



Crossref Metadata Sprint

Madrid 2025

AI-driven multilingual metadata



- Ramazan
 - Roxana
 - Sajad
 - Mohamed
- **Problem:** Crossref's metadata infrastructure structurally excludes non-English content, leading to poor discoverability, fragmented citation networks, and epistemic inequity.
 - **Diagnosis:** Core issues include monolingual indexing bias, lack of multilingual matching, fragmented DOIs for translations, and absent multilingual metadata structures.
 - **Impact:** This marginalizes Global South scholarship, violates FAIR/Open Science principles, and stifles multilingual technological innovation.
 - **Solution:** Introduce AI-driven multilingual metadata layers via Schema.org/Dublin Core mapping, multilingual profiles, enriched deposit workflows, and enhanced API/API schema to support inclusive scholarly visibility.

Future Steps

- **Define multilingual metadata standards** (Schema.org, DC, BCP47).
- **Deploy AI for translation and cross-language matching.**
- **Upgrade API and deposit flow for multilingual support.**
- **Advocating through member organizations**

Membership & Metadata

Our members often ask us questions about our member list, and we're interested in different ways to display analyses of our membership and metadata.

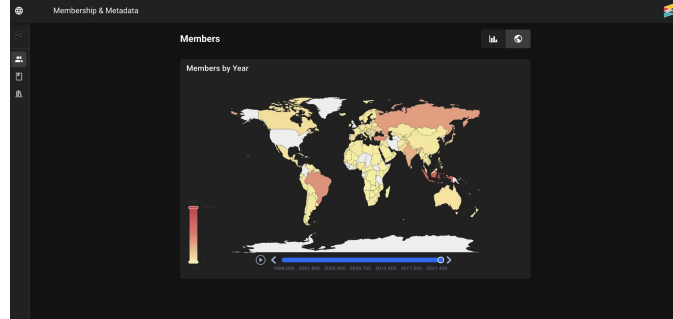
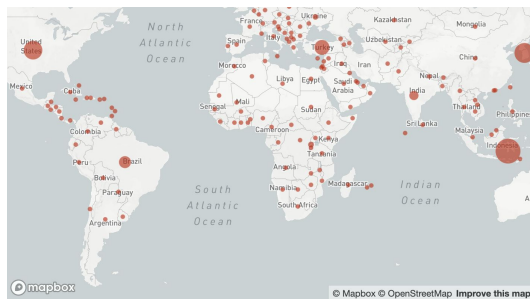
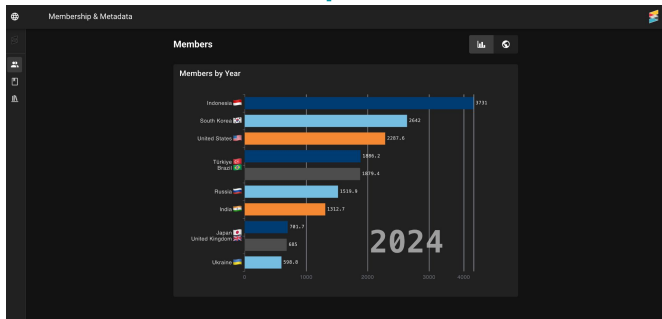
We have ~22k members based from all over the world and we're often asked "how many members are from our country? Who are they, and what are they depositing?".

In this project, we attempt to display some of this information in an interesting and accessible way.

- Paul Davis
- Patrick Vale

Progress achieved

- Transformed data from the Crossref REST API (using node) into a format for consumption by a charting library (echarts)
- Prototyped a member map in Streamlit
- Created chart components to display the data as a bar race and time series overlaid on a world map
- Created a Vue application to wrap the components
- <https://membership-data.netlify.app/members>



Potential next steps

- Extra charting components
- More datasets
- Data cleanup for some datasets
- Add filtering
- Deploy to gitlab pages

Join the Crossref

Analyzing Abstract Coverage in Crossref DOI Articles

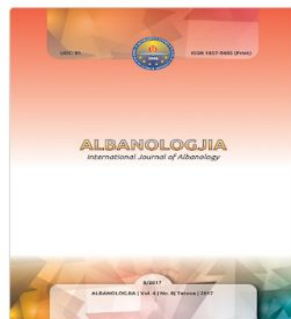
A Data-driven Approach to Study Abstract Presence in University of Tetova Journal Articles

- Agon Memeti, University of Tetova, North Macedonia

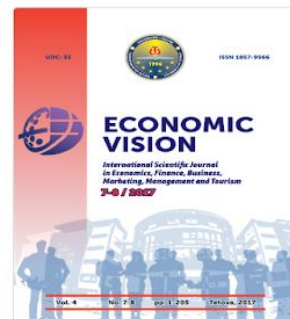
University of Tetova - Journals



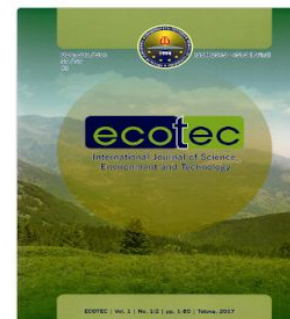
**ACTA MEDICA
BALKANICA**
*International Journal of
Medical Sciences*



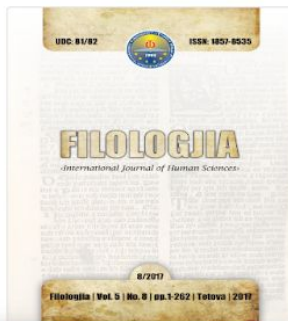
ALBANOLOGJIA
*International Journal of
Albanology*



ECONOMIC VISION
*International Scientific Journal
in Economics, Finance,
Business, Marketing,
Management and Tourism*



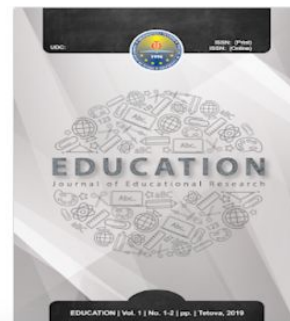
ECOTEC
*International Journal of
Sciences, Environment and
Technology*



FILOLOGJIA
International Journal of Human Sciences
8/2017
Filologjia | Vol. 5 | No. 1 | pp. 1-262 | Tetova, 2017



FREEDOM
*Journal for Peacebuilding and
Transcultural Communication*
Vol. 1 | No. 1-2 | pp. 1-164 | Tetova, 2015



EDUCATION
Journal of Educational Research
Vol. 1 | No. 1-2 | pp. 1-164 | Tetova, 2015



**International Journal of Food
Technology and Nutrition**
Vol. 1 | No. 1-2 | pp. 1-164 | Tetova, 2015



Vol. 11 | No. 21-22, 2024

NDIHMESA E MITHAT FRASHËRIT PËR GJUHËN SHQIPE

Abdurrahim MAXHUNI

Abstract

Prurjen e familjarëve të tij rilindës e vazhdoi edhe Mithat Frashëri, i cili që nga viti 1905 filloi të botonte shkrime të ndryshme politike e gjuhësore. Përmes tyre, ai do të trajtonte probleme që prekun palcën e kombit, atë të gjuhës dhe të historisë së tij. I rritur në një familje ku gjuha shqipe lartësohej, nuk do të mund të qëndronte indiferent karshi gjuhës, karshi laramanisë së përdorimit të saj, karshi shkarjeve të saj. Pjesëmarrja në dy kongrese të rëndësishme, brenga e tij për gjuhën, kultivimi i saj, shpërfaqja e vlerave të saj dhe ruajtja e saj, janë elemente që e përcaktojnë identitetin e tij prej rilindësi. Në këtë kumtesë do të paraqesim ndihmesën e tij në disa shkrime të botuara në përmbledhjen Vepra të zgjedhura 1-8, të botuara në Tiranë.

Pages: 10 - 14

DOI: <https://doi.org/10.62792/ut.albanologjia.v11.i21-22.p2578>

Download

Introduction

- **Objective**

- Analyze University of Tetova Journal articles for 2024.
- Focus on metadata and abstracts using XML files with DOI numbers.

- **Tools**

- Python, Pandas, Matplotlib, and other relevant libraries for XML and data analysis.

Overview of Data

- **Data Sources:**
 - XML files of journal articles for 2024 from University of Tetova Journal.
 - DOI data provided for each article.
- **Data Structure:**
 - Each XML file contains metadata such as title, authors, abstract, publication date, etc.

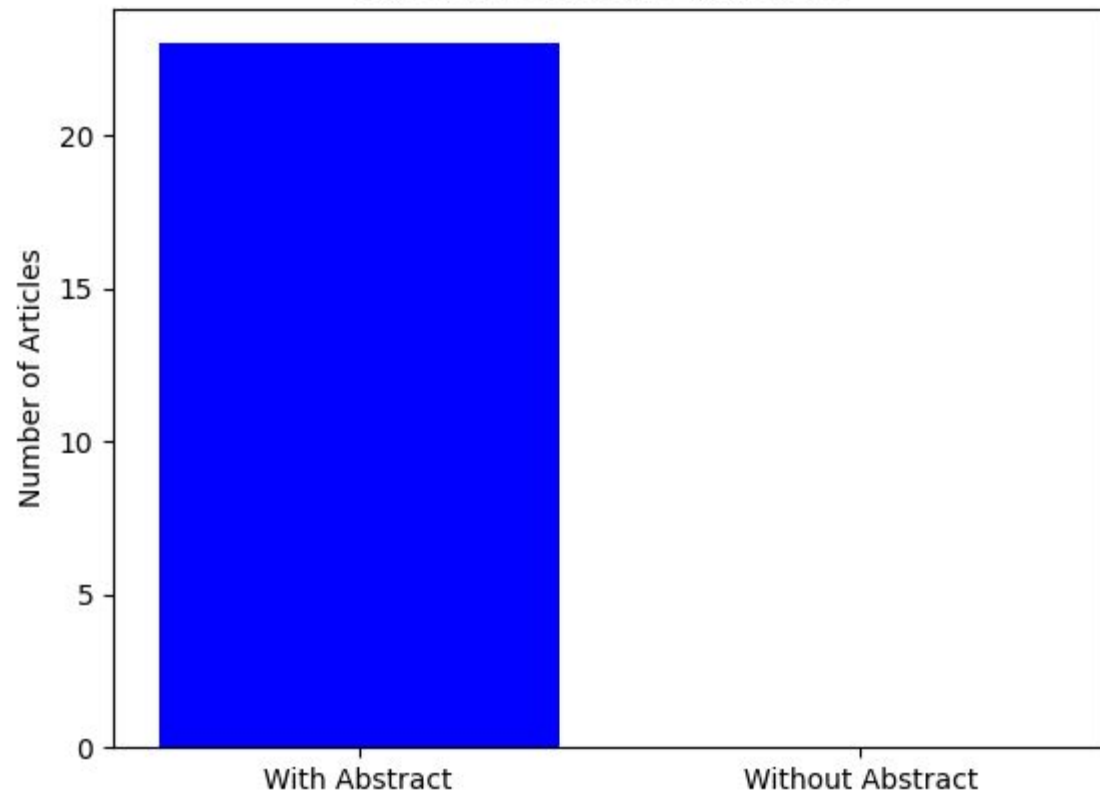
Data Extraction with Python

- Python Script:
 - Extract relevant metadata (DOI, title, abstract, authors) from each XML file.
- Libraries Used:
 - `xml.etree.ElementTree` for XML parsing.
 - `pandas` for data organization and analysis.
 - `matplotlib` for visualizations

Analysis Plan

- What to Analyze:
 - Total number of articles (13 UT Journals - 2024)
 - Articles with and without abstracts.

Abstract Coverage in Articles



Conclusions and Next Steps

- Refine the analysis for other metadata
(e.g., authorship, publication date).
- Create more detailed visualizations.

You are Crossref

Analyzing Authorship Patterns in University of Tetova Journal Articles (2024): Metadata Enrichment and Visualization

- Agon Memeti, University of Tetova, North Macedonia

Introduction

- **Objective:** To analyze the authorship distribution and citations of University of Tetova Journal articles published in 2024.
- **Data Source:** CSV file from Crossref containing DOIs registered to the University's journals.
- **Methodology Overview:** Metadata retrieval, data integration, author analysis, and visualization

UT-2024 CSV File with Metadata

Sheet 1	
1	Authors
2	Hasan SALIU, Zejri MAZLLAMI
3	Gani PLLANA, Agron DURO
4	POPUJVE* TË ISMAIL KADARESE
5	Donika BAKIU
6	Rini USEINI
7	Artresha ZENKI-DALIPI
8	Fatjona ABDULLAHI
9	ALA EL-MEARI AND DIVINA COMMEDIA (DIVINE COMEDY) BY Zejri MAZLLAMI, Hasan SALIU
10	Zymbyle AMETI
11	Mirushe HOKHA
12	Abdurrahim MAXHUNI
13	ANTON NIKË BERISHËS MAKSIMILIAN LAMBERCI PËR LETË Arsim HALILI
14	Marina DANILOVSKA
15	(1918-1929) Zeqirja IDRIZI
16	CONCLUSIONS
17	Nora FAZLIU ABAZI, Natasa TOSESKA SPASOVA, Manushaqe AJVAZI JASHARI
18	Sheqibe BEADINI, Nexhbedin BEADINI, Altin BEGOLLI, Aleina ARSLANI
19	S Sabina TOSUNI, Elona GAXHJA, Entela DRIZAJ, Rezarta STENA
20	Driton SELMANI, Bekim ISMAILI, Ismail SAIDI, Nexhibe NUHII, Qahil IBRAHIMI, Agnesa EMURLI, Elona EMURLI, Afrim ALIU
21	RISTICS: EMPIRICAL ANALYSIS OF THE EXPERIENCES OF T Milazim KAMBERI, Laureta ABAZI, Hyjnor KICA, Adriatik KAMBERI
22	Artresha IBRAHIMI, Marija MILOSEVSKA JANAKIESKA, Vjosa SARAÇINI
23	Imer ZENKU, Merita DEMA, Nexhbedin ISMAILI
24	Liljana D. SOFRONIEVSKA, Bojan KARANAKOV, Ognjen MARINA, Teodora MIHAJLOVSKA
25	Zoran JOŠEVIĆ, Pero STEFANOVIĆ, Feta SINANI, Mile STOJIMENOV
26	Teodora MIHAJLOVSKA, Dobre MIKOLOVSKI, Ana TROMBEVA-GAVRILOSKA, Liljana DIMEVSKA SOFRONIEVSKA
27	Florinë MUSHICA, Agon MEMETI
28	Mirjeta MEDII ARIFI, Mimoza RISTOVA, Vesna GERSHAN
29	IIA Hajdar KIČAJ, Aurora BAKAJ, Mariola ISMAILAJ, Xhuljana ARAPAJ, Megi MALAJ, Juna TAFILI

UT-2024 CSV File with Metadata

Sheet 1		
Fatjona HYSI, Eglantina DERVISHI	0	
Drenusha KAMBERI, Hava REXHEPI, Gojart KAMBERI	0	
Blerta ALILI, Daniela DIMITROVA RADOJICQ	0	
Kateryna IVANOVA	0	
Isa SPAHIU	0	
Mirlinda SHAQIRI, Teuta ILJAZI	0	
Mirela SHELLA	0	
Brian D. JOSEPH	0	
Arbana AJDINI	0	
Angjela MARKOVIC, Nalle MALA-IMAMI	0	
	0	
Arlind FARIZI, Valdet HYSENAJ, REXHEP HOTI	0	
Nurie EMRULLAI	0	
Shiret ELEZI, Milena BOSHKOSKA KLISAROSK, Marija MIDOVSKA PETKOSKA	0	
Faton SHABANI	0	
Artina KAMBERI, Shenaj HAXHIMUSTafa	0	
, Muareme AJDARI, Osman ISMAILI,	0	
, Sonja NOVOTNI, Mihailo MARKOVIC,	0	
	0	
, Sunchica TRIFUNOVSKA-JANIC	0	
, Irena GJORGJESKA, Sonja NOVOTNI, , Mihailo MARKOVIC,	0	
, Hana DARDHISHTA, Marina SPASOVSKA,	0	
, Shejnaze AJDINI-MURTEZI, Dëtrim SALIU,	0	
, Esmeralda HIDRI	0	
, Kamila AMMOUR	0	
Albulena BEADINI, Sheqibe BEADINI, Nexhbedin BEADINI, Blandi LOKAJ, Festina DOBRA, Greta KASTRATI, Ardit ADEMI, Adriana DUSHAJ	0	
Adelina ELEZI, Sheqibe BEADINI, Irena KOSTOVSKA, Egzona ZIBERI	0	
Albin BEADINI, Adelina ELEZI, Koço ÇAKARALOSKI, Learta HASANI	0	
Albulena BEADINI, Irfan AMETI, Adelina ELEZI, Albin BEADINI, Avdi NAZIFI, Egzona ZIBERI	0	
Art ZYLBEARI, Kocho CHAKALAROSKI, Milka ZDRAVKOVSKA	0	
Ignac SIVEC, Vladimir KRPAČ, Mladen KUČINIĆ	0	
Endora CELOHOXHAJ, Michele LEMONNIER-DARCEMONT, Christian DARCEMONT, Elton HALIMI	0	
Gresa AHMA, Kushtrim GASHI, Mereme IDRIZI, Hanife RUSTEMI-AMETI, Xhezair IDRIZI, Rejhana LUMA, Hava MIFTARI, Durim ALLJA, Eljesa	1	

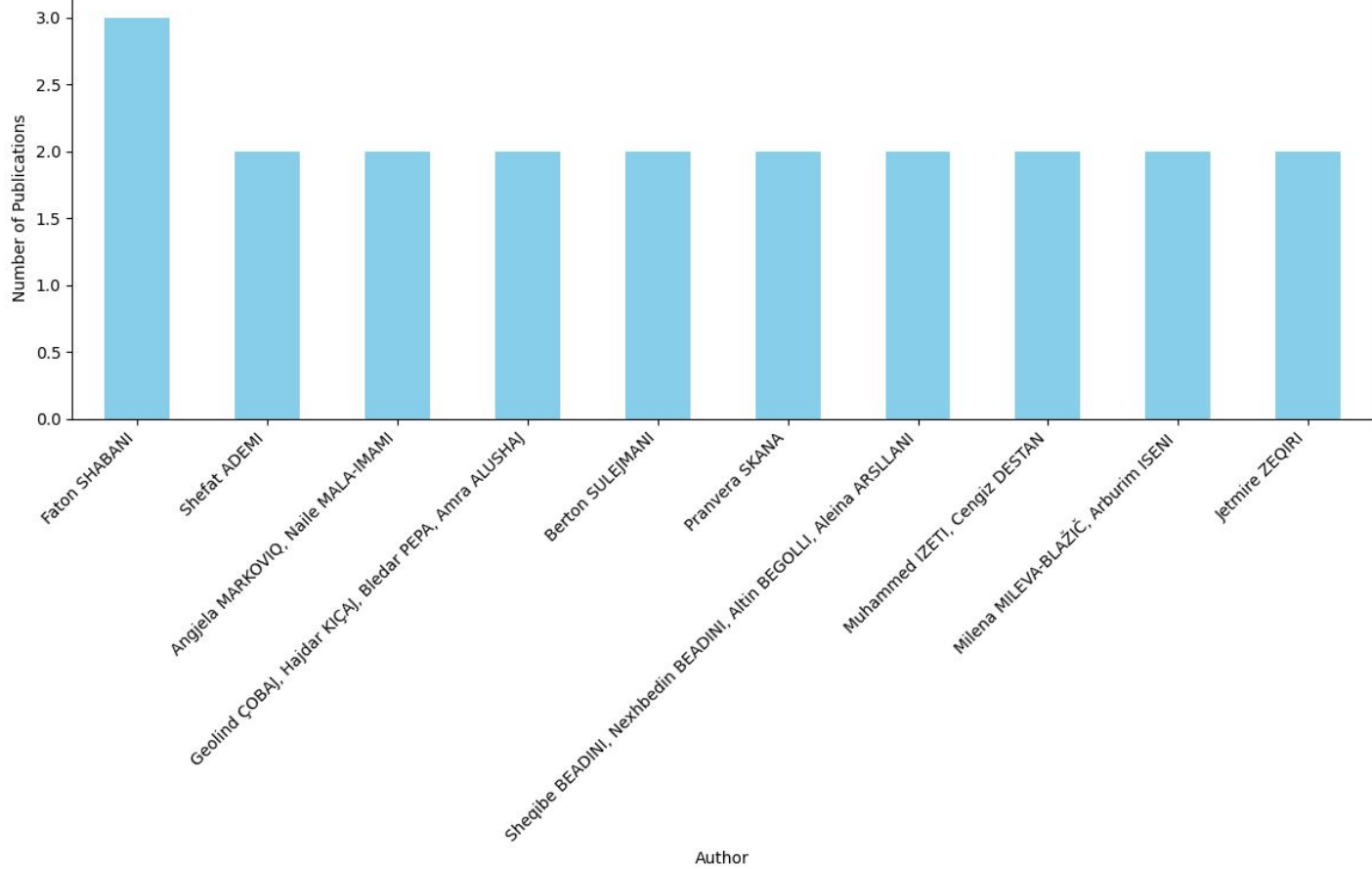
Metadata Retrieval and Data Integration

- **Metadata Retrieval:** Utilized the Crossref REST API to fetch metadata (titles, authors and citations) for each DOI.
- **Data Integration:** Merged retrieved metadata with the original dataset based on the DOI column, ensuring a comprehensive DataFrame.

Author Analysis

- **Data Processing:** Split the 'Authors' field (semicolon-separated names) into individual entries using `str.split(';')` and `explode()`.
- **Publication Count:** Determined the number of publications per author, and citations per paper.

Top 10 Authors by Number of Publications



Author changes between preprints and published articles

- Mohamed

Progress achieved

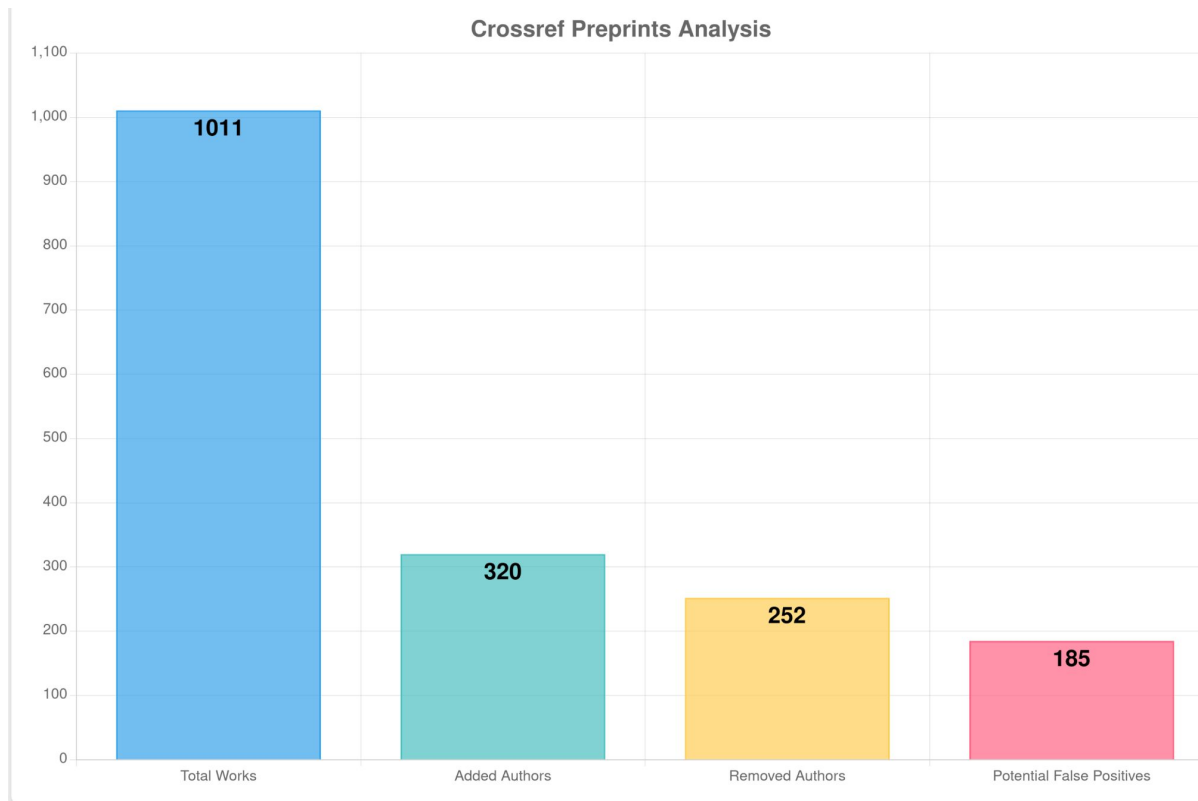
Gitlab repo: <https://gitlab.com/mselim1/ppr-journal-auth-checker>

This project checks Author changes between preprints and research articles in crossref. The code ran on **1011 pairs** from CSHL.

There is some tolerance in author matching:

- . Case insensitive
- . Handle names moved across given name and family name
- . Remove hyphens and dots
- . Ignore if first name is written as initial letter in either version
- . If number of removed authors and added are the same, assume it is potential false positive

Progress achieved



Progress achieved

Potential genuine author **removal 6.7%** compared to author **addition 13.5%** in published version

Conclusions:

- Author addition in published version is more likely than removal
- Corrupted data entering makes matching challenging leading to false positives
- More fuzzy logic required to match authors as permutations of writing names in full or only initials needs to be considered
- In some cases the preprint include author consortia (team name) and in published version it has all individual names which makes author matching impossible

Example:

Published version;

<https://api.crossref.org/works/10.1038/s41531-024-00838-4>

Preprint version:

<https://api.crossref.org/works/10.1101/2023.10.18.23297218>

Potential next steps

1. Handle scalability issue as script blocked after almost 1K request
2. Make doi prefix dynamic
3. Allow adding API tokens for members
4. Using more fuzzy logic like tokenisation or generating different author permutations and check if at least one exists

Join the Crossref

Enriching citations with Crossref metadata



- Jack Ekinsmyth
- Maria de la Paz Cardona
- Ana Bermejo

Context

- The Funding Impact Measures Team aims to understand the impact of Wellcome funding against Wellcome's strategic mission
- We regularly assess the impact of large scale repositories of publications, policy documents, patents etc which cite scholarly literature.
- However a key issue is data quality when matching document references lists to scholarly DOIs.

Aims

- Create a tool which:
 - Parses XML documents to extract their references
 - Formats and queries the Crossref API
 - Validates the quality of the matches
- To allow for more streamlined impact analyses

Progress achieved

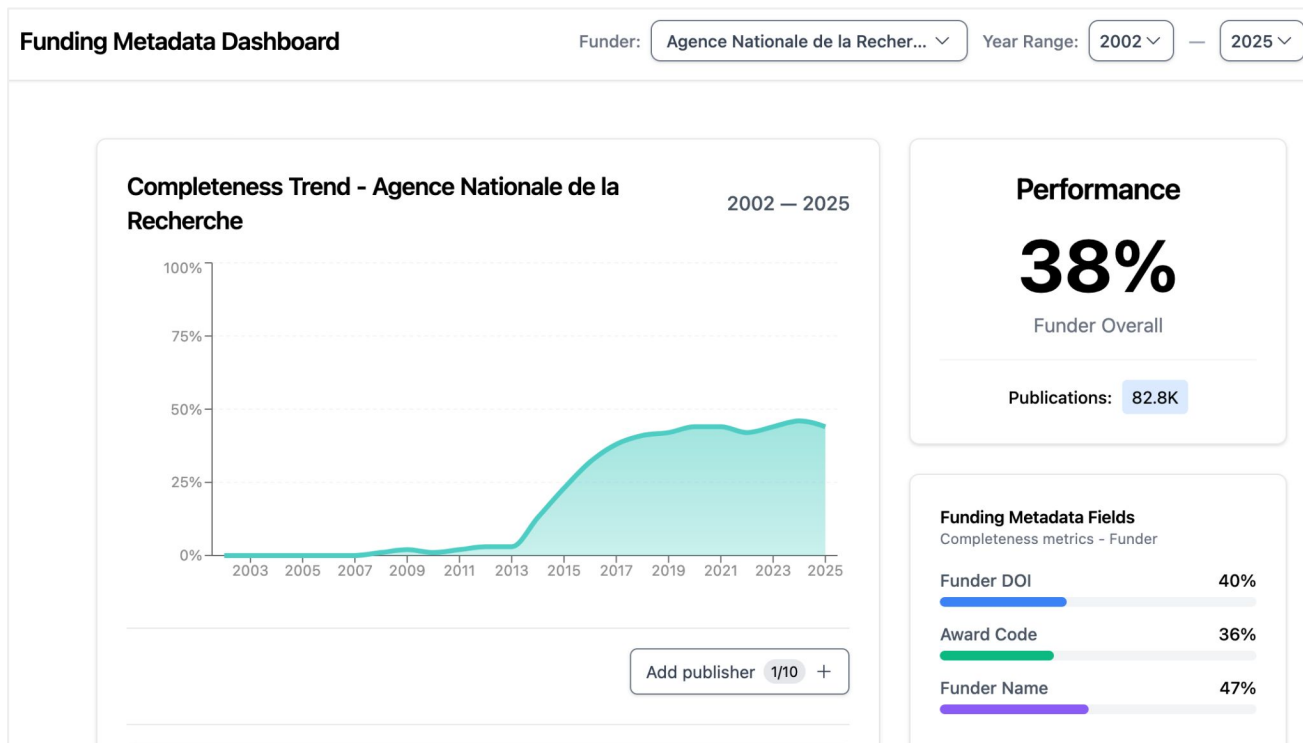
- Parsing XML citations to create Harvard-style references that Crossref can understand
- Fuzzy matching retrieved results from Crossref to identify true matches
- Using an LLM agent to determine whether low ratios from fuzzy match actually correspond to same or different publications

Potential next steps

- Further explore the case of review articles which crossref confuses with the original publication.
- Fine tune the LLM agent to become better at discriminating between original and review.

You're Crossref

Funding Completeness



Dashboard: <https://funding-metadata-dashboard-ui.vercel.app/>

Progress achieved

- Generalizing and documenting the whole pipeline
 - <https://github.com/adambuttrick/anr-funding-metadata-analysis/tree/202504-adam-hackathon>
- Tested improvements by adding new funder entry

Potential next steps

- Add additional funders
- Expand dashboard with other API data

Join the Crossref

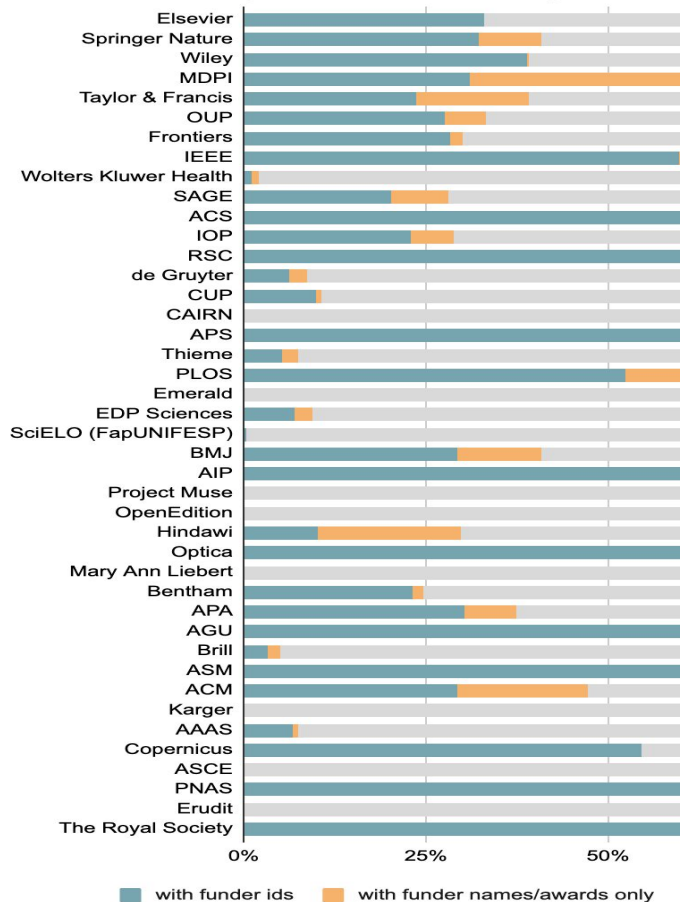
Funding Completeness (2)

- Proportion of DOIs per publisher* with funding information, including assigned Funder ID
- Proportion of funding records with Funder ID assigned by publisher vs by Crossref
- Funder IDs assigned by publisher: variance across funders?

**selected publishers shown, journal articles 2022-2024*

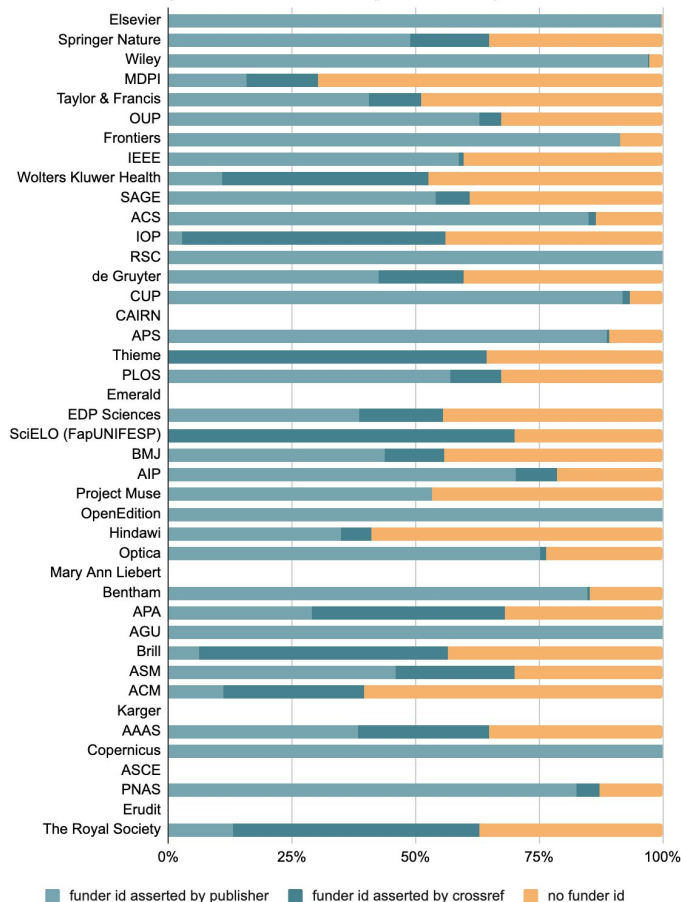
Selected publishers - Crossref DOIs with

journal articles 2022-2024 (per 2025-02-28)



Selected publishers - funding records in Crosref DOIs

journal articles 2022-2024 (per 2025-02-28)





Relative proportion of funder IDs assigned by publisher (vs by Crossref)

Potential next steps

- Map publisher-level data completeness to submission / publishing systems in use
- Discuss with funders, publishers and service providers how to improve capture and propagation of funder metadata (part of Barcelona Declaration working group)

Join the Crossref

Improving the Public Data File

- Iñaki Úcar
- Carlos del Ojo

Identified issues

- Providing alternative formats
- Providing complementary features
- Diff to update the info from year to year

Progress

- GitHub repo: <https://github.com/Enchufa2/metadata-sprint-madrid-2025>
- Explored the suitability of Parquet as a new file format
- Explored the suitability of DuckDB as analytical engine
- Provided converters
- Extracted the complete schema
- Provided benchmarks that prove that Parquet + DuckDB in a single server rivals with Spark queries in a cluster
- Explored options for data versioning such as Apache Iceberg and lakeFS

Conclusions and Next Steps

- Parquet provides more compression, flexibility, and accessibility
- The DuckDB engine provides a high-performance as well as easy to use SQL interface
- Partitioning would allow distributing much smaller diffs
- More advanced solutions should be explored

Volunteered Crossref

Retraction Data Mash-Up

Didi, Blessing

Goal

Connect Retraction Watch (RW) data set with metadata from CrossRef and others (ROR, OpenAlex) for analysis; e.g. by publisher, journal, country, etc.

You're Crossed

Artifacts

Pipelines

- Load daily RW csv dump from GitLab
- Process RW data against ROR API
- Process RW data against CrossRef API

Web App for user to query data and generate charts

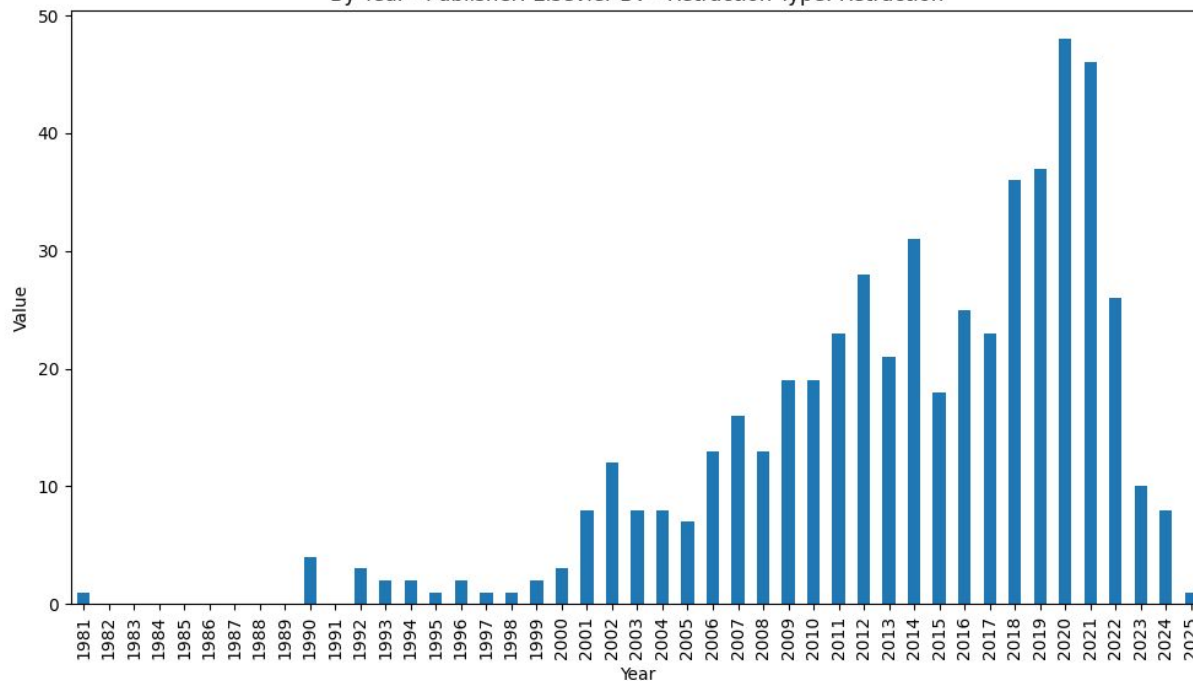
Volunteered Crossref

Limitations

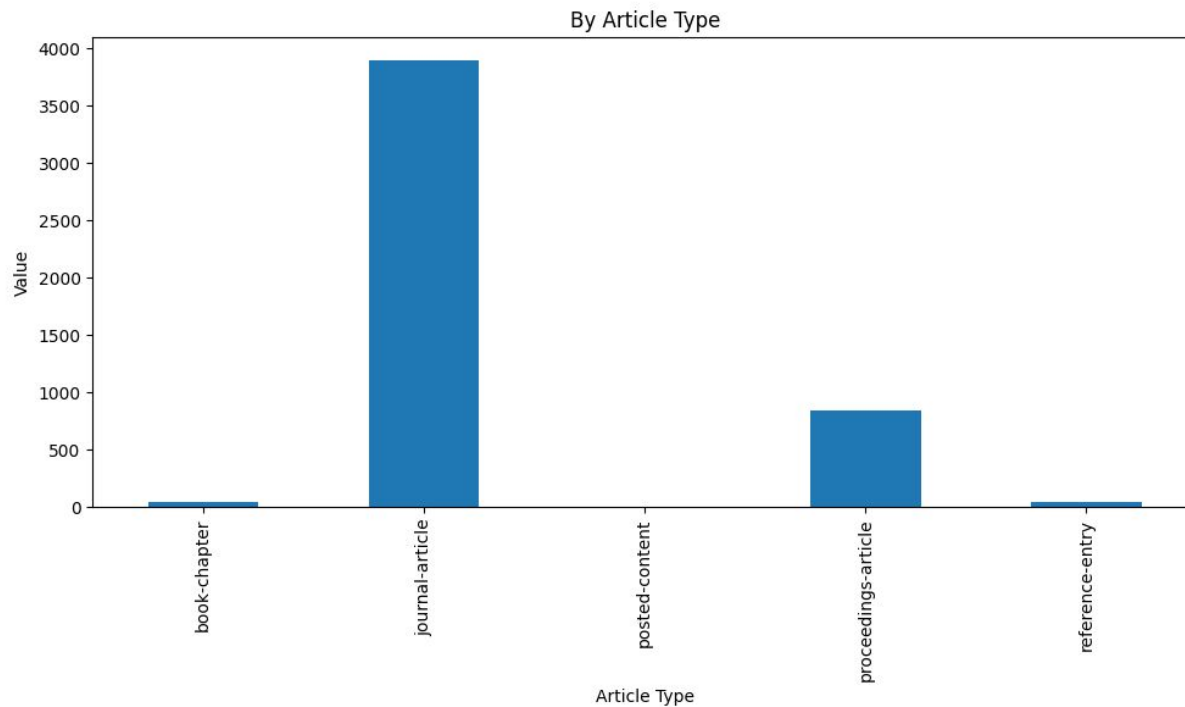
- ROR data is mostly in arrays; which we **mangled to string** when **dumping to parquet** files
- Funder data is too diverse; need to **reconciliate on higher funder hierarchy level** to be meaningful
- We **sampled RW data to 5000 items** due API look-ups and time constraints

Volunteered Crossref

By Year - Publisher: Elsevier BV - Retraction Type: Retraction



Based on limited data sample



Based on limited data sample

Powered by Crossref

Retraction Analysis

Crossref Metadata Sprint 2025

Contents

- Team
- Challenge
- Dataset Stats
- Crossref & Open Citations (1 paper, proof of concept)
- OpenAlex (wider stats)

Team

- Alexandra Málaga
- Cyril Labbe
- Robert Bianchi
- Yagmur Ozturk



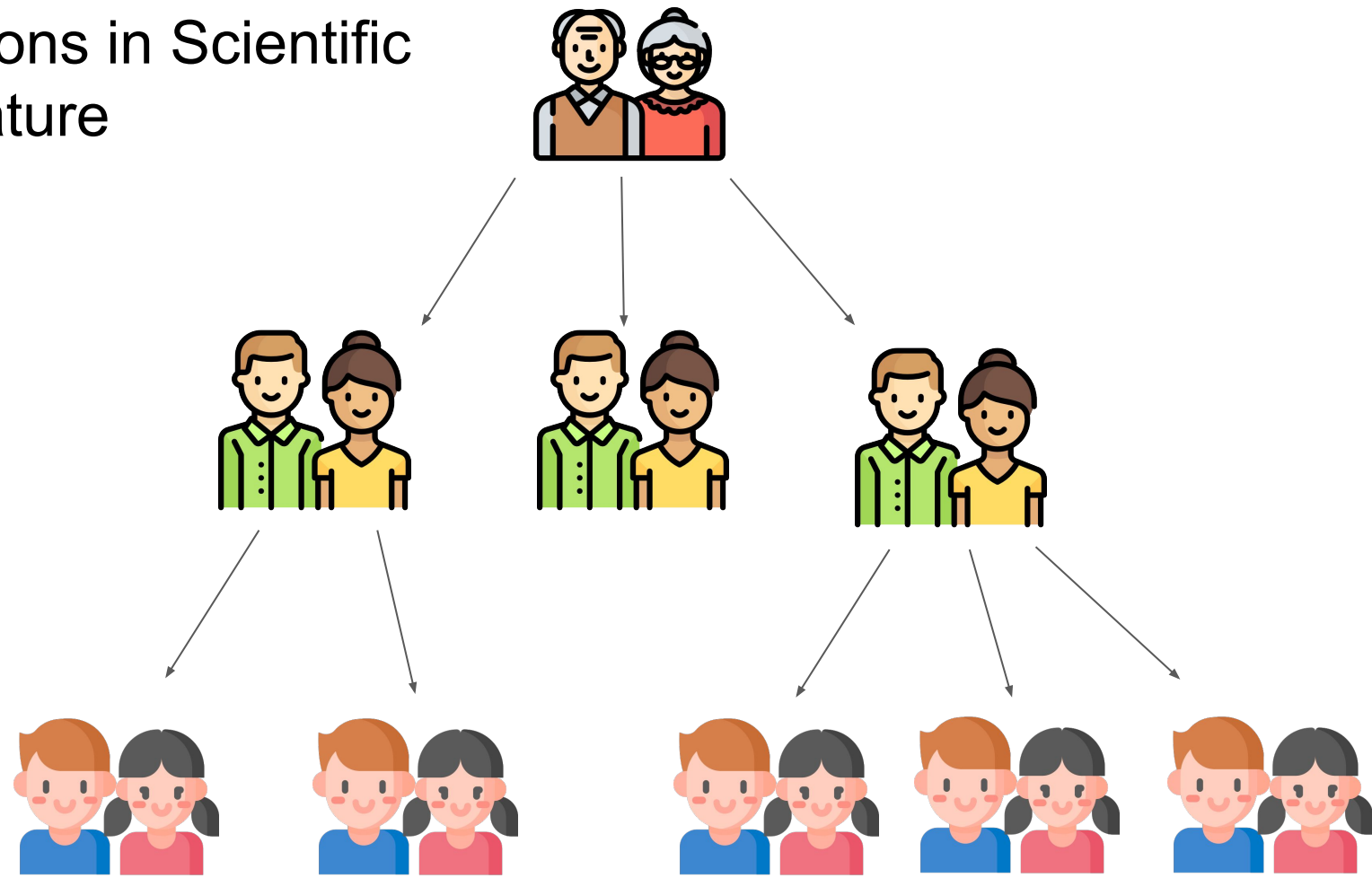
Challenge

- Visibility of retractions across citation generations gets lost
 - Papers can cite “clean” (not retracted) papers which themselves are dependent on retracted papers. These dependencies can quickly get hidden and lost in the citation network.
- Analysis of citation networks quickly grows across generations

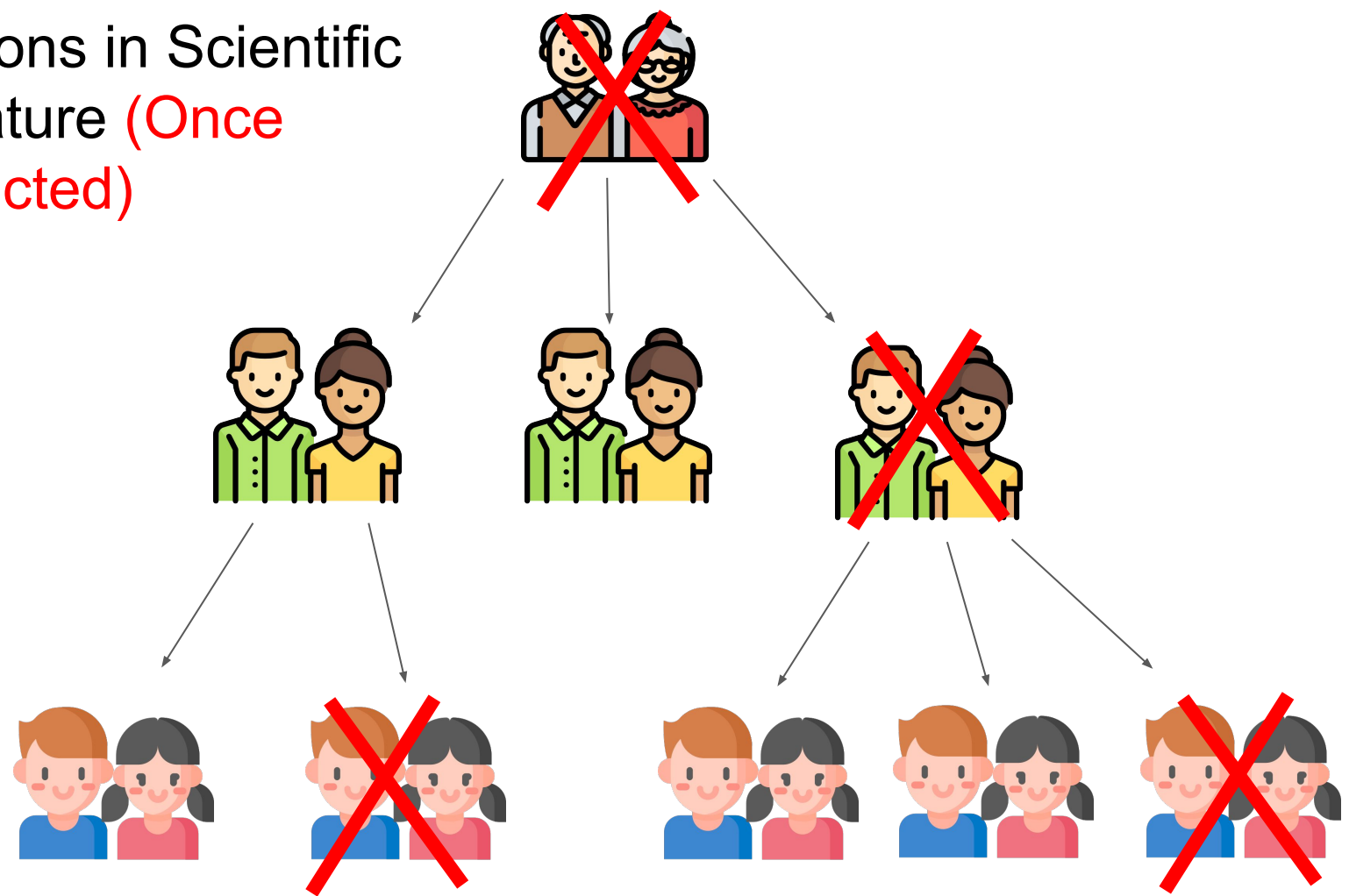
Tasks

- Test analytics for 1 example using various data sources (Crossref, OpenCitations, OpenAlex)
- Investigate scale of task across 1 full dataset (OpenAlex)
- Pipelines to repeat the data collection process

Citations in Scientific Literature



Citations in Scientific Literature (Once Retracted)



Pipeline to extract metadata of citations to retracted articles (OpenCitations and Crossref data)

1. Given a DOI
2. Extract DOIs of papers that cite the RETRACTED SOURCE paper



3. Extract DOIs of papers that cite the parents



4. Get JSON dumps from Crossref to analyze these papers

Database Stats Comparison (Citations to Wakefield et al.)

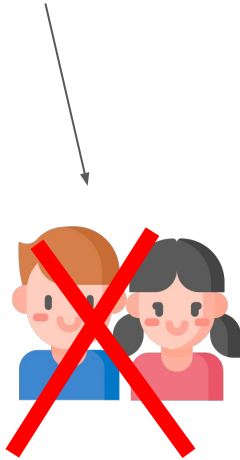
	OpenCitations	Crossref
All Records	32855	49244
Unique Papers	25093	31726
Generation 1 (sons, visible in the Cited By list)	1779	1869
Retracted in generation 1 (visible)	3	2
Generation 2	24633	32098
Retracted in generation 2 (not visible)	41	82

How many retractions/corrections in the lineage of Wakefield et al.? [Crossref data]



["Correction"]	6
["Erratum"]	4
["New version"]	4
["Erratum", "Retraction"]	1
["Erratum", "Correction"]	1
["Corrigendum"]	1

How many retractions/corrections in the lineage of Wakefield et al.? [Crossref data]



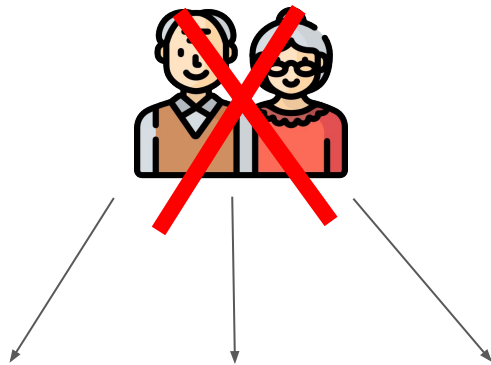
["Correction"]	63
["Erratum"]	26
["New version"]	24
["Retraction"]	11
["Retraction", "Retraction"]	6
["Corrigendum"]	5
["New edition"]	5
["Erratum", "Erratum"]	2
["Withdrawal", "Retraction"]	2
["Expression of concern"]	2
["Retraction", "Erratum", "Retraction"]	1
["New version", "Retraction"]	1
["Withdrawal"]	1
["Erratum", "Retraction"]	1
["New version", "Correction"]	1

Size of the cake (OpenAlex)

Total grandpas = 51k (whole Retraction Watch DB)
(50 998)

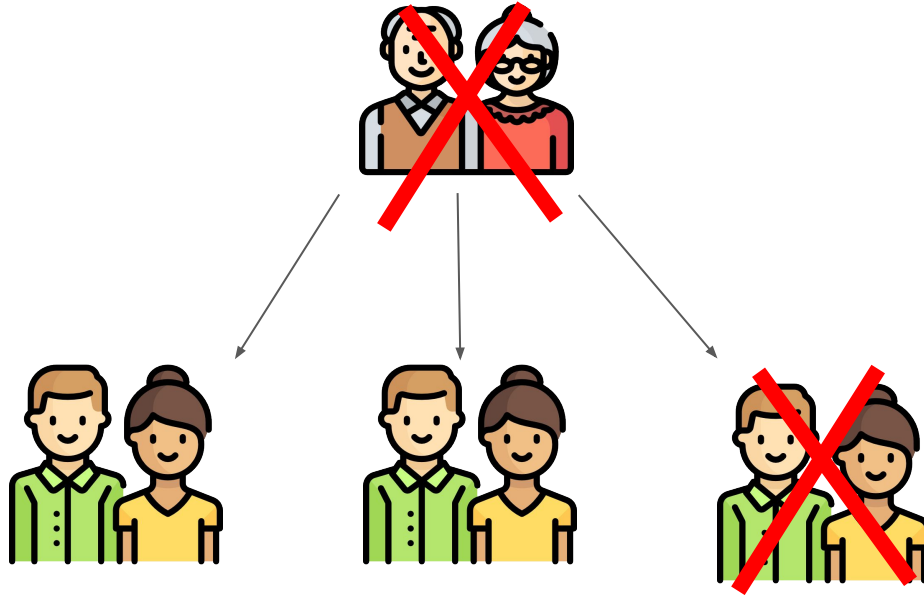


Lonely grandpas = 12k
(12 244 - 24%)



Family grandpas = 38k
(38 754 - 76%)

Size of the cake



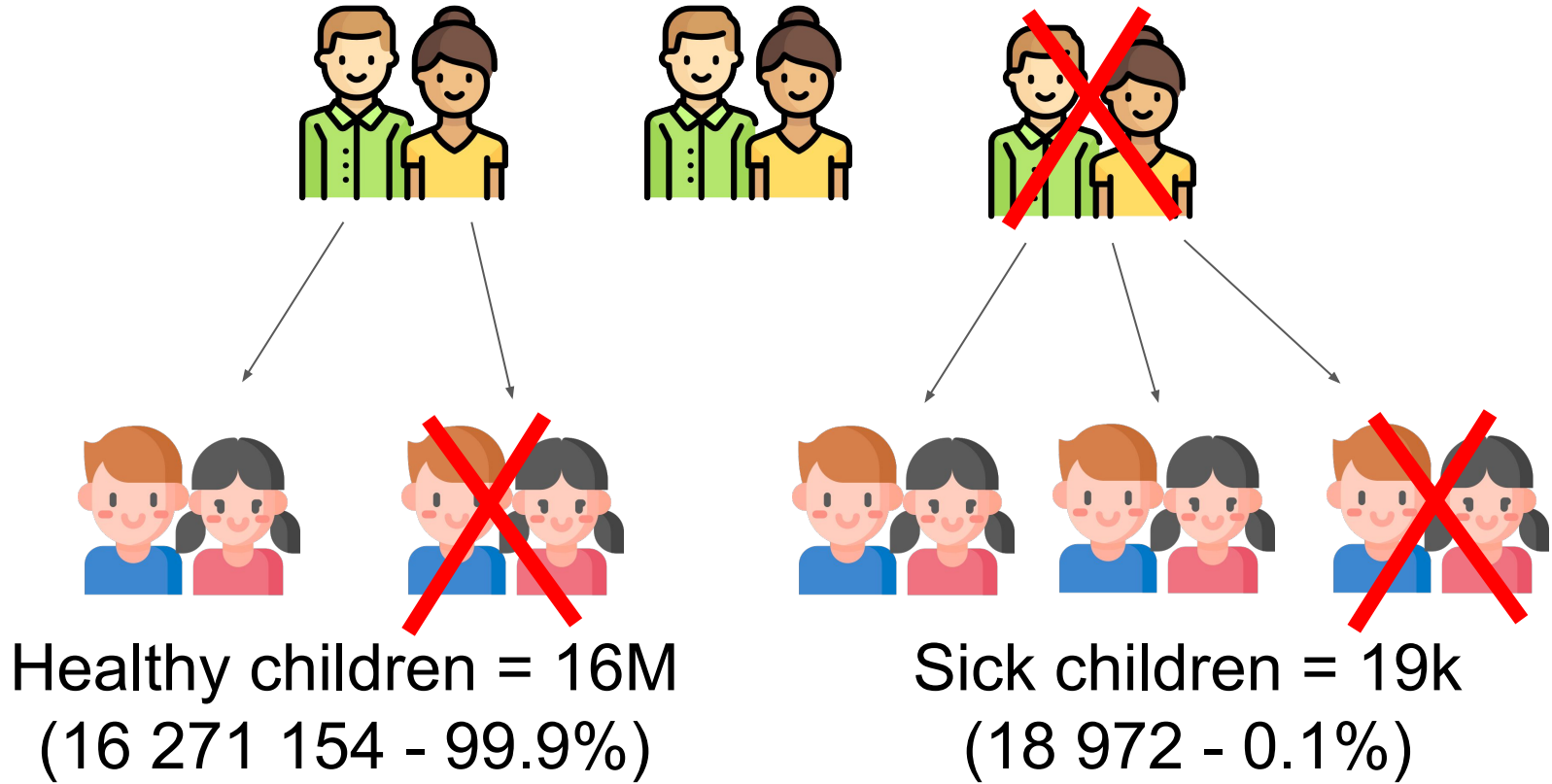
Healthy parents = 702k
(701 925 - 99%)

Sick parents = 8k
(8 225 - 1%)

Size of the cake

	Parent not retracted (%)	Parent retracted (%)
1 sick grandpa	85	60
2 sick grandpas	10	20
3 sick grandpas	3	8
+4 sick grandpas	2	12
	100	100

Size of the cake



Size of the cake

grandpas	Number of sons	Number of grandsons
W3046275966	7318	80355
W2166000498	3961	389584
W3120387768	3948	116547

<https://openalex.org/works/W3046275966>