). We only keep bigrams without any stop words.

Table 3 – Content of the file "papers_bigrams.csv"

Variable	Type	Short description
PaperID	numeric	MAG identifier. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
Bigrams_Title	string	Processed bigrams in the title text.
Bigrams_Abstract	string	Processed bigrams in the abstract text.

The csv file "**papers_trigrams**" contains one row and three columns for each paper published between 1800 and 2020 (n=83,490,800). We only keep trigrams without any stop word or with a stop word in the middle of the trigram.

Table 4 – Content of the file "papers_trigrams.csv"

Variable	Type	Short description
PaperID	numeric	MAG identifier. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
Trigrams_Title	string	Processed trigrams in the title text.

Trigrams_Abstract	string	Processed trigrams in the abstract text.

The csv file "**new_words**" contains one row and three columns for each new word (unigram) introduced for the first time in history by papers published between 1901 and 2020 and reused at least once in future papers (n= 7,305,080). Papers published between 1880 and 1900 are used to establish a baseline dictionary of words. We removed stop words.

Table 5 – Content of the file "new_words.csv"

Variable	Type	Short description
Word	string	New word.
PaperID	numeric	MAG identifier of the paper introducing the new word for the first time based on the publication date. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
Reuse	numeric	Total number of future papers which use the word in their title and/or abstract.

The csv file "**new_bigrams**" contains one row and three columns for each new bigram (two consecutive words) introduced for the first time in history by papers published between 1901 and 2020 and reused at least once in future papers (n= 32,104,336). Papers published between 1880 and 1900 are used to establish a baseline dictionary of bigrams. We only keep bigrams without any stop words.

Table 6 – Content of the file "new_bigrams.csv"

Variable	Type	Short description
Bigram	string	Two words part of the new bigram separated by an underscore '_'.
PaperID	numeric	MAG identifier of the paper introducing the new bigram for the first time based on the publication date. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
Reuse	numeric	Total number of future papers which use the bigram in their title and/or abstract.

The csv file "**new_trigrams**" contains one row and three columns for each new trigram (three consecutive words) introduced for the first time in history by papers published between 1901 and 2020 and reused at least once in future papers (n= 75,573,271). Papers published between 1880 and 1900 are used to establish a baseline dictionary of trigrams. We only keep trigrams without any stop word or with a stop word in the middle of the trigram.

Table 7 – Content of the file "new_trigrams.csv"

Variable	Type	Short description
Trigram	string	Three words part of the new trigram separated by an underscore '_'.
PaperID	numeric	MAG identifier of the paper introducing the new trigram for the first time based on the publication date. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
Reuse	numeric	Total number of future papers which use the trigram in their title and/or abstract.

The csv file "**new_word_combs**" contains one row and four columns for each new pairwise combination of words introduced for the first time in history by papers published between 1901 and 2020 and reused at least once in future papers (n=1,129,014,727). To do so, we calculate all possible pairs between any of the words of a paper. In contrast to bigrams and trigrams, it does not matter where the words appear in the paper, nor the order in which they appear. Papers published between 1880 and 1900 are used to establish a baseline dictionary of new word combinations. We only keep new word combinations without any stop words.

Table 8 – Content of the file "new_word_combs.csv"

Variable	Type	Short description
Word1	string	First word part of the new word combination.
Word2	string	Second word part of the new word combination.
PaperID	numeric	MAG identifier of the paper introducing the new word combination for the first time based on the publication date. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
Reuse	numeric	Total number of future papers which use the new word combination in their title and/or abstract.

The csv file "**papers_textual_metrics**" contains one row and fourteen columns for each paper published between 1901 and 2020 (n=72,245,396). This file contains the text measures at paper level. In addition, we use SPECTER to transform each paper title and abstract (not processed) into a 768-size embedding vector to represent the semantic content of the paper (Cohan et al., 2020), and calculate the pairwise cosine similarity between each focal paper and every prior paper published in the preceding 5 years. Next, we calculate *cosine_max* (10^{th} column in dataset) as one minus the maximum cosine similarity between the focal paper and all prior papers. Thus, this measure reflects the semantic distance between the focal paper and the prior paper from the entire population in the previous five years that is closest in scientific content to the focal paper. Alternatively, we calculate *cosine_avg* (11^{th} column in dataset) as one minus the average pairwise cosine similarity between the document-embedding vector of a focal paper and the document-embedding vectors of all papers published in the five years before the focal paper. Hence, this measure reflects the average semantic distance between the focal paper and all prior papers from the entire population in the previous five years.

Table 9 – Content of the file "papers_textual_metrics.csv"

Variable	Type	Short description
PaperID	numeric	MAG identifier. Add 'W' to the numerical value of the identifier to use it in OpenAlex.
new_word	numeric	Number of new words (unigrams) introduced by the paper for the first time.
new_word_reuse	numeric	Number of new words (unigrams) introduced by the paper for the first time weighted by their future reuse, i.e. the sum of the number of new words and their reuse.
new_bigram	numeric	Number of new bigrams (two consecutive words) introduced by the paper for the first time.
new_bigram_reuse	numeric	Number of new bigrams introduced by the paper for the first time weighted by their future reuse.
new_trigram	numeric	Number of new trigrams (three consecutive keywords) introduced by the paper for the first time.
new_trigram_reuse	numeric	Number of new trigrams introduced by the paper for the first time weighted by their future reuse.
new_word_comb	numeric	Number of new pairwise word combinations introduced by the paper for the first time.
new_word_comb_reuse	numeric	Number of new pairwise word combinations introduced by the paper for the first time weighted by their future reuse.
cosine_max	numeric	One minus the maximum cosine similarity between the focal paper and all prior papers published in the previous five years.
cosine_avg	numeric	One minus the average cosine similarity between the focal paper and all prior papers published in the previous five years.
n_words	numeric	Number of unique processed words in the title and abstract.
n_bigrams	numeric	Number of unique processed bigrams in the title and abstract.
n_trigrams	numeric	Number of unique processed trigrams in the title and abstract.

The csv file "**stopwords**" contains the list of stop words used to preprocess the text of papers. The file contains one column and one row for each stop word (n=67,150).

Table 10 – Content of the file "stopwords.csv"

Variable	Type	Short description
Stopword	string	Stopword removed from the title and abstract of the papers.

References

Arts S, Hou J, Gomez JC (2021) Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. ***Research Policy*** 50(2):104-144.

Arts S, Melluso N, Veugelers R (2023). Beyond citations: Text-based metrics for assessing novelty and its impact in scientific publications. https://doi.org/10.48550/arXiv.2309.16437

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level Representation Learning using Citation-informed Transformers. In ***Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*** (pp. 2270-2282).

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., and Wang, K. (2015). An overview of Microsoft academic service (mas) and applications. In ***Proceedings of the 24th international conference on world wide web***, 243–246.