Alexandria3k: Researching the world's knowledge on your laptop

Diomidis Spinellis

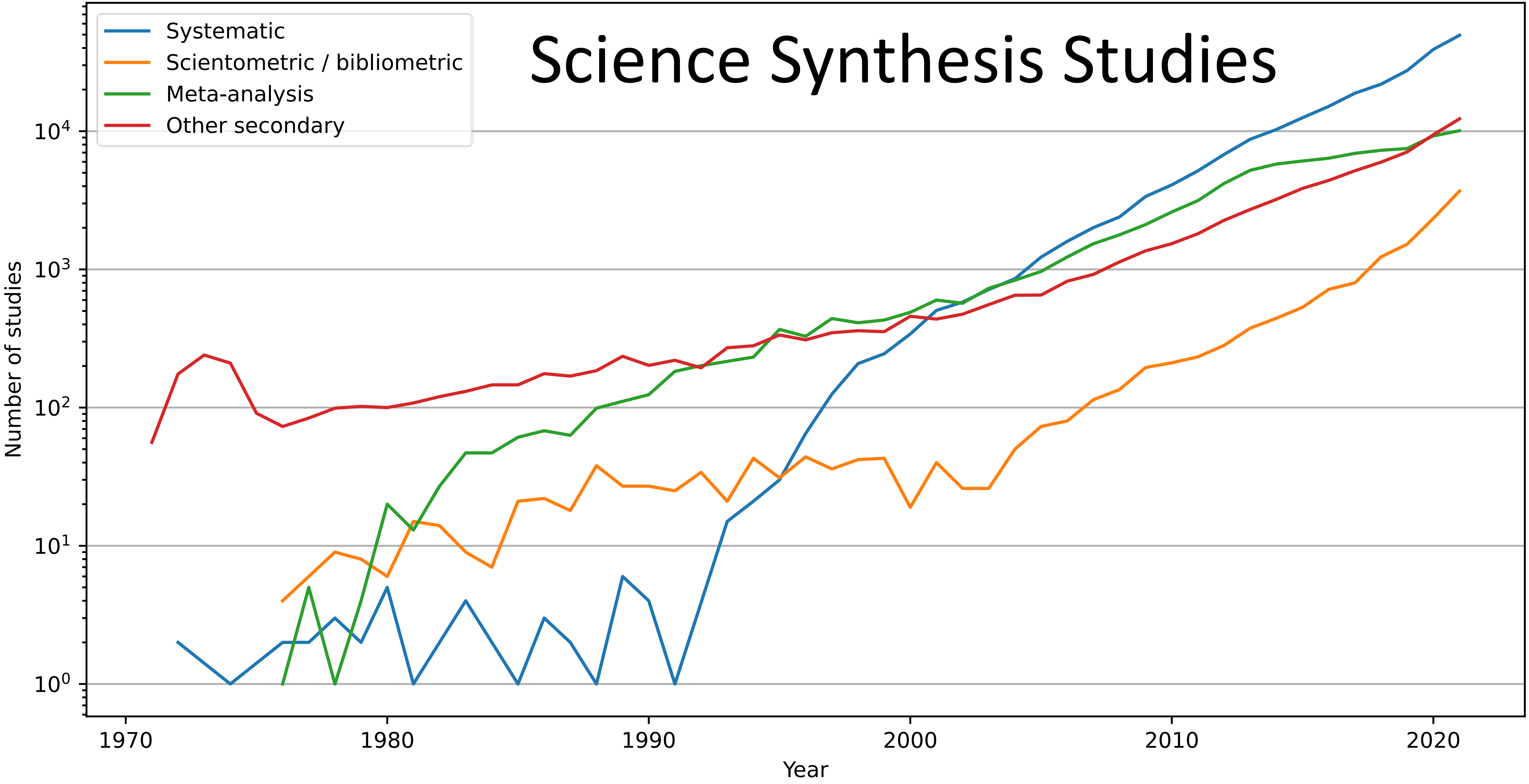
Department of Management Science and Technology
Athens University of Economics and Business

Department of Software Technology
Delft University of Technology

www.spinellis.gr

𝕏 @CoolSWEng

@CoolSWEng@mastodon.acm.org


Science Synthesis Studies

Standing on shoulders or feet? An extended study on the usage of the MSR data papers

Kotti, Kravvaritis, Dritsa, Spinellis

Empirical Software Engineering (2020)

DOI: 10.1007/s10664-020-09834-7

🏆 ACM SIGSOFT Distinguished Paper Award in MSR 2019

Empirical Software Engineering (2020) 25:3288–3322
https://doi.org/10.1007/s10664-020-09834-7

Check for updates

Standing on shoulders or feet? An extended study on the usage of the MSR data papers

Zoe Kotti[1] · Konstantinos Kravvaritis[1] · Konstantina Dritsa[1] · Diomidis Spinellis[1]

Published online: 18 July 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

The establishment of the Mining Software Repositories (MSR) data showcase conference track has encouraged researchers to provide data sets as a basis for further empirical studies. The objective of this study is to examine the usage of data papers published in the MSR proceedings in terms of use frequency, users, and use purpose. Data track papers were collected from the MSR data showcase track and through the manual inspection of older MSR proceedings. The use of data papers was established through manual citation searching followed by reading the citing studies and dividing them into strong and weak citations. Contrary to weak, strong citations truly use the data set of a data paper. Data papers were then manually clustered based on their content, whereas their strong citations were classified by hand according to the knowledge areas of the Guide to the Software Engineering Body of Knowledge. A survey study on 108 authors and users of data papers provided further insights regarding motivation and effort in data paper production, encouraging and discouraging factors in data set use, and future desired direction regarding data papers. We found that 65% of the data papers have been used in other studies, with a long-tail distribution in the number of strong citations. Weak citations to data papers usually refer to them as an example. MSR data papers are cited in total less than other MSR papers. A considerable number of the strong citations stem from the teams that authored the data papers. Publications providing Version Control System (VCS) primary and derived data are the most frequent data papers and the most often strongly cited ones. Enhanced developer data papers are the least common ones, and the second least frequently strongly cited. Data paper authors tend to gather data in the context of other research. Users of data sets appreciate high data quality and are discouraged by lack of replicability of data set construction. Data related to machine learning or derived from the manufacturing sector are two suggestions of the respondents for future data papers. Overall, data papers have provided the foundation for a significant number of studies, but there is room for improvement in their utilization. This can be done by setting a higher bar for their publication, by encouraging their use, by

Communicated by: Yasutaka Kamei and Andy Zaidman

✉ Zoe Kotti
zoekotti@aueb.gr

Extended author information available on the last page of the article.

Springer

Impact of SE Research in Practice: A Patent and Author Survey Analysis

Kotti, Gousios, Spinellis

IEEE Transactions on Software Engineering (2022)

DOI: 10.1109/TSE.2022.3208210

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING 1

Impact of Software Engineering Research in Practice: A Patent and Author Survey Analysis

Zoe Kotti, Georgios Gousios, and Diomidis Spinellis, *Senior Member, IEEE*

Abstract—Existing work on the practical impact of software engineering (SE) research examines industrial relevance rather than adoption of study results, hence the question of how results have been practically applied remains open. To answer this and investigate the outcomes of impactful research, we performed a quantitative and qualitative analysis of 4 354 SE patents citing 1 690 SE papers published in four leading SE venues between 1975–2017. Moreover, we conducted a survey on 475 authors of 593 top-cited and awarded publications, achieving 26% response rate. Overall, researchers have equipped practitioners with various tools, processes, and methods, and improved many existing products. SE practice values knowledge-seeking research and is impacted by diverse cross-disciplinary SE areas. Practitioner-oriented publication venues appear more impactful than researcher-oriented ones, while industry-related tracks in conferences could enhance their impact. Some research works did not reach a wide footprint due to limited funding resources or unfavorable cost-benefit tradeoff of the proposed solutions. The need for higher SE research funding could be corroborated through a dedicated empirical study. In general, the assessment of impact is subject to its definition. Therefore, academia and industry could jointly agree on a formal description to set a common ground for subsequent research on the topic.

Index Terms—software engineering, practical impact, empirical study, survey, patent citations

1 INTRODUCTION

IN 2018, the field of software engineering (SE) marked the 50th anniversary of its first two-year conference series—the 1968–69 NATO Conferences on Software Engineering [1], [2]. Despite its relatively short period of existence, a lot of research has been performed in SE during these 50 years, composing a large body of information. In the meantime, numerous software and technology-related companies have emerged, partially as a result of hardware advancement and cloud computing [3], forming a multi-trillion dollar industry [4]. This growth both in terms of knowledge and market share raises the question of how these two relate, and to what extent research may have impacted industry. In this context, we define as *impact* **the direct or indirect incorporation of a software engineering study's output in an industrial setting, for example, in an industrial software development tool, process, marketable product, or service**.

In the scope of this study, we consider *software engineering* the discipline that systematically employs computer science knowledge and principles to develop new methods and tools to improve software development. The discipline's areas include software requirements, design, construction, testing, maintenance, configuration management, quality, SE management, SE models and methods, and SE process [5]. The application process is based on systematic, disciplined, and quantifiable SE approaches, and is influenced by cross-disciplinary areas, namely, mathematics, general management, project management, and systems engineering [5]. Note that our definition distinguishes foundational computer science research (*e.g.*, devising a new static analysis method, a test prioritization algorithm, or a requirements definition language) from that performed in SE contexts. For the described examples to be considered SE research, we expect them to be accompanied with empirical evaluation through, for example, repository mining, a developer survey, or a case study.

Existing work on the practical impact of SE research examines industrial relevance rather than adoption of study results. A variety of interviews and literature reviews have been conducted, mainly in domain-specific contexts such as the ACM SIGSOFT Impact Project [6], to assess the relation of research to industrial needs, highlight gaps between the two, and suggest best practices for collaborative projects. However, the question of how research results have been practically applied remains open.

To tackle this question and investigate the outcomes of impactful SE research, we performed a quantitative and qualitative analysis of SE patents citing SE research from four leading SE venues. Patents are by definition practical applications of technology, and are frequently employed as an estimator of the academic research impact (*e.g.*, in the works by Narin *et al.* [7], Estublier *et al.* [8], and the National Academy of Engineering [9]). Software patents have increased rapidly in number, comprising 15% of all patents [10]. Most of them are acquired by large manufacturing firms from the computers, electronics, and machinery industries [10]. Furthermore, we conducted a survey on authors of highly recognized SE publications to examine impactful types, areas, methods, and outcomes of SE research as well as their footprint on information technology, society, and industry.

Our findings suggest that SE researchers have equipped practitioners with various tools, processes, and methods,

- *D. Spinellis and G. Gousios are with the Department of Software Technology, Delft University of Technology, The Netherlands. E-mail: {D.Spinellis,G.Gousios}@tudelft.nl*
- *D. Spinellis and Z. Kotti are with the Department of Management Science and Technology, Athens University of Economics and Business, Greece. E-mail: {dds,zoekotti}@aueb.gr*

ML4SE: A Tertiary Study

Kotti, Galanopoulou, Spinellis

ACM Computing Surveys, 2023

arXiv:2211.09425v1 [cs.SE] 17 Nov 2022

Machine Learning for Software Engineering: A Tertiary Study

ZOE KOTTI, RAFAILA GALANOPOULOU, and DIOMIDIS SPINELLIS, Athens University of Economics and Business, Greece

Machine learning (ML) techniques increase the effectiveness of software engineering (SE) lifecycle activities. We systematically collected, quality-assessed, summarized, and categorized 83 reviews in ML for SE published between 2009–2022, covering 6 117 primary studies. The SE areas most tackled with ML are software quality and testing, while human-centered areas appear more challenging for ML. We propose a number of ML for SE research challenges and actions including: conducting further empirical validation and industrial studies on ML; reconsidering deficient SE methods; documenting and automating data collection and pipeline processes; reexamining how industrial practitioners distribute their proprietary data; and implementing incremental ML approaches.

CCS Concepts: • **Software and its engineering** → Extra-functional properties; Automatic programming; • **General and reference** → Surveys and overviews; • **Computing methodologies** → Machine learning approaches; Machine learning algorithms.

Additional Key Words and Phrases: Tertiary study, machine learning, software engineering, systematic literature review

ACM Reference Format:

Zoe Kotti, Rafaila Galanopoulou, and Diomidis Spinellis. 2022. Machine Learning for Software Engineering: A Tertiary Study. *ACM Comput. Surv.* 1, 1 (November 2022), 37 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine learning (ML) is a thriving discipline with various practical applications and active research topics, many of which nowadays entangle the discipline of software engineering (SE) [113]. Through ML we can address SE problems that cannot be completely algorithmically modeled, or for which existing solutions do not provide satisfactory results yet (*e.g.*, defect/fault detection [16, 165, 180]). In addition, ML finds application in SE tasks where data cannot be easily analyzed with other algorithms (*e.g.*, software requirements, code comments, code reviews, issues [9, 91, 174]). Another important aspect of ML is that it can significantly reduce manual effort in common SE tasks (*e.g.*, automatic program repair [157], code suggestion [61], defect prediction [19], malware detection [147], feature location [40]) with great accuracy results [146, 164]. In fields such as health informatics ML and SE are considered complementary disciplines, since the growing scale and complexity of healthcare datasets have posed a challenge for clinical practice and medical research, requiring new engineering approaches from both fields [38].

In the early nineties, Huff and Selfridge [68] recognized the need for creating software systems that partially take some responsibility for their own evolution, offering the ability to implement, measure, and assess changes easily. These changes should also contribute to the overall improvement of the corresponding systems [142]. Around the same time, Brooks [29] prompted software practitioners to investigate evolutionary advancements rather than waiting for

Authors' address: Zoe Kotti; Rafaila Galanopoulou; Diomidis Spinellis, {zoekotti,rgalanopoulou,dds}@aueb.gr, Athens University of Economics and Business, Patission 76, Athens, Greece, 10434.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

Manuscript submitted to ACM

Semantic Scholar

dblp
computer science bibliography

ACM DL DIGITAL LIBRARY

Scopus®

ScienceDirect®

Clarivate™

Google Scholar

ResearchGate

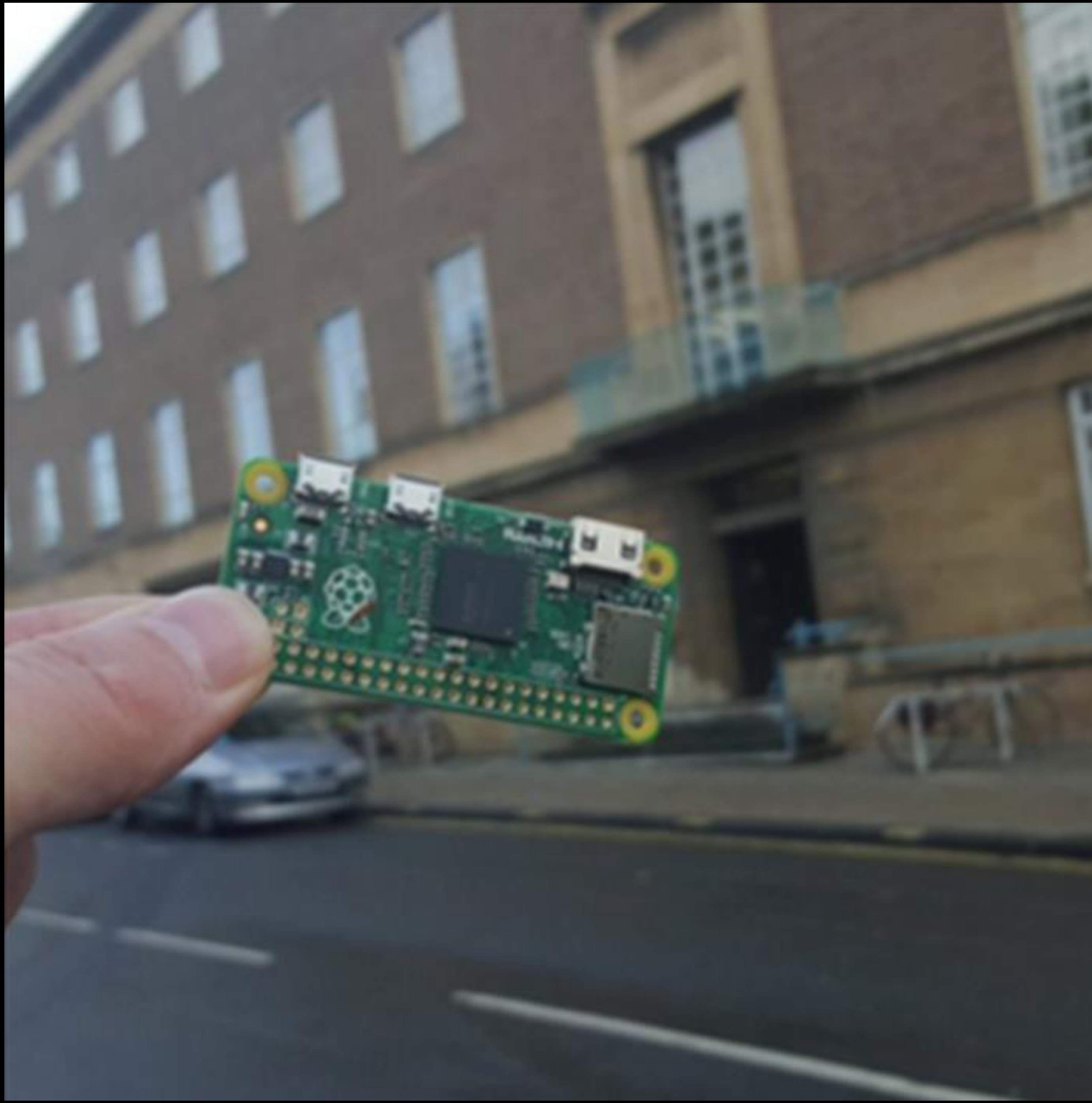
PubMed.gov

Issues

- Lack of transparency, repeatability, reproducibility
- High latency, low bandwidth
- Rate limits
- Proprietary and restricted query languages
- Limited coverage
- Availability and cost

2

ELLIOTT
ELECTRONIC COMPUTER
CITY TREASURERS DEPARTMENT



Measure	Elliott 405	Raspberry Pi Zero
Year	1957	2015
Price	£85,000 (1957) — €2M (2018)	$5
Instruction cycle time	10.71–0.918 ms (93-1089 Hz)	1 ns (1 GHz clock)
Main memory	16 kB drum store	512 MB LPDDR2 SDRAM
Fast memory	1280 bytes (nickel delay lines)	32 kB (16 kB I + 16kB D L1 cache)
Secondary memory	1.2 MB (300,000 word magnetic film)	8 GB (typical micro SD flash card)
Output bandwidth	25 characters/s	373 MB/s (1080p60 HDMI)
Weight	3–6 tons	9 g
Size	21 cabinets, each 2m x 77cm x 77cm	65mm x 30mm x 5.4mm
Operating power	10 kW	0.7 W

Crossref

ORCID

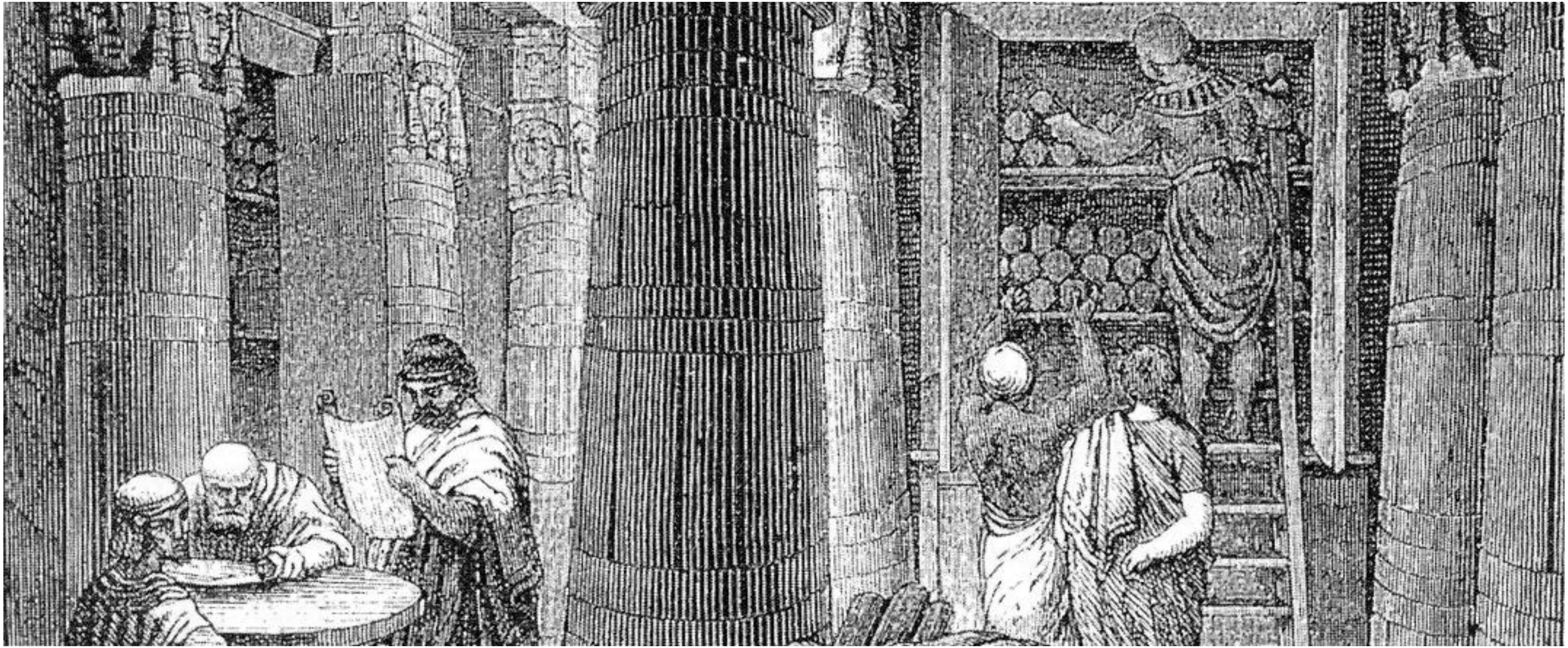
Connecting Research
and Researchers

UNITED STATES PATENT AND TRADEMARK OFFICE
DEPARTMENT OF COMMERCE

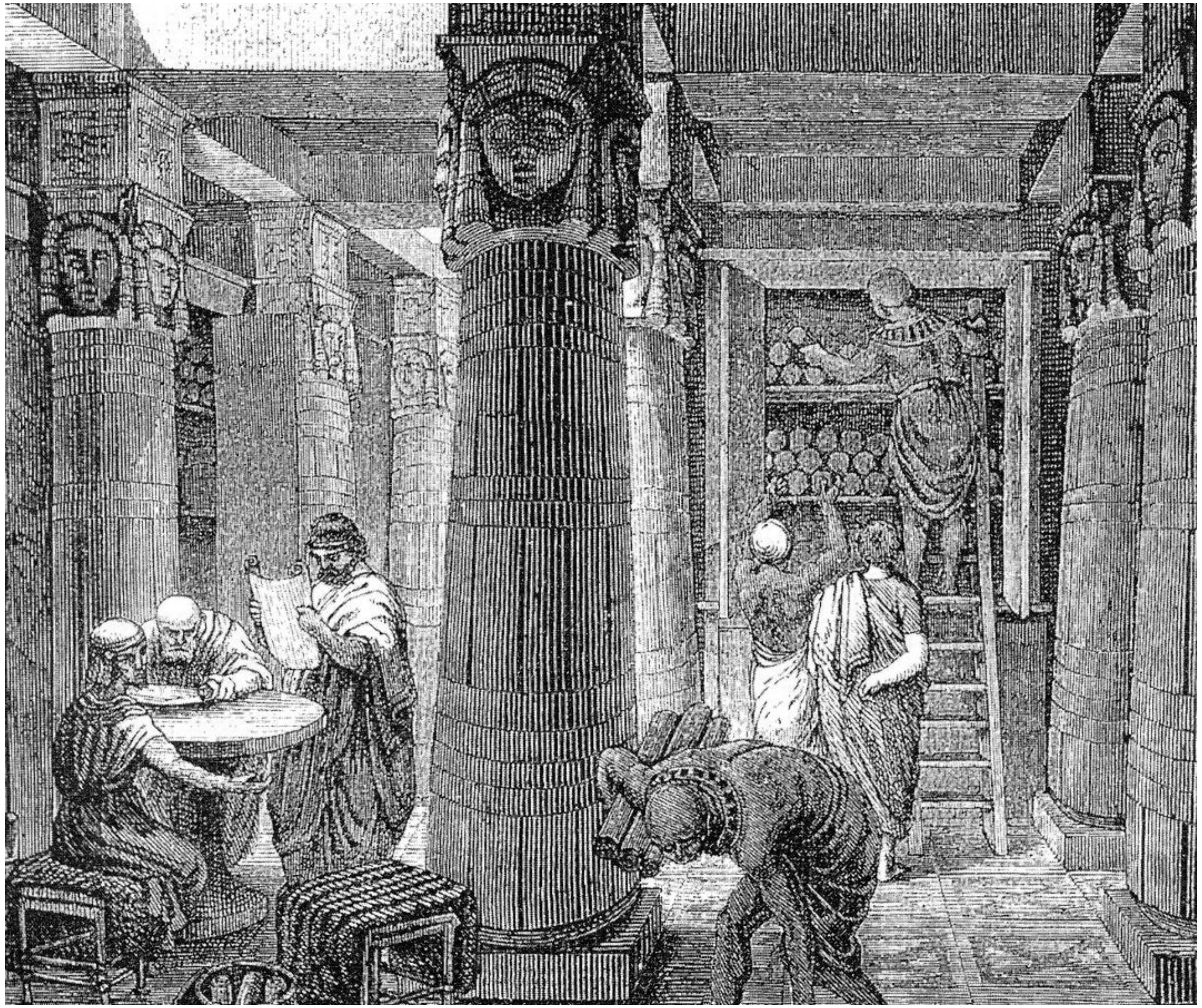
PubMed.gov

ROR

Alexandria3k



Publication metadata analytics on the desktop

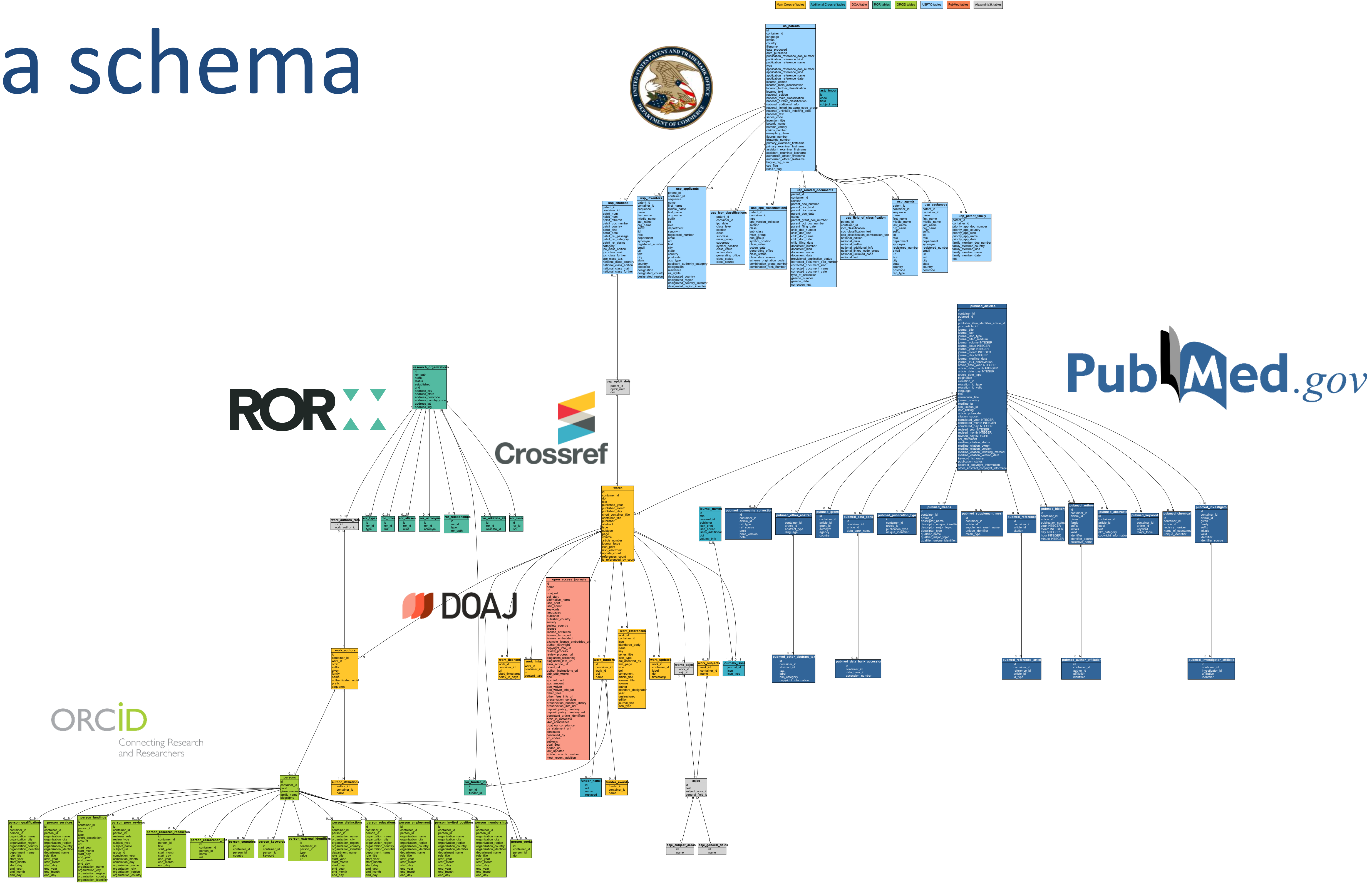


- Relational access to 1.9 TB of data
- 4.2 billion records in 74 tables
- Installed as a single Python module
- No (graph) database / cluster to install / maintain
- Efficient
  - Data sample queries run in minutes
  - Data building of full data slices in 5 h–2 days
  - Then queries run in seconds
  - Space requirements start at 157 GB for downloaded data

Agenda

- Data model and data
- Alexandria3k in practice
- Implementation
- Issues and limitations
- Way forward

Data schema



us_patents
id
container_id
language
status
country
filename
date_produced
date_published
publication_reference_doc_number
publication_reference_kind
publication_reference_name
type
application_reference_doc_number
application_reference_kind
application_reference_name
application_reference_date
locarno_edition
locarno_main_classification
locarno_further_classification
locarno_text
national_edition
national_main_classification
national_further_classification
national_additional_info
national_linked_indexing_code_group
national_unlinked_indexing_code
national_text
series_code
invention_title
botanic_name
botanic_variety
claims_number
exemplary_claim
figures_number
drawings_number
primary_examiner_firstname
primary_examiner_lastname
assistant_examiner_firstname
assistant_examiner_lastname
authorized_officer_firstname
authorized_officer_lastname
hague_reg_num
cpa_flag
rule47_flag

asjc_import
id
code
field
subject_area

usp_citations
patent_id
container_id
patcit_num
nplcit_num
nplcit_othercit
patcit_doc_number
patcit_country
patcit_kind
patcit_date
patcit_rel_passage
patcit_rel_category
patcit_rel_claims
category
ipc_class_edition
ipc_class_main
ipc_class_further
cpc_class_text
national_class_country
national_class_edition
national_class_main
national_class_further

usp_inventors
patent_id
container_id
sequence
name
first_name
middle_name
last_name
org_name
suffix
iid
role
department
synonym
registered_number
email
url
text
city
state
country
postcode
designation
designated_country
designated_region

usp_applicants
patent_id
container_id
sequence
name
first_name
middle_name
last_name
org_name
suffix
iid
role
department
synonym
registered_number
email
url
text
city
state
country
postcode
app_type
applicant_authority_category
designation
residence
us_rights
designated_country
designated_region
designated_country_inventor
designated_region_inventor

usp_icpr_classifications
patent_id
container_id
ipc_date
class_level
section
class
subclass
main_group
subgroup
symbol_position
class_value
action_date
generating_office
class_status
class_source

usp_cpc_classifications
patent_id
container_id
type
cpc_version_indicator
section
class
sub_class
main_group
sub_group
symbol_position
class_value
action_date
generating_office
class_status
class_data_source
scheme_origination_code
combination_group_number
combination_rank_number

usp_related_documents
patent_id
container_id
relation
parent_doc_number
parent_doc_kind
parent_doc_name
parent_doc_date
status
parent_grant_doc_number
parent_pct_doc_number
parent_filing_date
child_doc_number
child_doc_kind
child_doc_name
child_doc_date
child_filing_date
document_number
document_kind
document_name
document_date
provisional_application_status
corrected_document_doc_number
corrected_document_kind
corrected_document_name
corrected_document_date
type_of_correction
gazette_number
gazette_date
correction_text

usp_field_of_classification
patent_id
container_id
ipcr_classification
cpc_classification_text
cpc_classification_combination_text
national_edition
national_main
national_further
national_additional_info
national_linked_code_group
national_unlinked_code
national_text

usp_agents
patent_id
container_id
sequence
name
first_name
middle_name
last_name
org_name
suffix
iid
role
department
synonym
registered_number
email
url
text
city
state
country
postcode
rep_type

usp_assignees
patent_id
container_id
name
first_name
middle_name
last_name
org_name
suffix
iid
role
department
synonym
registered_number
email
url
text
city
state
country
postcode

usp_patent_family
patent_id
container_id
priority_app_doc_number
priority_app_country
priority_app_kind
priority_app_name
priority_app_date
family_member_doc_number
family_member_country
family_member_kind
family_member_name
family_member_date
text

usp_nplcit_dois
patent_id
nplcit_num
doi

research_organizations
id
ror_path
name
status
established
grid
address_city
address_state
address_postcode
address_country_code
address_lat
address_lng

work_authors_rors
ror_id
work_author_id

ror_types
id
ror_id
type

ror_links
id
ror_id
link

ror_aliases
id
ror_id
alias

ror_acronyms
id
ror_id
acronym

ror_relationships
id
ror_id
type
ror_path

ror_wikidata_ids
id
ror_id
wikidata_id

ror_isnis
id
ror_id
isni

works
id
container_id
doi
title
published_year
published_month
published_day
short_container_title
container_title
publisher
abstract
type
subtype
page
volume
article_number
journal_issue
issn_print
issn_electronic
update_count
references_count
is_referenced_by_count

open_access_journals
id
name
url
doaj_url
oaj_start
alternative_name
issn_print
issn_eprint
keywords
languages
publisher
pubisher_country
society
society_country
license
license_attributes
license_terms_url
license_embedded
example_license_embedded_url
author_copyright
copyright_info_url
review_process
review_process_url
plagiarism_screening
plagiarism_info_url
aims_scope_url
board_url
author_instructions_url
sub_pub_weeks
apc
apc_info_url
apc_amount
apc_waiver
apc_waiver_info_url
other_fees
other_fees_info_url
preservation_services
preservation_national_library
preservation_info_url
deposit_policy_directory
deposit_policy_directory_url
persistent_article_identifiers
orcid_in_metadata
i4oc_compliance
doaj_oa_compliance
oa_statement_url
continues
continued_by
lcc_codes
subjects
doaj_Seal
added_on
last_updated
article_records_number
most_recent_addition

work_authors
id
container_id
work_id
orcid
suffix
given
family
name
authenticated_orcid
prefix
sequence

work_licenses
work_id
container_id
url
start_timestamp
delay_in_days

work_links
work_id
container_id
url
content_type

work_funders
id
container_id
work_id
doi
name

work_references
work_id
container_id
issn
standards_body
issue
key
series_title
isbn_type
doi_asserted_by
first_page
isbn
doi
component
article_title
volume_title
volume
author
standard_designator
year
unstructured
edition
journal_title
issn_type

work_updates
work_id
container_id
label
doi
timestamp

works_asjcs
work_id
asjc_id

work_subjects
work_id
container_id
name

journals_issns
journal_id
issn
issn_type

journal_names
id
title
crossref_id
publisher
issn_print
issn_eprint
issns_additional
doi
volume_info

author_affiliations
author_id
container_id
name

ror_funder_ids
id
ror_id
funder_id

funder_names
id
url
name
replaced

funder_awards
funder_id
container_id
name

asjcs
id
field
subject_area_id
general_field_id

asjc_subject_areas
id
name

asjc_general_fields
id
name

persons
id
container_id
orcid
given_names
family_name
biography

person_qualifications
id
container_id
person_id
organization_name
organization_city
organization_region
organization_country
organization_identifier
department_name
role_title
start_year
start_month
start_day
end_year
end_month
end_day

person_services
id
container_id
person_id
organization_name
organization_city
organization_region
organization_country
organization_identifier
department_name
role_title
start_year
start_month
start_day
end_year
end_month
end_day

person_fundings
id
container_id
person_id
title
type
short_description
amount
url
start_year
start_month
start_day
end_year
end_month
end_day
organization_name
organization_city
organization_region
organization_country
organization_identifier

person_peer_reviews
id
container_id
person_id
reviewer_role
review_type
subject_type
subject_name
subject_url
group_id
completion_year
completion_month
completion_day
organization_name
organization_city
organization_region
organization_country

person_research_resources
id
container_id
person_id
title
start_year
start_month
start_day
end_year
end_month
end_day

person_researcher_urls
id
container_id
person_id
name
url

person_countries
id
container_id
person_id
country

person_keywords
id
container_id
person_id
keyword

person_external_identifiers
id
container_id
person_id
type
value
url

person_distinctions
id
container_id
person_id
organization_name
organization_city
organization_region
organization_country
organization_identifier
department_name
role_title
start_year
start_month
start_day
end_year
end_month
end_day

person_educations
id
container_id
person_id
organization_name
organization_city
organization_region
organization_country
organization_identifier
department_name
role_title
start_year
start_month
start_day
end_year
end_month
end_day

person_employments
id
container_id
person_id
organization_name
organization_city
organization_region
organization_country
organization_identifier
department_name
role_title
start_year
start_month
start_day
end_year
end_month
end_day

person_invited_positions
id
container_id
person_id
organization_name
organization_city
organization_region
organization_country
organization_identifier
department_name
role_title
start_year
start_month
start_day
end_year
end_month
end_day

person_memberships
id
container_id
person_id
organization_name
organization_city
organization_region
organization_country
organization_identifier
department_name
role_title
start_year
start_month
start_day
end_year
end_month
end_day

person_works
id
container_id
person_id
doi

pubmed_articles
id
container_id
pubmed_id
doi
publisher_item_identifier_article_id
pmc_article_id
journal_title
journal_issn
journal_issn_type
journal_cited_medium
journal_volume INTEGER
journal_issue INTEGER
journal_year INTEGER
journal_month INTEGER
journal_day INTEGER
journal_medline_date
journal_ISO_abbreviation
article_date_year INTEGER
article_date_month INTEGER
article_date_day INTEGER
article_date_type
pagination
elocation_id
elocation_id_type
elocation_id_valid
language
title
vernacular_title
journal_country
medline_ta
nlm_unique_id
issn_linking
article_pubmodel
citation_subset
completed_year INTEGER
completed_month INTEGER
completed_day INTEGER
revised_year INTEGER
revised_month INTEGER
revised_day INTEGER
coi_statement
medline_citation_status
medline_citation_owner
medline_citation_version
medline_citation_indexing_method
medline_citation_version_date
keyword_list_owner
publication_status
abstract_copyright_information
other_abstract_copyright_informatio

pubmed_comments_correction
id
container_id
article_id
ref_type
ref_source
pmid
pmid_version
note

pubmed_other_abstract
id
container_id
article_id
abstract_type
language

pubmed_grants
id
container_id
article_id
grant_id
acronym
agency
country

pubmed_data_bank
id
container_id
article_id
data_bank_name

pubmed_publication_type
id
container_id
article_id
publication_type
unique_identifier

pubmed_meshs
id
container_id
article_id
descriptor_name
descriptor_unique_identifie
descriptor_major_topic
descriptor_type
qualifier_name
qualifier_major_topic
qualifier_unique_identifier

pubmed_supplement_mesh
id
container_id
article_id
supplement_mesh_name
unique_identifier
mesh_type

pubmed_reference
id
container_id
article_id
citation

pubmed_history
id
container_id
article_id
publication_status
year INTEGER
month INTEGER
day INTEGER
hour INTEGER
minute INTEGER

pubmed_authors
id
container_id
article_id
given
family
suffix
initials
valid
identifier
identifier_source
collective_name

pubmed_abstracts
id
container_id
article_id
label
text
nlm_category
copyright_information

pubmed_keywords
id
container_id
article_id
keyword
major_topic

pubmed_chemicals
id
container_id
article_id
registry_number
name_of_substance
unique_identifier

pubmed_investigator
id
container_id
article_id
given
family
suffix
initials
valid
identifier
identifier_source

pubmed_other_abstract_tex
id
container_id
abstract_id
text
label
nlm_category
copyright_information

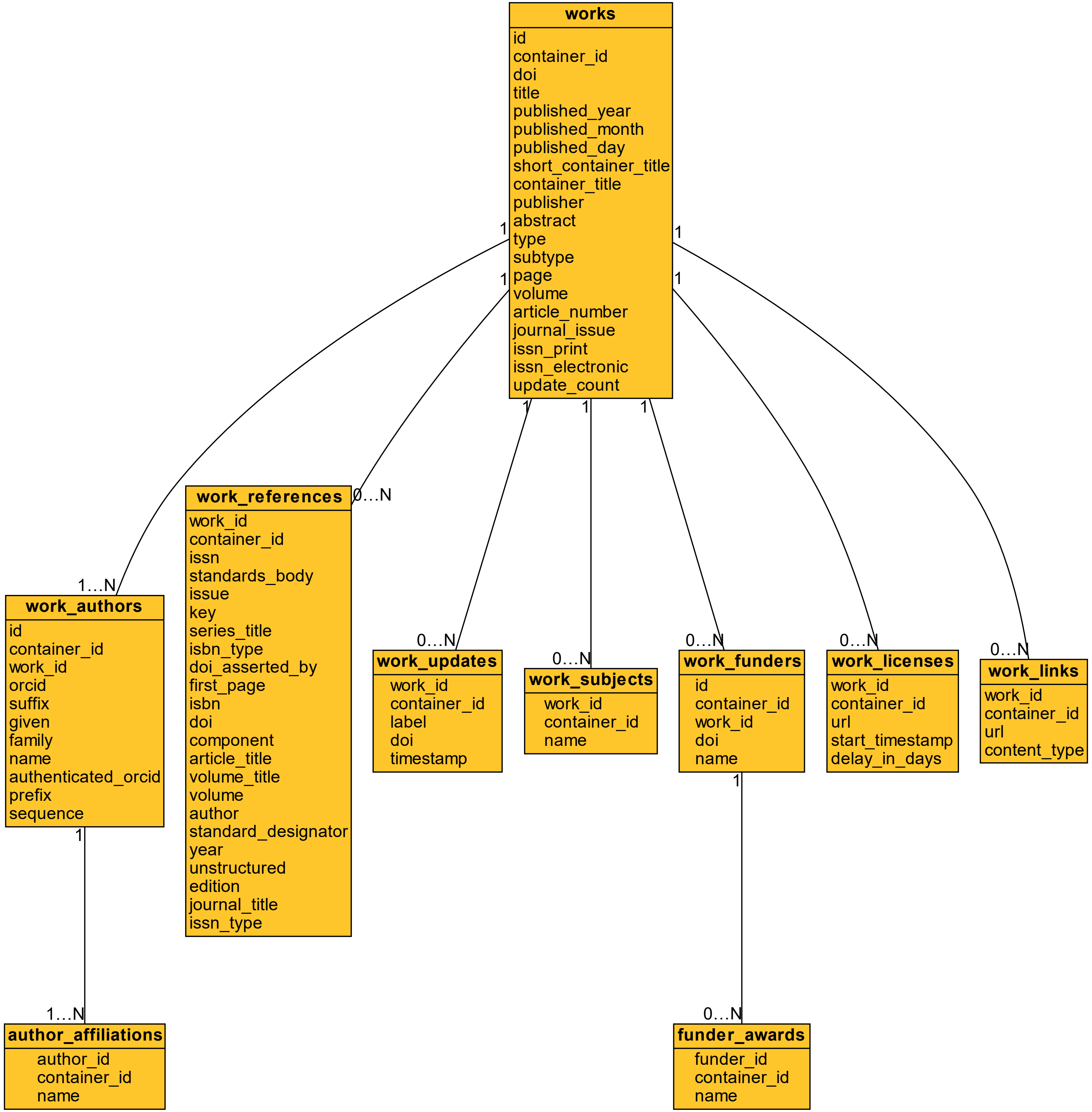
pubmed_data_bank_accession
id
container_id
data_bank_id
accession_number

pubmed_reference_articl
id
container_id
reference_id
article_id
id_type

pubmed_author_affiliation
id
container_id
author_id
affiliation
identifier

pubmed_investigator_affiliatio
id
container_id
investigator_id
affiliation
identifier

Crossref



works
- id
- container_id
- doi
- title
- published_year
- published_month
- published_day
- short_container_title
- container_title
- publisher
- abstract
- type
- subtype
- page
- volume
- article_number
- journal_issue
- issn_print
- issn_electronic
- update_count

work_authors
- id
- container_id
- work_id
- orcid
- suffix
- given
- family
- name
- authenticated_orcid
- prefix
- sequence

work_references
- work_id
- container_id
- issn
- standards_body
- issue
- key
- series_title
- isbn_type
- doi_asserted_by
- first_page
- isbn
- doi
- component
- article_title
- volume_title
- volume
- author
- standard_designator
- year
- unstructured
- edition
- journal_title
- issn_type

work_updates
- work_id
- container_id
- label
- doi
- timestamp

work_subjects
- work_id
- container_id
- name

work_funders
- id
- container_id
- work_id
- doi
- name

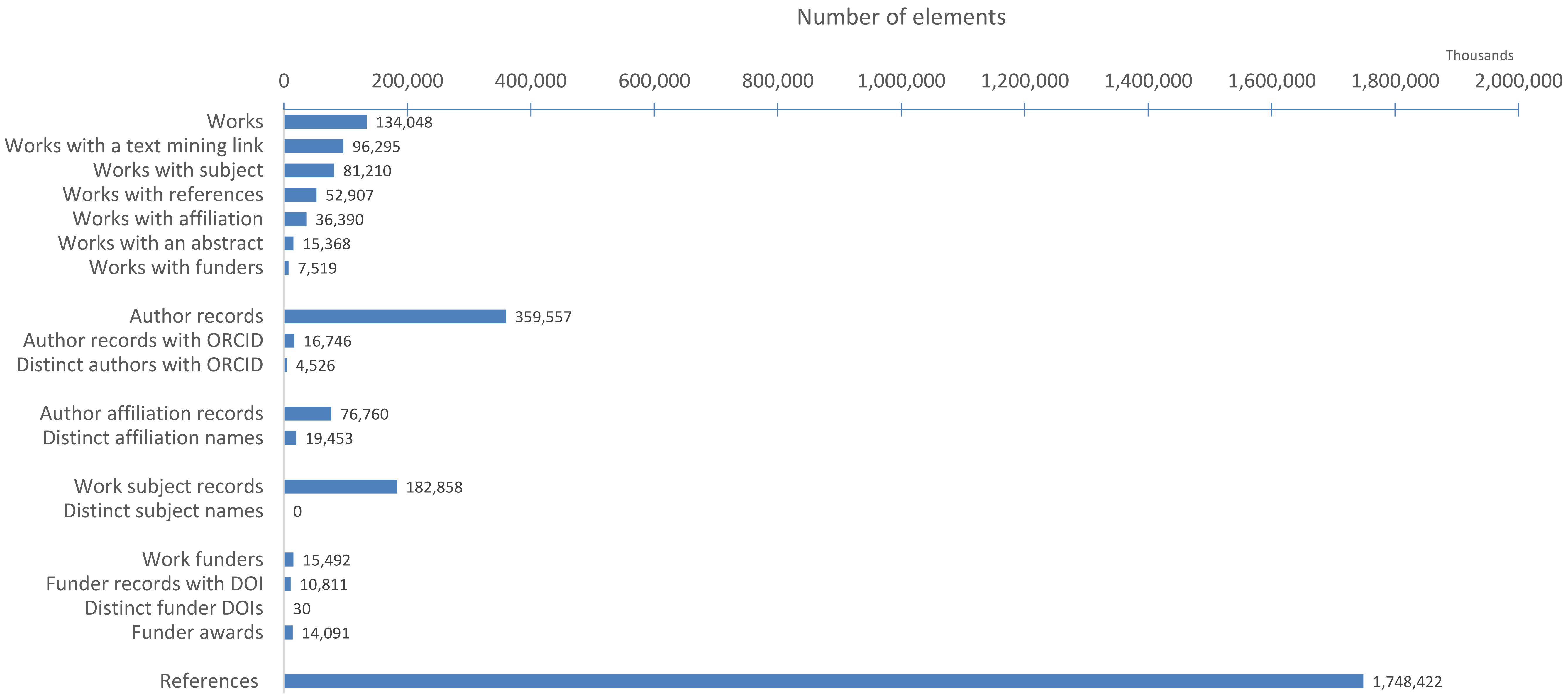
work_licenses
- work_id
- container_id
- url
- start_timestamp
- delay_in_days

work_links
- work_id
- container_id
- url
- content_type

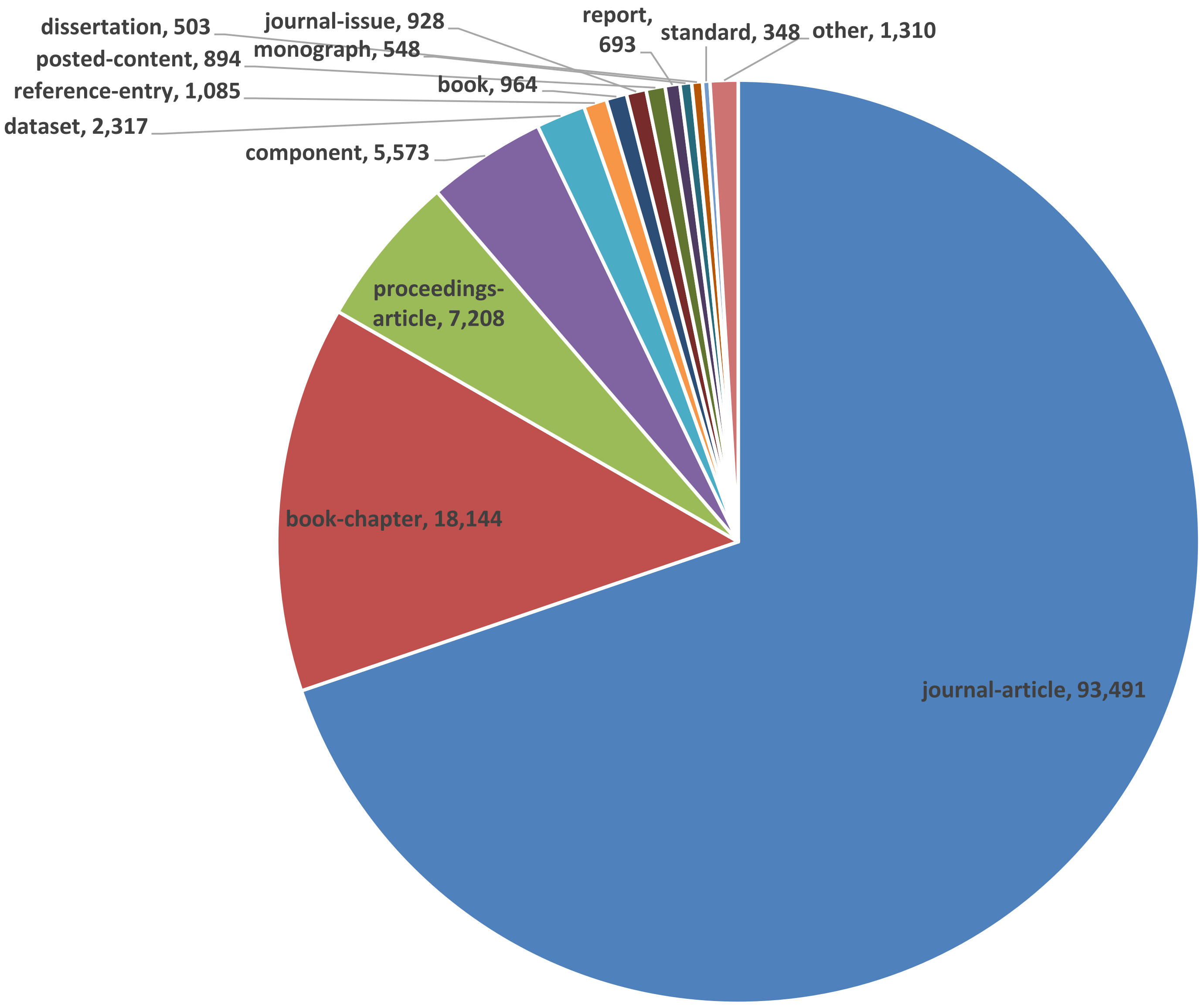
author_affiliations
- author_id
- container_id
- name

funder_awards
- funder_id
- container_id
- name

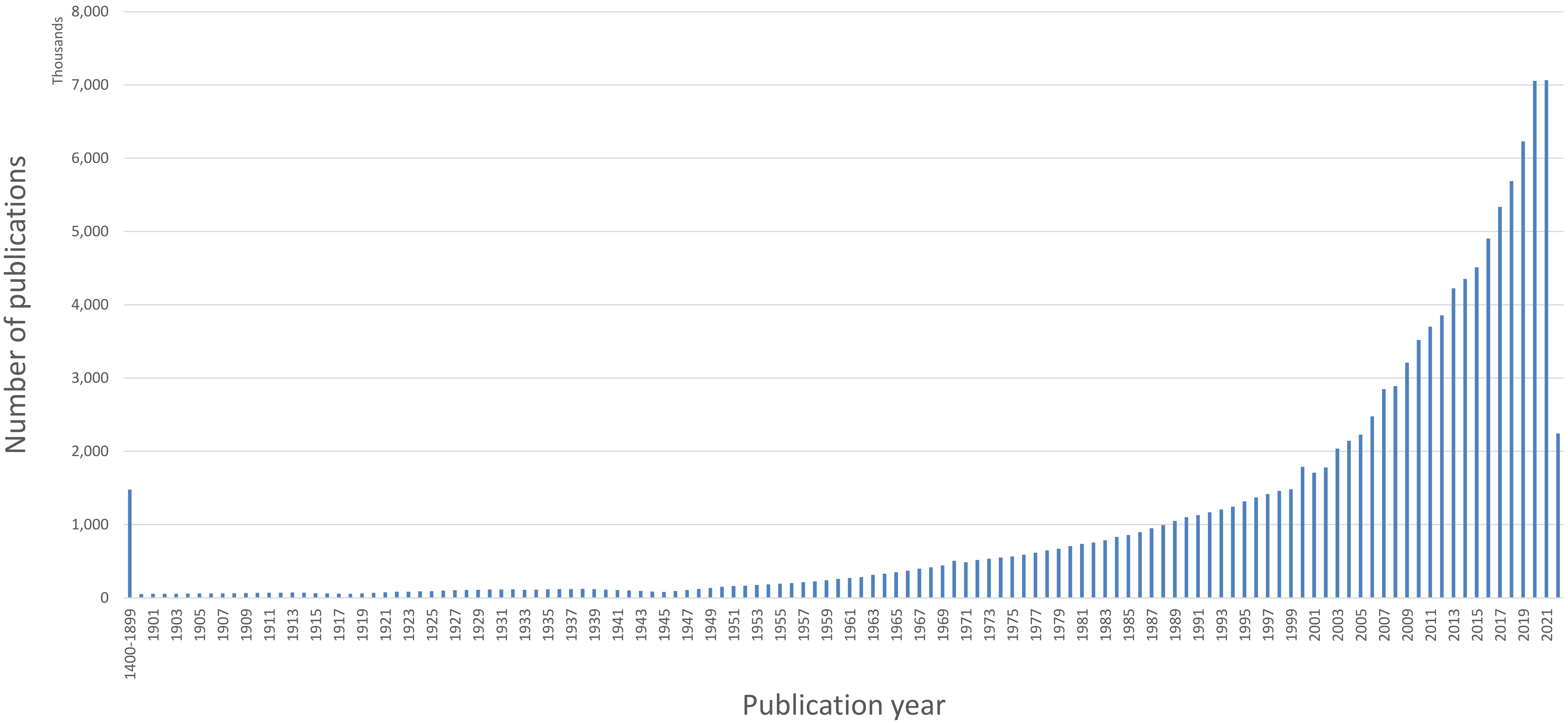
Crossref data in numbers



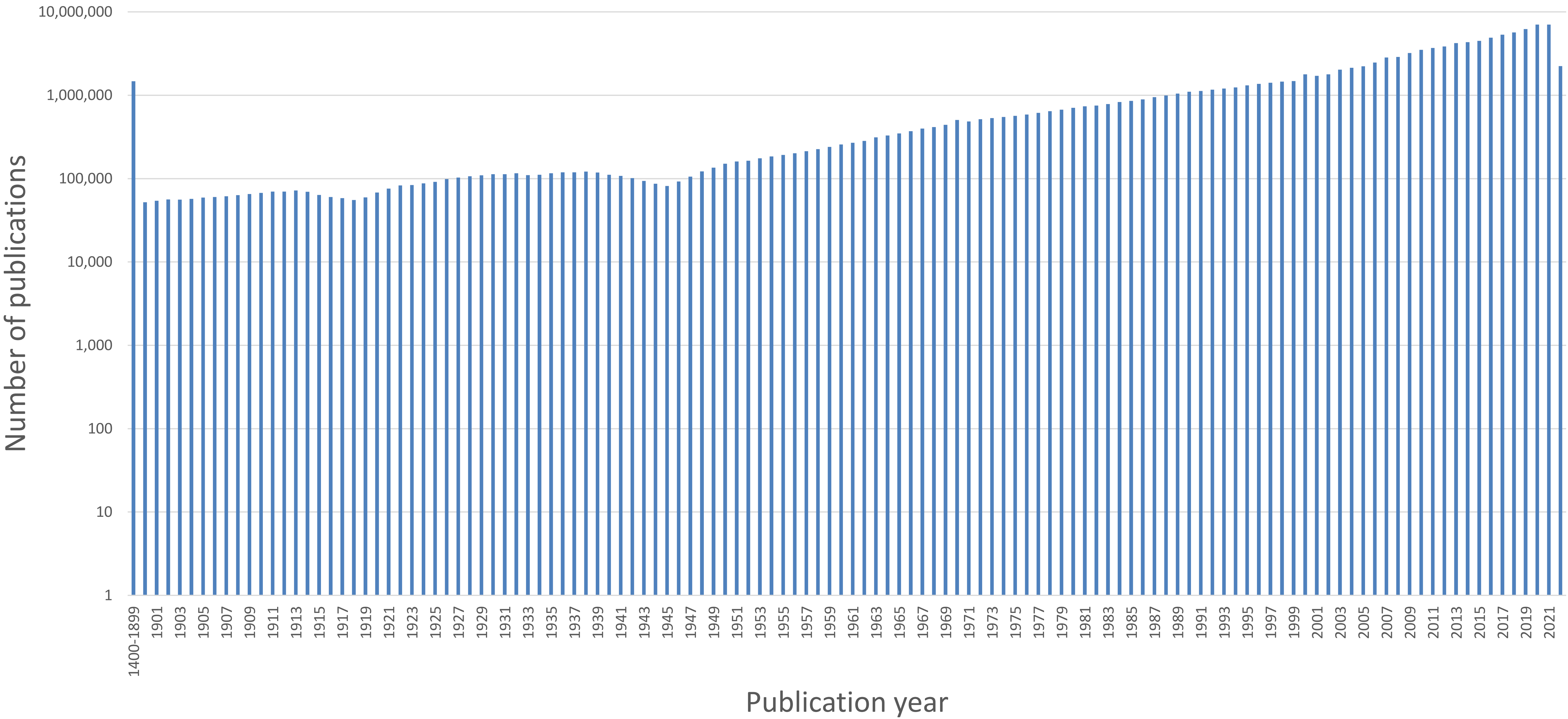
Crossref record types (thousands)

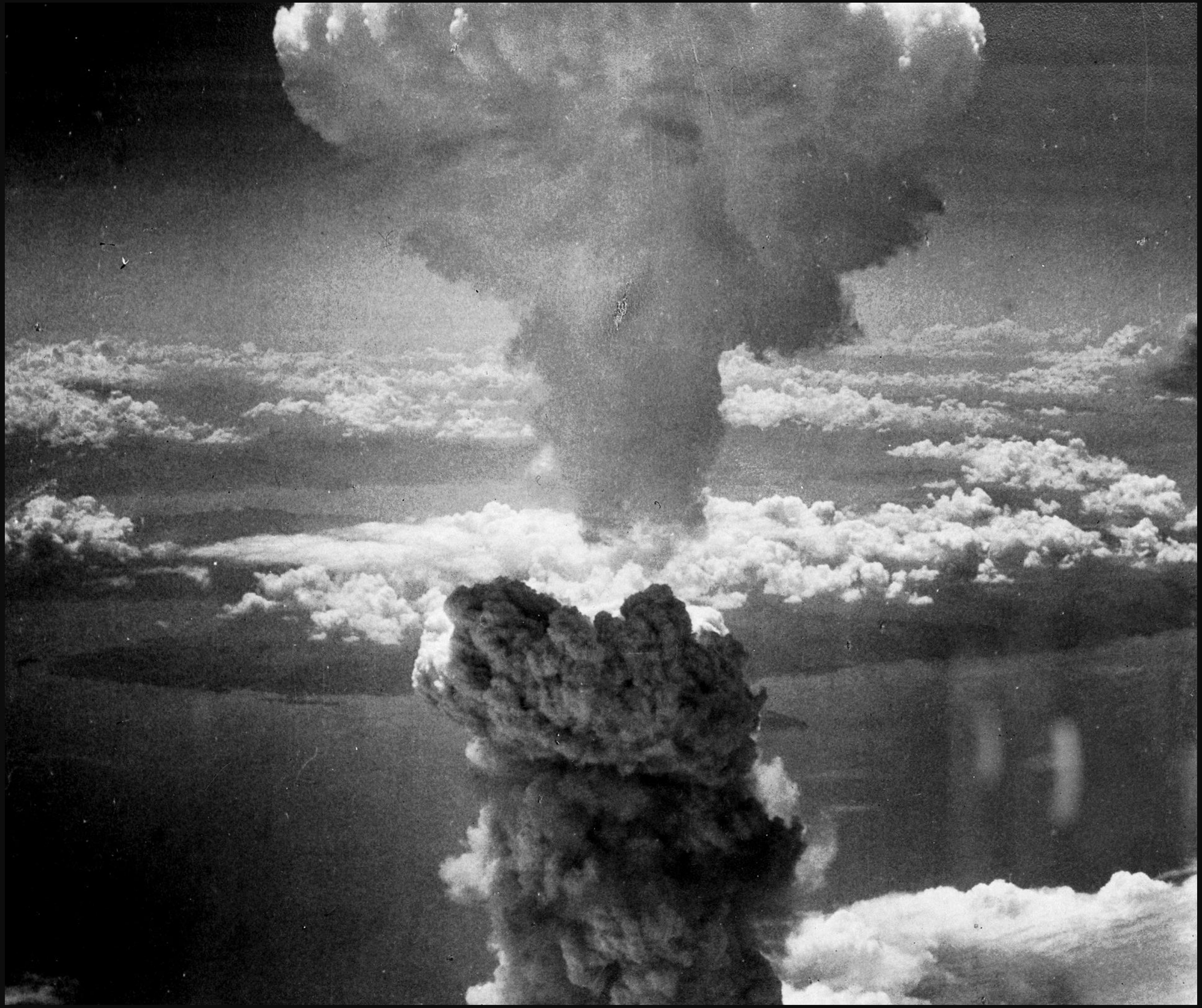


Crossref publications per year

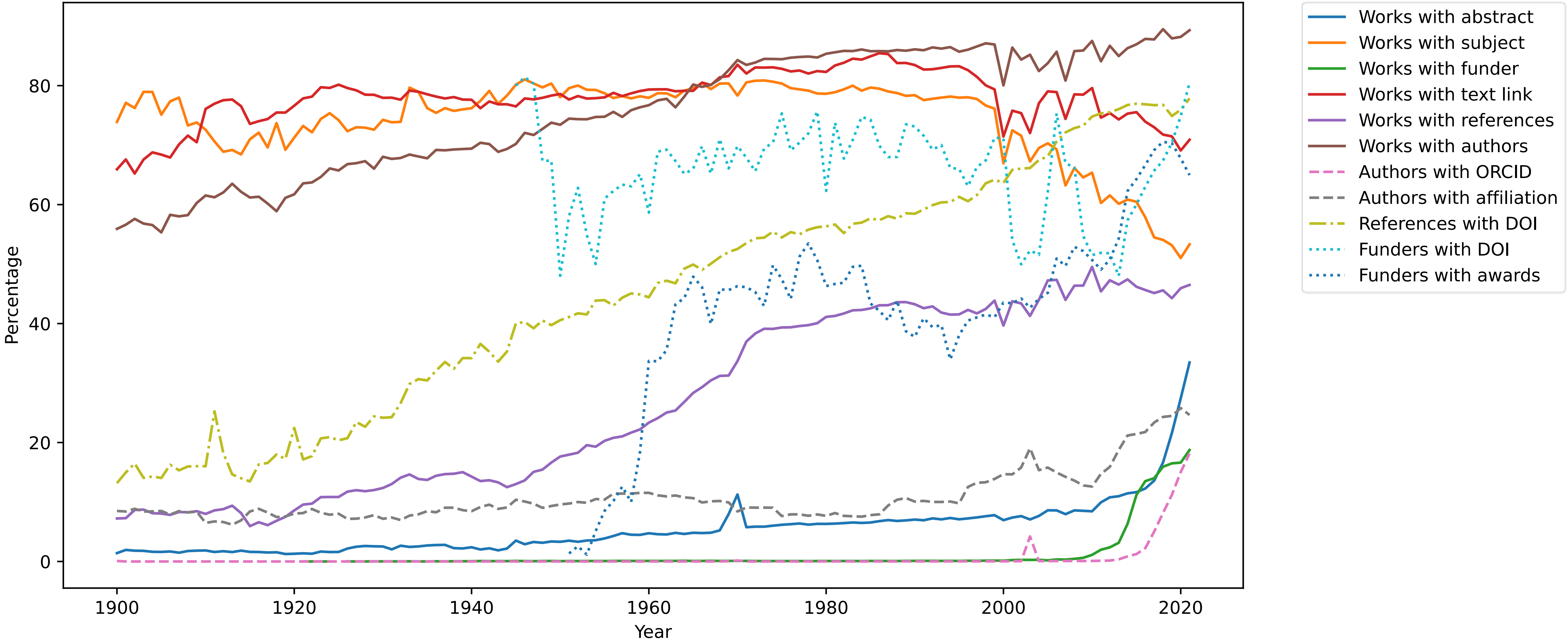


(Log scale)

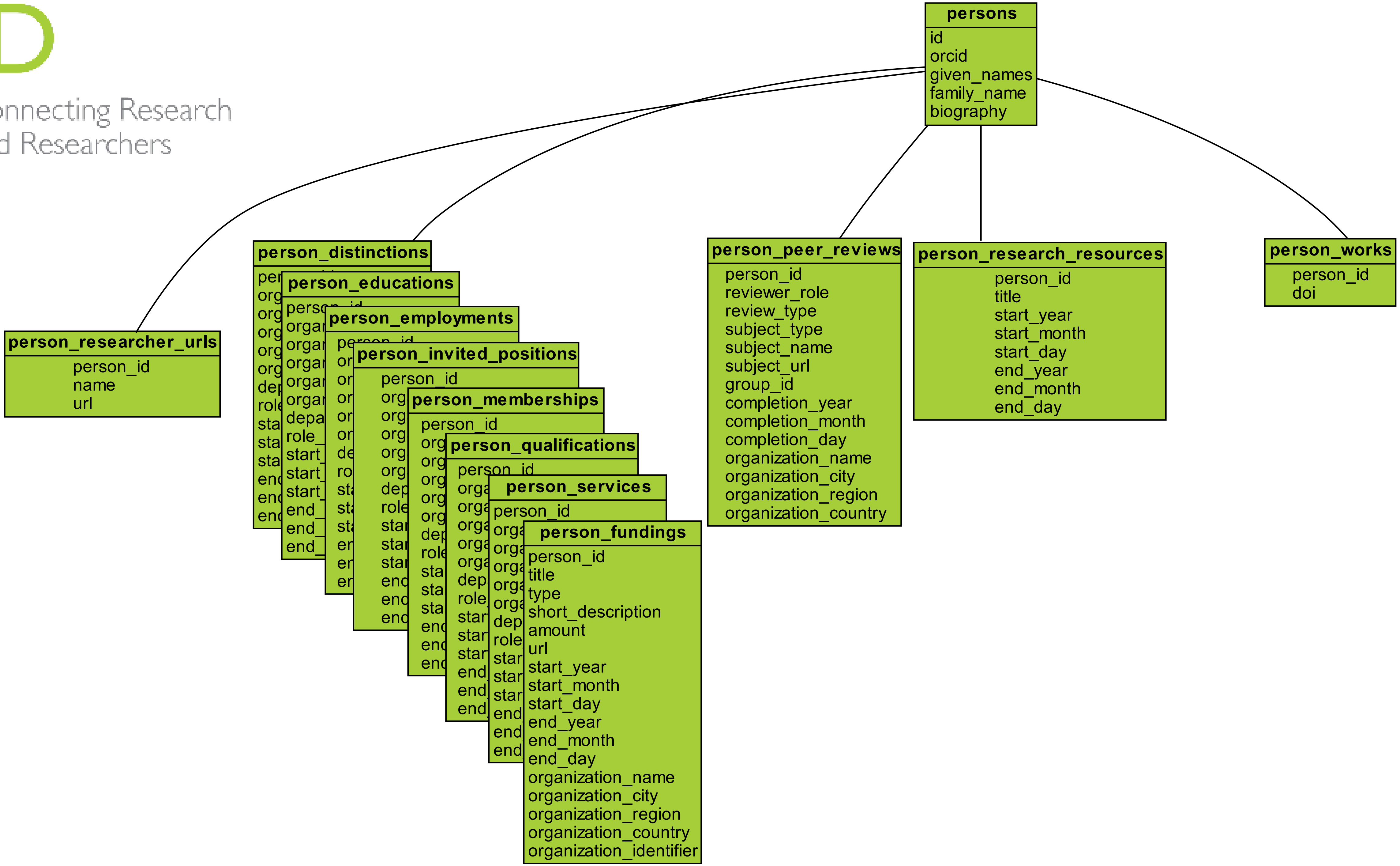

Number of publications
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
1400-1899
1901
1903
1905
1907
1909
1911
1913
1915
1917
1919
1921
1923
1925
1927
1929
1931
1933
1935
1937
1939
1941
1943
1945
1947
1949
1951
1953
1955
1957
1959
1961
1963
1965
1967
1969
1971
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
2021
Publication year



Data availability



Connecting Research
and Researchers



persons
- id
- orcid
- given_names
- family_name
- biography

person_researcher_urls
- person_id
- name
- url

person_distinctions

person_educations

person_employments

person_invited_positions
- person_id

person_memberships
- person_id

person_qualifications
- person_id

person_services
- person_id

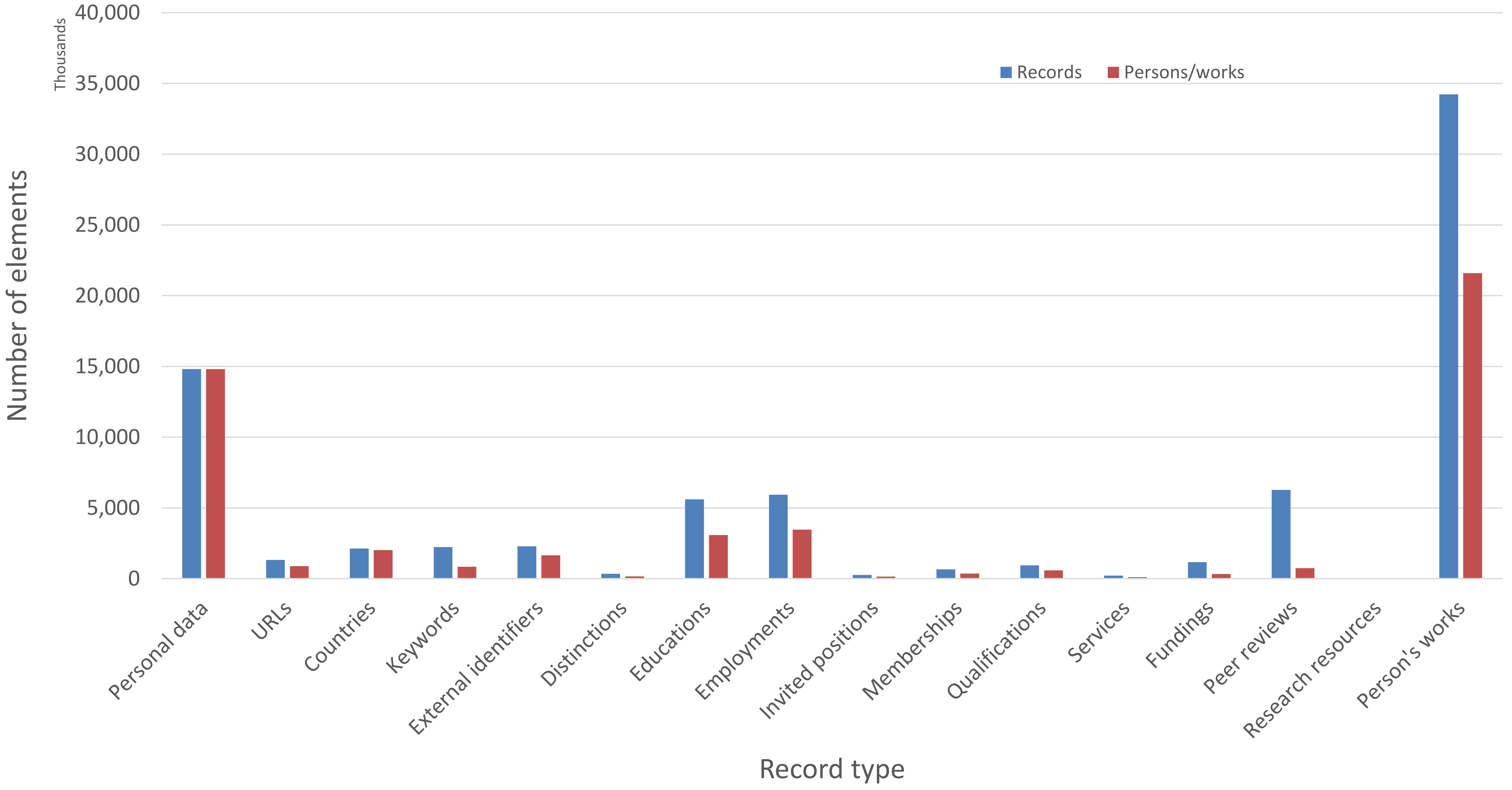
person_fundings
- person_id
- title
- type
- short_description
- amount
- url
- start_year
- start_month
- start_day
- end_year
- end_month
- end_day
- organization_name
- organization_city
- organization_region
- organization_country
- organization_identifier

person_peer_reviews
- person_id
- reviewer_role
- review_type
- subject_type
- subject_name
- subject_url
- group_id
- completion_year
- completion_month
- completion_day
- organization_name
- organization_city
- organization_region
- organization_country

person_research_resources
- person_id
- title
- start_year
- start_month
- start_day
- end_year
- end_month
- end_day

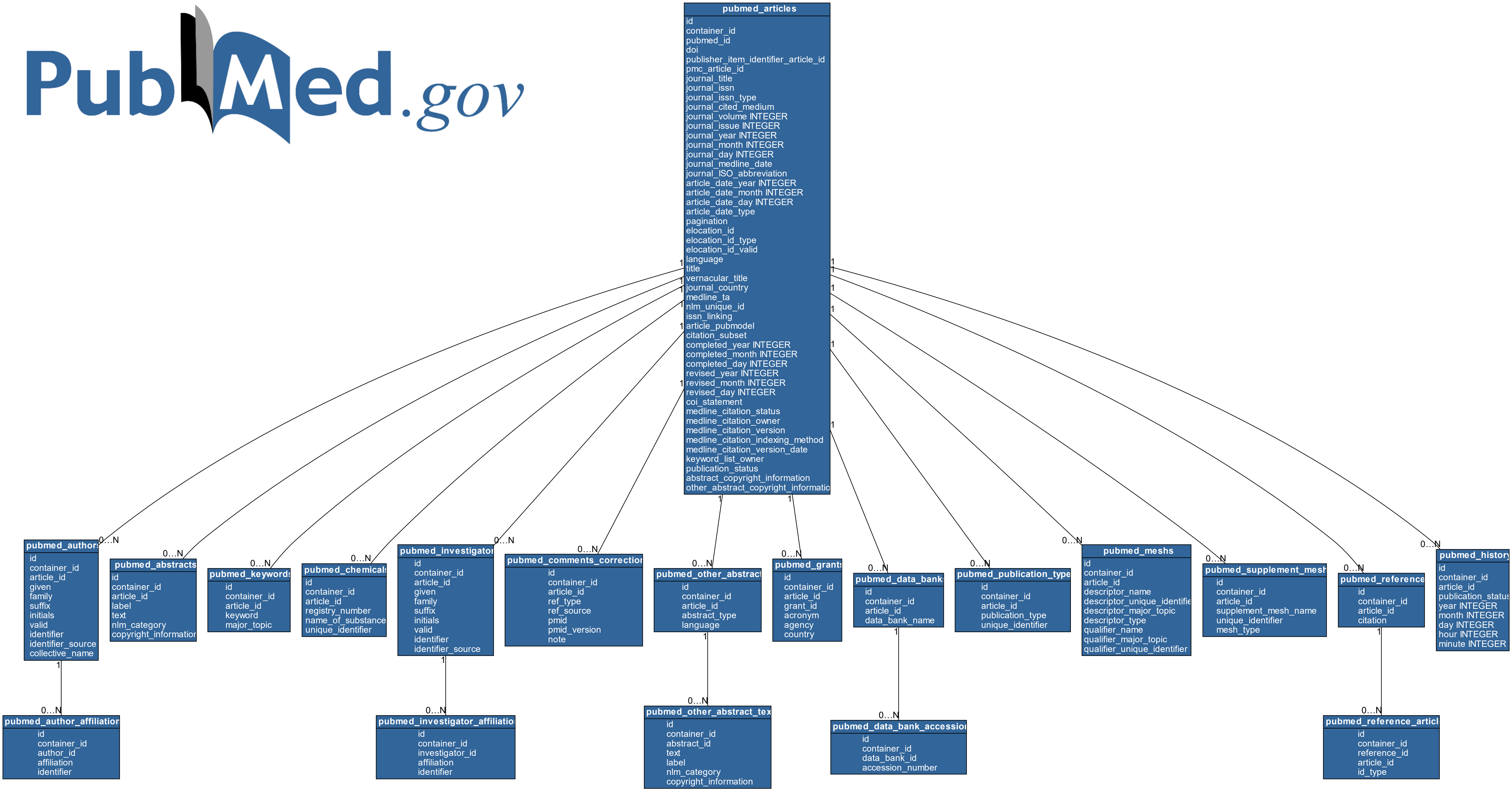
person_works
- person_id
- doi

ORCID data



Thousands
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
Number of elements
🟦 Records
🟥 Persons/works
Personal data
URLs
Countries
Keywords
External identifiers
Distinctions
Educations
Employments
Invited positions
Memberships
Qualifications
Services
Fundings
Peer reviews
Research resources
Person's works
Record type

PubMed.gov



pubmed_articles
id
container_id
pubmed_id
doi
publisher_item_identifier_article_id
pmc_article_id
journal_title
journal_issn
journal_issn_type
journal_cited_medium
journal_volume INTEGER
journal_issue INTEGER
journal_year INTEGER
journal_month INTEGER
journal_day INTEGER
journal_medline_date
journal_ISO_abbreviation
article_date_year INTEGER
article_date_month INTEGER
article_date_day INTEGER
article_date_type
pagination
elocation_id
elocation_id_type
elocation_id_valid
language
title
vernacular_title
journal_country
medline_ta
nlm_unique_id
issn_linking
article_pubmodel
citation_subset
completed_year INTEGER
completed_month INTEGER
completed_day INTEGER
revised_year INTEGER
revised_month INTEGER
revised_day INTEGER
coi_statement
medline_citation_status
medline_citation_owner
medline_citation_version
medline_citation_indexing_method
medline_citation_version_date
keyword_list_owner
publication_status
abstract_copyright_information
other_abstract_copyright_informatio

pubmed_authors
id
container_id
article_id
given
family
suffix
initials
valid
identifier
identifier_source
collective_name

pubmed_abstracts
id
container_id
article_id
label
text
nlm_category
copyright_informatior

pubmed_keywords
id
container_id
article_id
keyword
major_topic

pubmed_chemicals
id
container_id
article_id
registry_number
name_of_substance
unique_identifier

pubmed_investigator
id
container_id
article_id
given
family
suffix
initials
valid
identifier
identifier_source

pubmed_comments_correctior
id
container_id
article_id
ref_type
ref_source
pmid
pmid_version
note

pubmed_other_abstract
id
container_id
article_id
abstract_type
language

pubmed_grants
id
container_id
article_id
grant_id
acronym
agency
country

pubmed_data_banks
id
container_id
article_id
data_bank_name

pubmed_publication_type
id
container_id
article_id
publication_type
unique_identifier

pubmed_meshs
id
container_id
article_id
descriptor_name
descriptor_unique_identifie
descriptor_major_topic
descriptor_type
qualifier_name
qualifier_major_topic
qualifier_unique_identifier

pubmed_supplement_mesh
id
container_id
article_id
supplement_mesh_name
unique_identifier
mesh_type

pubmed_reference
id
container_id
article_id
citation

pubmed_history
id
container_id
article_id
publication_status
year INTEGER
month INTEGER
day INTEGER
hour INTEGER
minute INTEGER

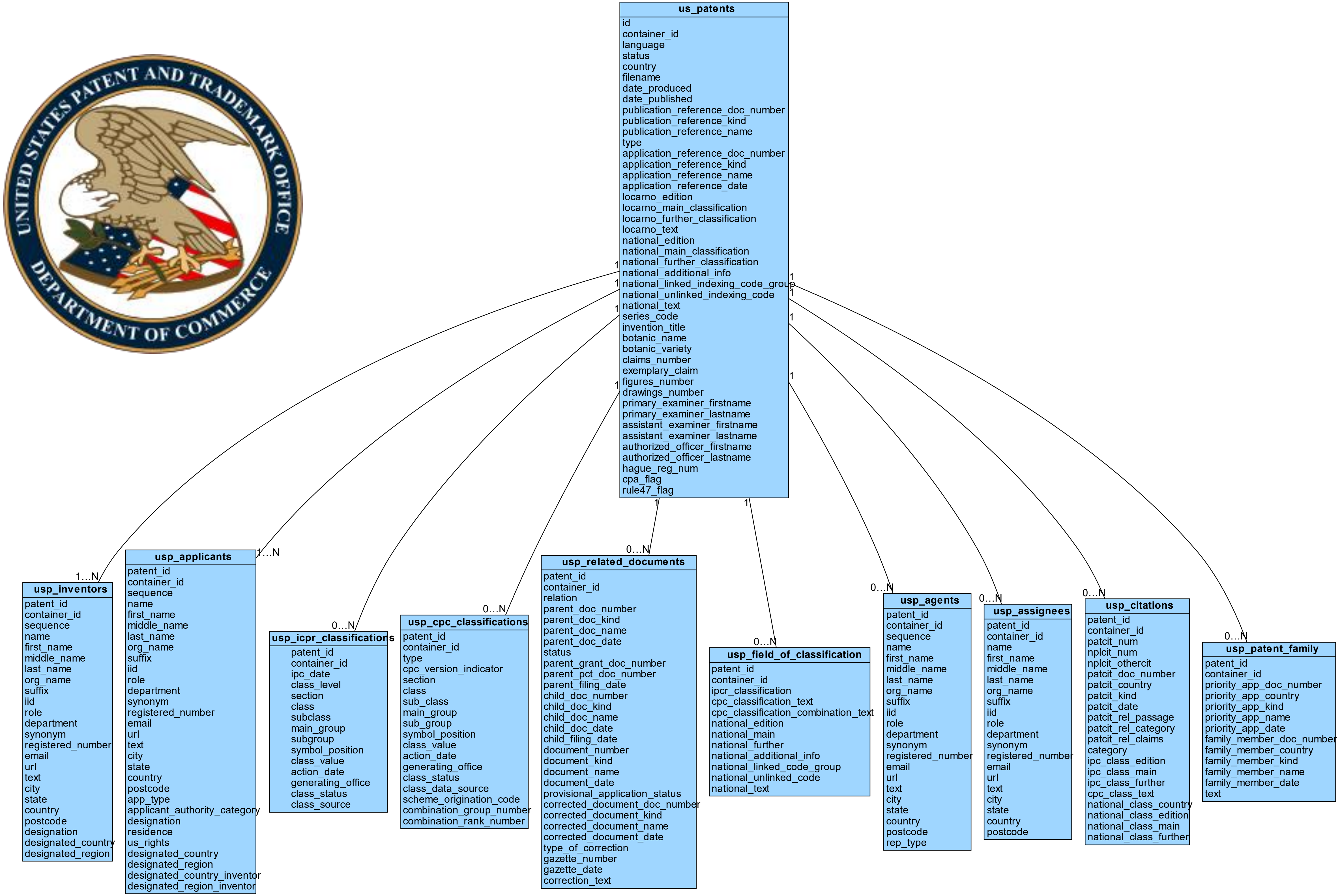
pubmed_author_affiliatior
id
container_id
author_id
affiliation
identifier

pubmed_investigator_affiliatio
id
container_id
investigator_id
affiliation
identifier

pubmed_other_abstract_tex
id
container_id
abstract_id
text
label
nlm_category
copyright_information

pubmed_data_bank_accessior
id
container_id
data_bank_id
accession_number

pubmed_reference_articl
id
container_id
reference_id
article_id
id_type



us_patents

id
container_id
language
status
country
filename
date_produced
date_published
publication_reference_doc_number
publication_reference_kind
publication_reference_name
type
application_reference_doc_number
application_reference_kind
application_reference_name
application_reference_date
locarno_edition
locarno_main_classification
locarno_further_classification
locarno_text
national_edition
national_main_classification
national_further_classification
national_additional_info
national_linked_indexing_code_group
national_unlinked_indexing_code
national_text
series_code
invention_title
botanic_name
botanic_variety
claims_number
exemplary_claim
figures_number
drawings_number
primary_examiner_firstname
primary_examiner_lastname
assistant_examiner_firstname
assistant_examiner_lastname
authorized_officer_firstname
authorized_officer_lastname
hague_reg_num
cpa_flag
rule47_flag

usp_inventors

patent_id
container_id
sequence
name
first_name
middle_name
last_name
org_name
suffix
iid
role
department
synonym
registered_number
email
url
text
city
state
country
postcode
designation
designated_country
designated_region

usp_applicants

patent_id
container_id
sequence
name
first_name
middle_name
last_name
org_name
suffix
iid
role
department
synonym
registered_number
email
url
text
city
state
country
postcode
app_type
applicant_authority_category
designation
residence
us_rights
designated_country
designated_region
designated_country_inventor
designated_region_inventor

usp_icpr_classifications

patent_id
container_id
ipc_date
class_level
section
class
subclass
main_group
subgroup
symbol_position
class_value
action_date
generating_office
class_status
class_source

usp_cpc_classifications

patent_id
container_id
type
cpc_version_indicator
section
class
sub_class
main_group
sub_group
symbol_position
class_value
action_date
generating_office
class_status
class_data_source
scheme_origination_code
combination_group_number
combination_rank_number

usp_related_documents

patent_id
container_id
relation
parent_doc_number
parent_doc_kind
parent_doc_name
parent_doc_date
status
parent_grant_doc_number
parent_pct_doc_number
parent_filing_date
child_doc_number
child_doc_kind
child_doc_name
child_doc_date
child_filing_date
document_number
document_kind
document_name
document_date
provisional_application_status
corrected_document_doc_number
corrected_document_kind
corrected_document_name
corrected_document_date
type_of_correction
gazette_number
gazette_date
correction_text

usp_field_of_classification

patent_id
container_id
ipcr_classification
cpc_classification_text
cpc_classification_combination_text
national_edition
national_main
national_further
national_additional_info
national_linked_code_group
national_unlinked_code
national_text

usp_agents

patent_id
container_id
sequence
name
first_name
middle_name
last_name
org_name
suffix
iid
role
department
synonym
registered_number
email
url
text
city
state
country
postcode
rep_type

usp_assignees

patent_id
container_id
name
first_name
middle_name
last_name
org_name
suffix
iid
role
department
synonym
registered_number
email
url
text
city
state
country
postcode

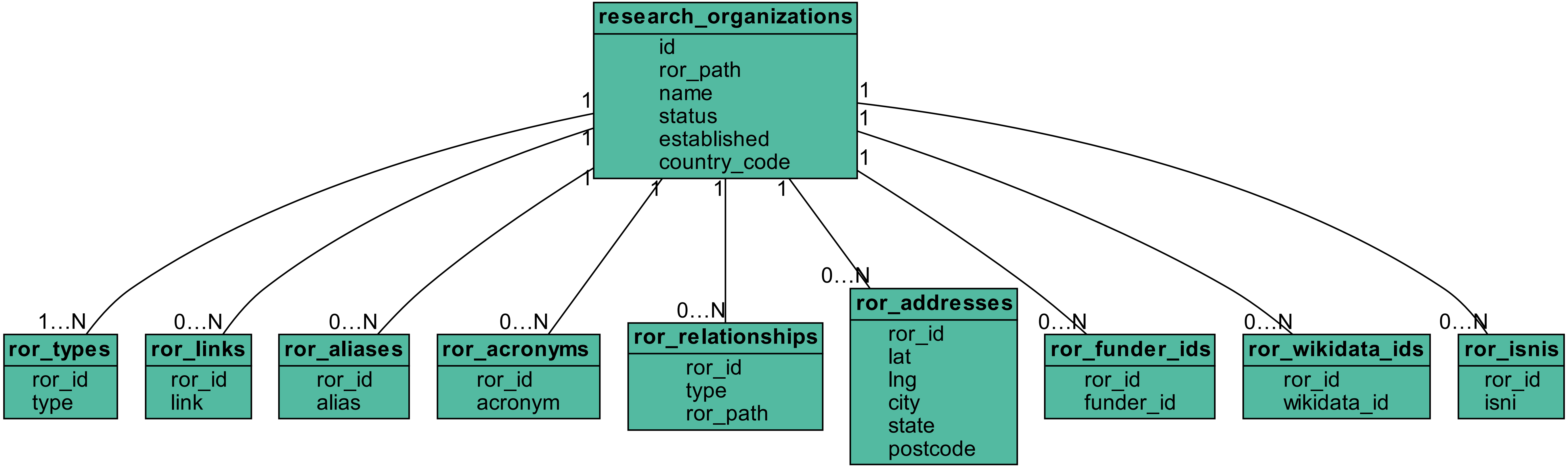
usp_citations

patent_id
container_id
patcit_num
nplcit_num
nplcit_othercit
patcit_doc_number
patcit_country
patcit_kind
patcit_date
patcit_rel_passage
patcit_rel_category
patcit_rel_claims
category
ipc_class_edition
ipc_class_main
ipc_class_further
cpc_class_text
national_class_country
national_class_edition
national_class_main
national_class_further

usp_patent_family

patent_id
container_id
priority_app_doc_number
priority_app_country
priority_app_kind
priority_app_name
priority_app_date
family_member_doc_number
family_member_country
family_member_kind
family_member_name
family_member_date
text

ROR



research_organizations
id
ror_path
name
status
established
country_code

ror_types
ror_id
type

ror_links
ror_id
link

ror_aliases
ror_id
alias

ror_acronyms
ror_id
acronym

ror_relationships
ror_id
type
ror_path

ror_addresses
ror_id
lat
lng
city
state
postcode

ror_funder_ids
ror_id
funder_id

ror_wikidata_ids
ror_id
wikidata_id

ror_isnis
ror_id
isni

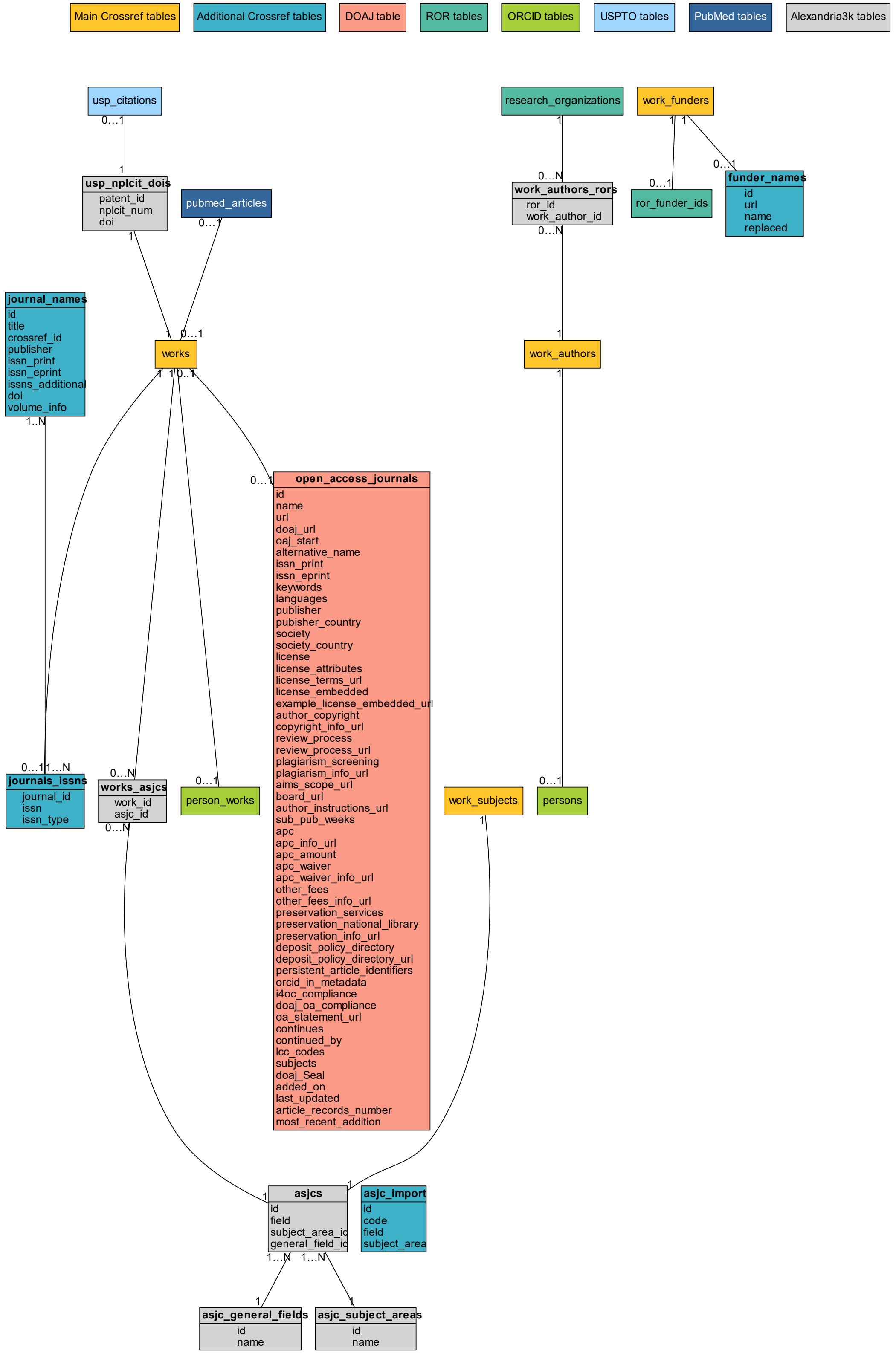
One m e thing...

Journals, Funders, Open Access

- Crossref journal names (109k records)
- Crossref funder names (21k records)
- DOAJ open access journal metadata (19k records)

Crossref

DOAJ


Main Crossref tables
Additional Crossref tables
DOAJ table
ROR tables
ORCID tables
USPTO tables
PubMed tables
Alexandria3k tables
usp_citations
0...1
1
usp_nplcit_dois
patent_id
nplcit_num
doi
1
pubmed_articles
0...1
research_organizations
1
0...N
work_authors_rors
ror_id
work_author_id
0...N
work_funders
1
1
0...1
ror_funder_ids
0...1
funder_names
id
url
name
replaced
journal_names
id
title
crossref_id
publisher
issn_print
issn_eprint
issns_additional
doi
volume_info
1..N
1
0...1
works
1
1
0..1
work_authors
1
1
0...1
open_access_journals
id
name
url
doaj_url
oaj_start
alternative_name
issn_print
issn_eprint
keywords
languages
publisher
pubisher_country
society
society_country
license
license_attributes
license_terms_url
license_embedded
example_license_embedded_url
author_copyright
copyright_info_url
review_process
review_process_url
plagiarism_screening
plagiarism_info_url
aims_scope_url
board_url
author_instructions_url
sub_pub_weeks
apc
apc_info_url
apc_amount
apc_waiver
apc_waiver_info_url
other_fees
other_fees_info_url
preservation_services
preservation_national_library
preservation_info_url
deposit_policy_directory
deposit_policy_directory_url
persistent_article_identifiers
orcid_in_metadata
i4oc_compliance
doaj_oa_compliance
oa_statement_url
continues
continued_by
lcc_codes
subjects
doaj_Seal
added_on
last_updated
article_records_number
most_recent_addition
0...1
1...N
journals_issns
journal_id
issn
issn_type
0...N
works_asjcs
work_id
asjc_id
0...N
0...1
person_works
work_subjects
1
0...1
persons
1
1
asjcs
id
field
subject_area_id
general_field_id
asjc_import
id
code
field
subject_area
1...N
1...N
1
1
asjc_general_fields
id
name
asjc_subject_areas
id
name

Alexandria3k in practice

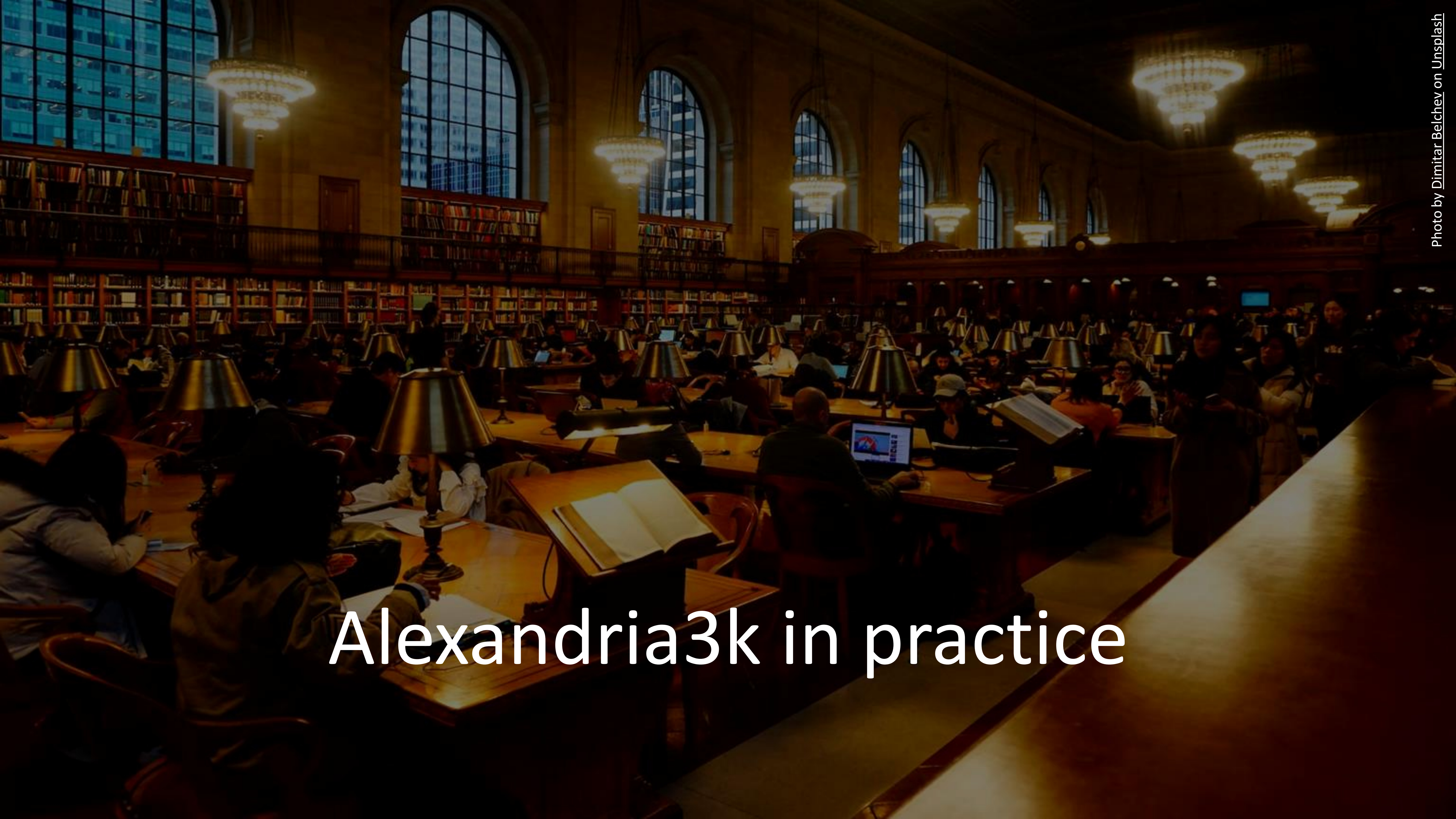


Photo by Dimitar Belchev on Unsplash

CLI usage

```
usage: a3k [-h] [-d DEBUG] [-v]
           {help,populate,process,query,list-processes,list-complete-schema,list-source-schema,list-process-schema,list-sources,version}
           ...

a3k: Relational interface to publication metadata

positional arguments:
  {help,populate,process,query,list-processes,list-complete-schema,list-source-schema,list-process-schema,list-sources,version}
                        Name of the a3k operation to perform.
    help                Show top-level help message.
    populate            Populate an SQLite database.
    process             Run a processing step on the specified database.
    query               Run a query directly on a data source.
    list-processes      List available data processes.
    list-complete-schema
                        List all data source and process schemas.
    list-source-schema  List all data source schemas (default) or the
                        specified one.
    list-process-schema
                        List the schema of all processes (default) or of the
                        specified one.
    list-sources        List available data sources
    version             Report program version

optional arguments:
  -h, --help            show this help message and exit
  -d DEBUG, --debug DEBUG
```

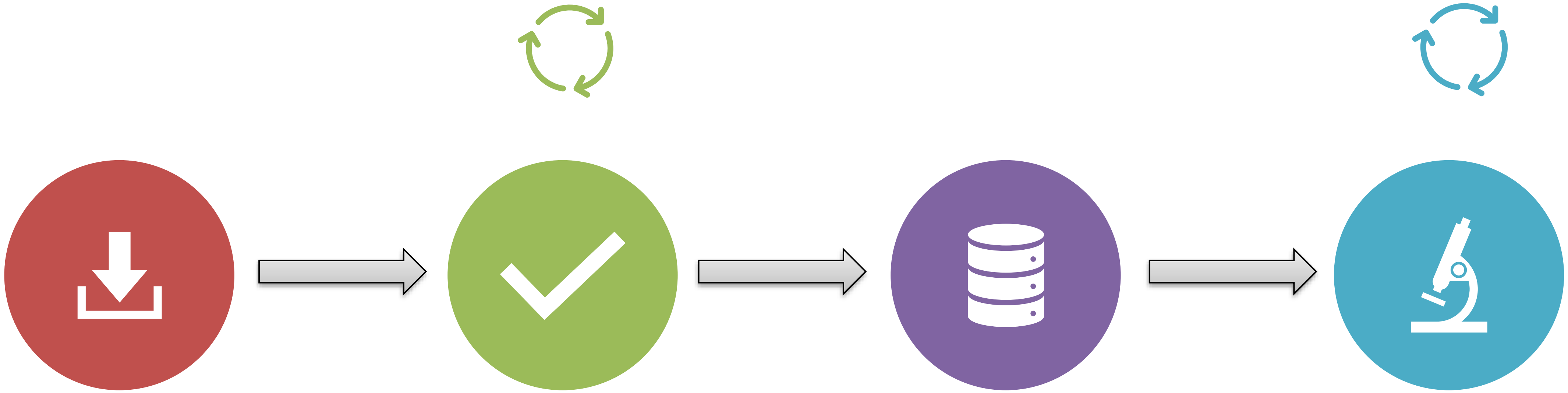
CLI invocation example

```
a3k populate covid.db \
   crossref 'April 2022 Public Data File from Crossref' \
   --row-selection "title like '%COVID%' OR abstract like '%COVID%' "
```

Python module example

```python
from alexandria3k.crossref import Crossref

crossref_instance = Crossref('April 2022 Public Data File from Crossref’)

crossref_instance.populate(
    "covid.db", condition="title like '%COVID%' OR abstract like '%COVID%'"
)
```

Typical workflow



Download data

< 3h
156 GiB

Run EDA queries directly on sample

2’ on 1%
8 records / s

Populate database

4–20 h
4–190 GiB

Develop, test, refine analysis queries

1’–7h
≤ 5GiB

Main use cases

- Run ad hoc SQL queries
- Populate SQLite databases
  - Select elements horizontally
    - SQL expression
    - Sampling
  - Select elements vertically
    - Table.Column
  - Building takes minutes, hours, or a couple of days
  - Then, SQLite database queries often run in seconds

Crossref publications by year

```
a3k query crossref 'April 2022 Public Data File from Crossref' \
   --query 'SELECT  published_year AS year, Count(*) AS number
            FROM    works
            GROUP by published_year' >results.csv
```

Crossref sampling

```
time alexandria3k query crossref 'April 2022 Public Data File from Crossref' \
  --query 'SELECT works.abstract is not null AS have_abstract, Count(*)
            FROM works GROUP BY have_abstract
          ' \
  --sample 'random.random() < 0.01 '
```

```
0   1218383
1    156617

real    2m6.488s
user    1m58.878s
sys     0m6.920s
```

Crossref population metrics

```
alexandria3k populate crossref 'April 2022 Public Data File from Crossref' graph.db \
   --columns  works.doi work_references.work_id work_references.doi work_funders.id \
    work_funders.work_id work_funders.doi funder_awards.funder_id funder_awards.name \
    author_affiliations.author_id author_affiliations.name work_subjects.work_id work_subjects.name \
    work_authors.id work_authors.work_id work_authors.orcid
```

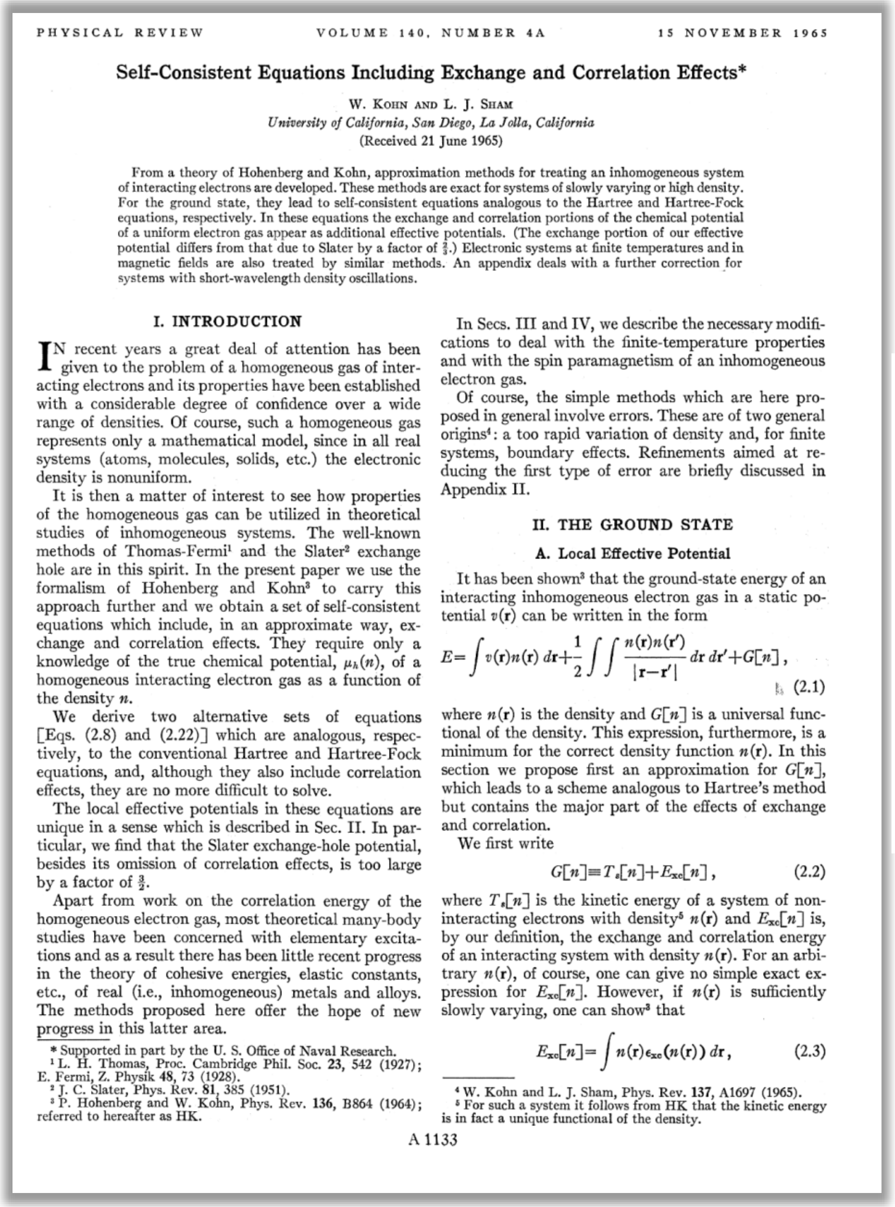
```sql
SELECT COUNT(*) FROM works;
SELECT COUNT(*) FROM (SELECT DISTINCT work_id FROM works_subjects);
SELECT COUNT(*) FROM (SELECT DISTINCT work_id FROM work_references);
SELECT COUNT(*) FROM affiliations_works;
SELECT COUNT(*) FROM (SELECT DISTINCT work_id FROM work_funders);

SELECT COUNT(*) FROM work_authors;
SELECT COUNT(*) FROM work_authors WHERE orcid is not null;
SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM work_authors);

SELECT COUNT(*) FROM authors_affiliations;
SELECT COUNT(*) FROM affiliation_names;

SELECT COUNT(*) FROM works_subjects;
SELECT COUNT(*) FROM subject_names;

SELECT COUNT(*) FROM work_funders;
SELECT COUNT(*) FROM funder_awards;

SELECT COUNT(*) FROM work_references;
```

Number of ORCID elements (for chart)

```
alexandria3k populate ORCID_2022_10_summaries.db \
  orcid ORCID_2022_10_summaries.tar.gz
```

```
SELECT "persons" AS type, (SELECT COUNT(*) FROM persons) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM persons)) AS persons UNION
SELECT "researcher_urls" AS type, (SELECT COUNT(*) FROM researcher_urls) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM researcher_urls)) AS persons UNION
SELECT "person_countries" AS type, (SELECT COUNT(*) FROM person_countries) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM person_countries)) AS persons UNION
SELECT "person_keywords" AS type, (SELECT COUNT(*) FROM person_keywords) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM person_keywords)) AS persons UNION
SELECT "person_external_identifiers" AS type, (SELECT COUNT(*) FROM person_external_identifiers) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM person_external_identifiers)) AS persons UNION
SELECT "distinctions" AS type, (SELECT COUNT(*) FROM distinctions) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM distinctions)) AS persons UNION
SELECT "educations" AS type, (SELECT COUNT(*) FROM educations) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM educations)) AS persons UNION
SELECT "employments" AS type, (SELECT COUNT(*) FROM employments) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM employments)) AS persons UNION
SELECT "invited_positions" AS type, (SELECT COUNT(*) FROM invited_positions) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM invited_positions)) AS persons UNION
SELECT "memberships" AS type, (SELECT COUNT(*) FROM memberships) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM memberships)) AS persons UNION
SELECT "qualifications" AS type, (SELECT COUNT(*) FROM qualifications) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM qualifications)) AS persons UNION
SELECT "services" AS type, (SELECT COUNT(*) FROM services) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM services)) AS persons UNION
SELECT "fundings" AS type, (SELECT COUNT(*) FROM fundings) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM fundings)) AS persons UNION
SELECT "peer_reviews" AS type, (SELECT COUNT(*) FROM peer_reviews) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM peer_reviews)) AS persons UNION
SELECT "research_resources" AS type, (SELECT COUNT(*) FROM research_resources) AS records,
  (SELECT COUNT(*) FROM (SELECT DISTINCT orcid FROM research_resources)) AS persons;
```

Consolidation / Disruption index



PHYSICAL REVIEW VOLUME 140, NUMBER 4A 15 NOVEMBER 1965

Self-Consistent Equations Including Exchange and Correlation Effects*

W. Kohn and L. J. Sham
University of California, San Diego, La Jolla, California

(Received 21 June 1965)

From a theory of Hohenberg and Kohn, approximation methods for treating an inhomogeneous system of interacting electrons are developed. These methods are exact for systems of slowly varying or high density. For the ground state, they lead to self-consistent equations analogous to the Hartree and Hartree-Fock equations, respectively. In these equations the exchange and correlation portions of the chemical potential of a uniform electron gas appear as additional effective potentials. (The exchange portion of our effective potential differs from that due to Slater by a factor of $\frac{2}{3}$.) Electronic systems at finite temperatures and in magnetic fields are also treated by similar methods. An appendix deals with a further correction for systems with short-wavelength density oscillations.

I. INTRODUCTION

In recent years a great deal of attention has been given to the problem of a homogeneous gas of interacting electrons and its properties have been established with a considerable degree of confidence over a wide range of densities. Of course, such a homogeneous gas represents only a mathematical model, since in all real systems (atoms, molecules, solids, etc.) the electronic density is nonuniform.

It is then a matter of interest to see how properties of the homogeneous gas can be utilized in theoretical studies of inhomogeneous systems. The well-known methods of Thomas-Fermi[1] and the Slater[2] exchange hole are in this spirit. In the present paper we use the formalism of Hohenberg and Kohn[3] to carry this approach further and we obtain a set of self-consistent equations which include, in an approximate way, exchange and correlation effects. They require only a knowledge of the true chemical potential, $\mu_h(n)$, of a homogeneous interacting electron gas as a function of the density n.

We derive two alternative sets of equations [Eqs. (2.8) and (2.22)] which are analogous, respectively, to the conventional Hartree and Hartree-Fock equations, and, although they also include correlation effects, they are no more difficult to solve.

The local effective potentials in these equations are unique in a sense which is described in Sec. II. In particular, we find that the Slater exchange-hole potential, besides its omission of correlation effects, is too large by a factor of $\frac{3}{2}$.

Apart from work on the correlation energy of the homogeneous electron gas, most theoretical many-body studies have been concerned with elementary excitations and as a result there has been little recent progress in the theory of cohesive energies, elastic constants, etc., of real (i.e., inhomogeneous) metals and alloys. The methods proposed here offer the hope of new progress in this latter area.

In Secs. III and IV, we describe the necessary modifications to deal with the finite-temperature properties and with the spin paramagnetism of an inhomogeneous electron gas.

Of course, the simple methods which are here proposed in general involve errors. These are of two general origins[4]: a too rapid variation of density and, for finite systems, boundary effects. Refinements aimed at reducing the first type of error are briefly discussed in Appendix II.

II. THE GROUND STATE

A. Local Effective Potential

It has been shown[3] that the ground-state energy of an interacting inhomogeneous electron gas in a static potential $v(\mathbf{r})$ can be written in the form

$$E=\int v(\mathbf{r})n(\mathbf{r})\,d\mathbf{r}+\frac{1}{2}\int\int\frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}'+G[n]\,, \tag{2.1}$$

where $n(\mathbf{r})$ is the density and $G[n]$ is a universal functional of the density. This expression, furthermore, is a minimum for the correct density function $n(\mathbf{r})$. In this section we propose first an approximation for $G[n]$, which leads to a scheme analogous to Hartree's method but contains the major part of the effects of exchange and correlation.

We first write

$$G[n]\equiv T_s[n]+E_{xc}[n]\,, \tag{2.2}$$

where $T_s[n]$ is the kinetic energy of a system of noninteracting electrons with density[5] $n(\mathbf{r})$ and $E_{xc}[n]$ is, by our definition, the exchange and correlation energy of an interacting system with density $n(\mathbf{r})$. For an arbitrary $n(\mathbf{r})$, of course, one can give no simple exact expression for $E_{xc}[n]$. However, if $n(\mathbf{r})$ is sufficiently slowly varying, one can show[3] that

$$E_{xc}[n]=\int n(\mathbf{r})\epsilon_{xc}(n(\mathbf{r}))\,d\mathbf{r}\,, \tag{2.3}$$

* Supported in part by the U. S. Office of Naval Research.

[1] L. H. Thomas, Proc. Cambridge Phil. Soc. **23**, 542 (1927); E. Fermi, Z. Physik **48**, 73 (1928).

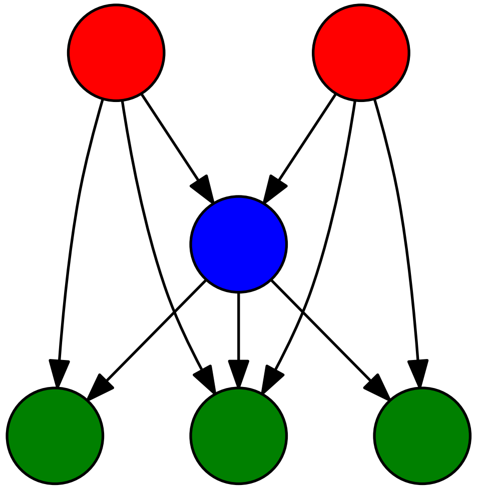
[2] J. C. Slater, Phys. Rev. **81**, 385 (1951).

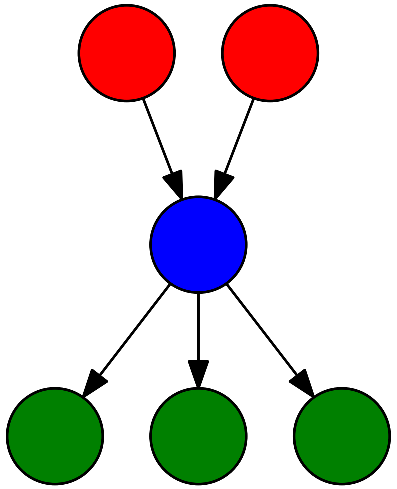
[3] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); referred to hereafter as HK.

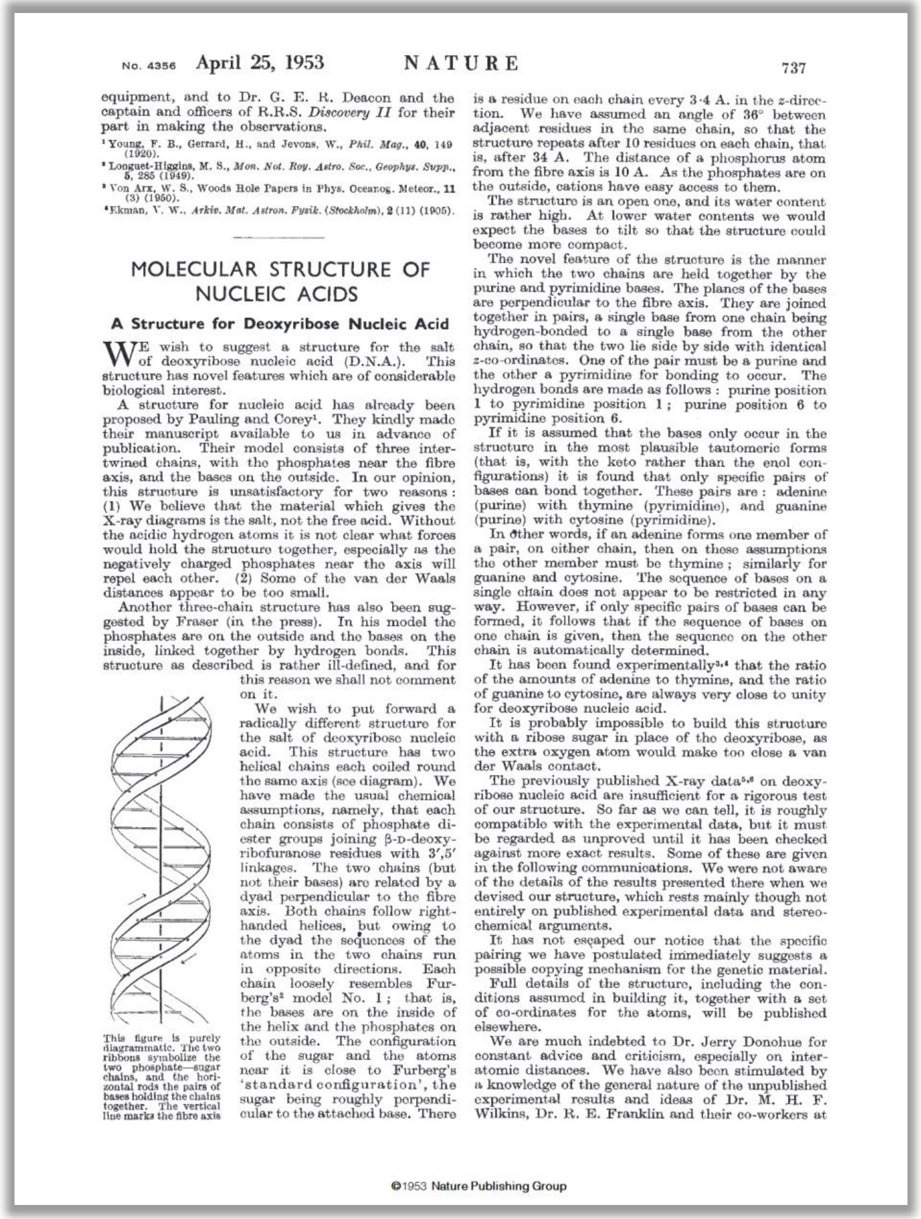
[4] W. Kohn and L. J. Sham, Phys. Rev. **137**, A1697 (1965).

[5] For such a system it follows from HK that the kinetic energy is in fact a unique functional of the density.

CD$_5$	Method	CD$_5$
−0.22	Nature	0.62
−0.25	Alexandria3k	0.57







No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).

[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).

[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).

[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

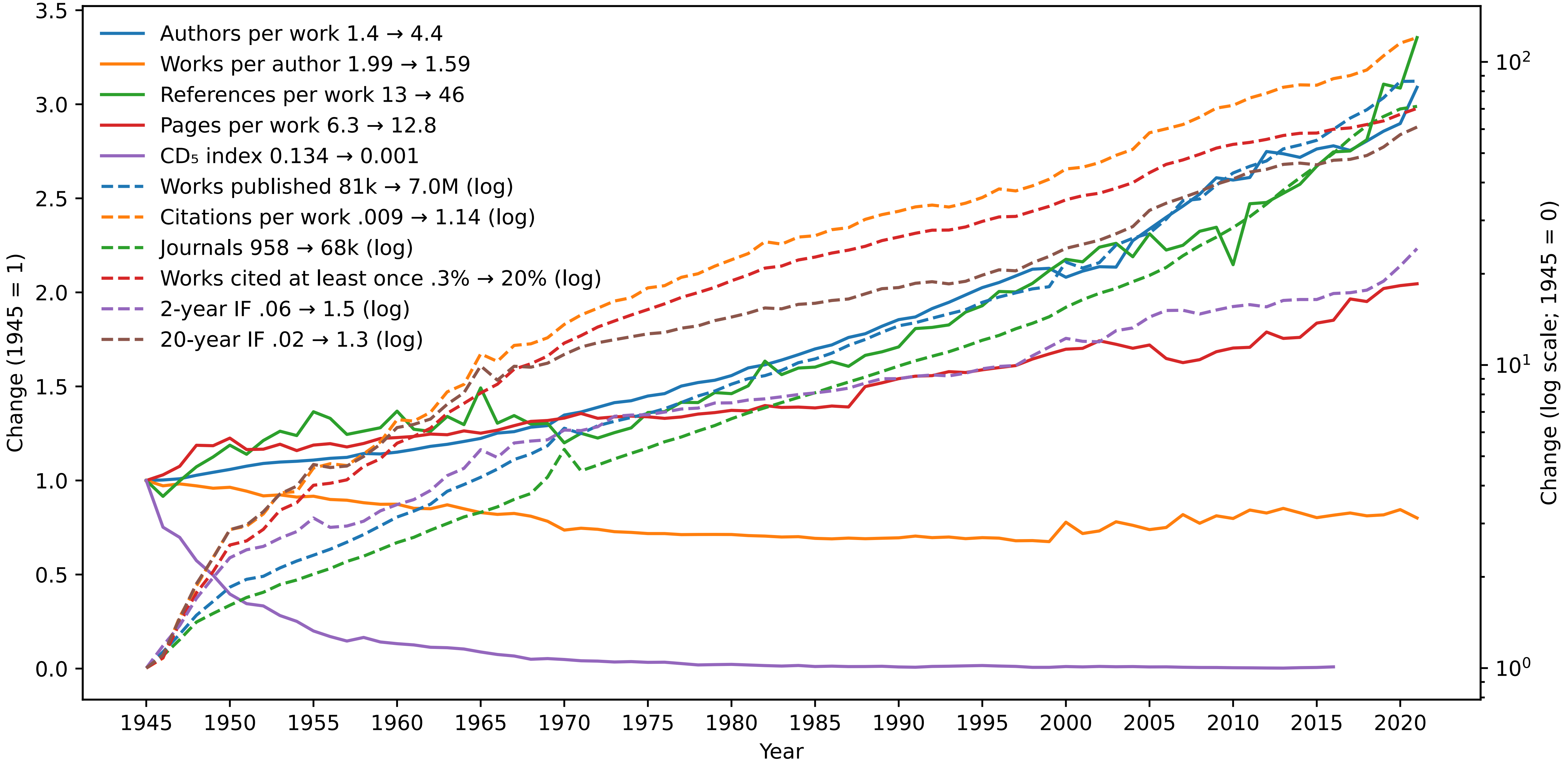
It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

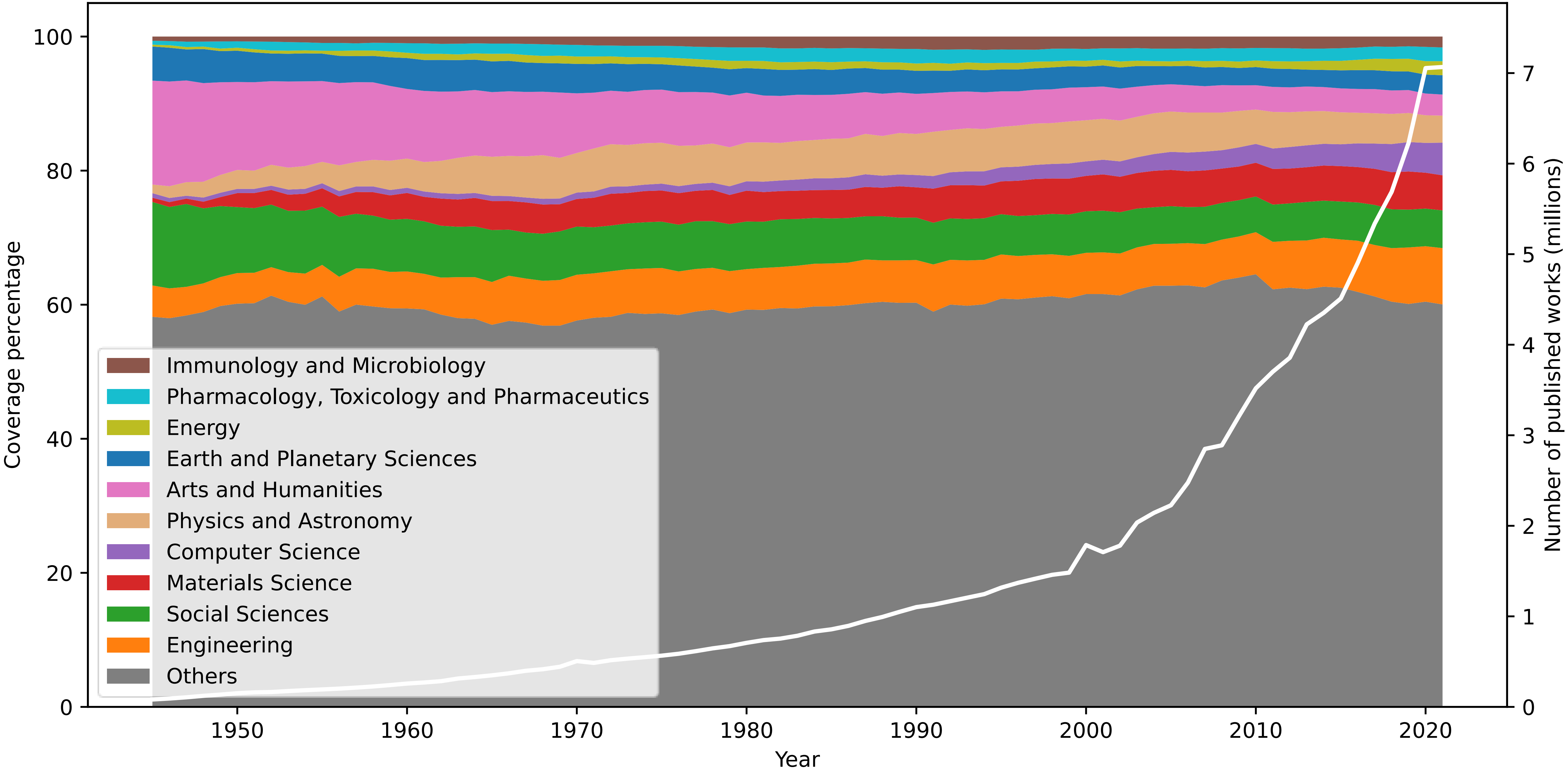
We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

©1953 Nature Publishing Group

Evolution of scientific publishing



Evolution in publications per field



```sql
-- Applicants Population by Country and year for the Top 5 Countries of 2022
WITH ranked_countries AS (
    SELECT
      SUBSTRING(date_published, 1, 4) AS year,
      usp_applicants.country AS country,
      COUNT(*) AS patent_count,
      ROW_NUMBER() OVER(PARTITION BY SUBSTRING(date_published, 1, 4) ORDER BY COUNT(*) DESC) AS country_rank
    FROM us_patents
    INNER JOIN usp_applicants
    ON us_patents.container_id = usp_applicants.patent_id
    GROUP BY
      year, usp_applicants.country
 ),
 top_5_2022 AS (
    SELECT country
    FROM ranked_countries
    WHERE
      year = '2022' AND country_rank <= 5
 )
SELECT
   rc.year,
   rc.country,
   rc.patent_count
 FROM ranked_countries rc
 JOIN top_5_2022 t5
 ON
   rc.country = t5.country
 ORDER BY
   rc.year, rc.country;
```

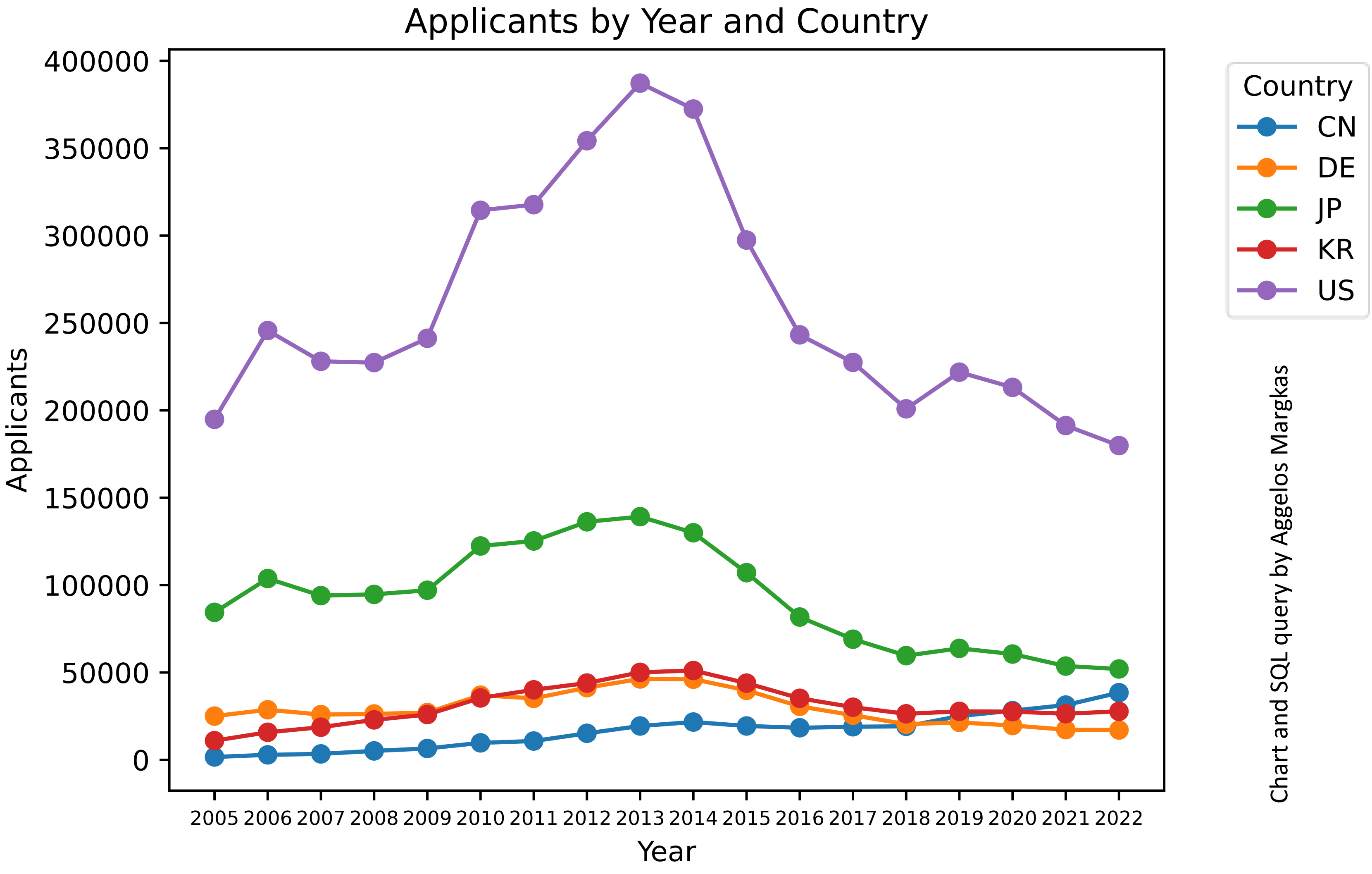
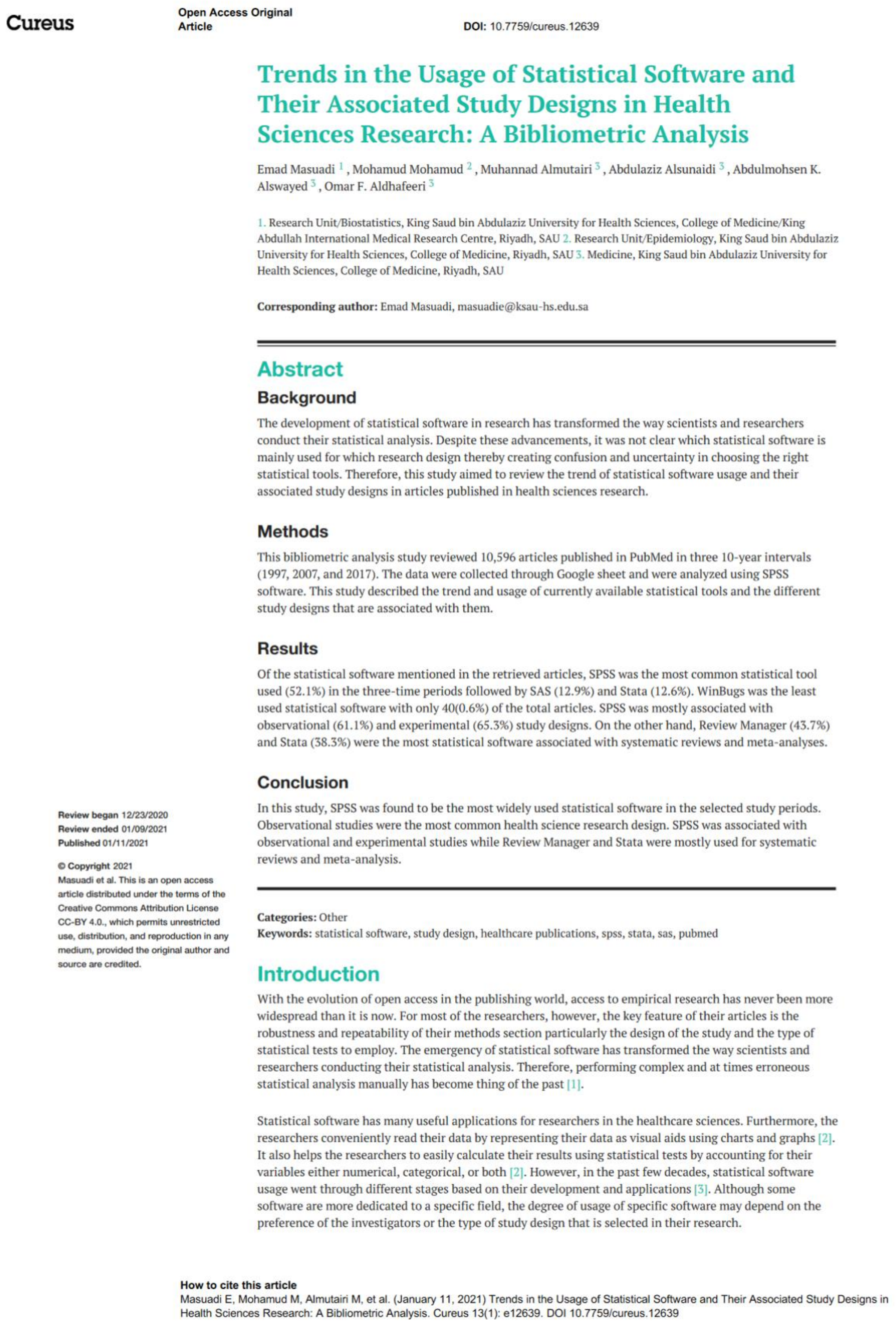


Chart and SQL query by Aggelos Margkas



Cureus

Open Access Original Article

DOI: 10.7759/cureus.12639

Trends in the Usage of Statistical Software and Their Associated Study Designs in Health Sciences Research: A Bibliometric Analysis

Emad Masuadi [1] , Mohamud Mohamud [2] , Muhannad Almutairi [3] , Abdulaziz Alsunaidi [3] , Abdulmohsen K. Alswayed [3] , Omar F. Aldhafeeri [3]

1. Research Unit/Biostatistics, King Saud bin Abdulaziz University for Health Sciences, College of Medicine/King Abdullah International Medical Research Centre, Riyadh, SAU 2. Research Unit/Epidemiology, King Saud bin Abdulaziz University for Health Sciences, College of Medicine, Riyadh, SAU 3. Medicine, King Saud bin Abdulaziz University for Health Sciences, College of Medicine, Riyadh, SAU

Corresponding author: Emad Masuadi, masuadie@ksau-hs.edu.sa

Abstract

Background

The development of statistical software in research has transformed the way scientists and researchers conduct their statistical analysis. Despite these advancements, it was not clear which statistical software is mainly used for which research design thereby creating confusion and uncertainty in choosing the right statistical tools. Therefore, this study aimed to review the trend of statistical software usage and their associated study designs in articles published in health sciences research.

Methods

This bibliometric analysis study reviewed 10,596 articles published in PubMed in three 10-year intervals (1997, 2007, and 2017). The data were collected through Google sheet and were analyzed using SPSS software. This study described the trend and usage of currently available statistical tools and the different study designs that are associated with them.

Results

Of the statistical software mentioned in the retrieved articles, SPSS was the most common statistical tool used (52.1%) in the three-time periods followed by SAS (12.9%) and Stata (12.6%). WinBugs was the least used statistical software with only 40(0.6%) of the total articles. SPSS was mostly associated with observational (61.1%) and experimental (65.3%) study designs. On the other hand, Review Manager (43.7%) and Stata (38.3%) were the most statistical software associated with systematic reviews and meta-analyses.

Conclusion

In this study, SPSS was found to be the most widely used statistical software in the selected study periods. Observational studies were the most common health science research design. SPSS was associated with observational and experimental studies while Review Manager and Stata were mostly used for systematic reviews and meta-analysis.

Review began 12/23/2020
Review ended 01/09/2021
Published 01/11/2021

© Copyright 2021
Masuadi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Categories: Other
Keywords: statistical software, study design, healthcare publications, spss, stata, sas, pubmed

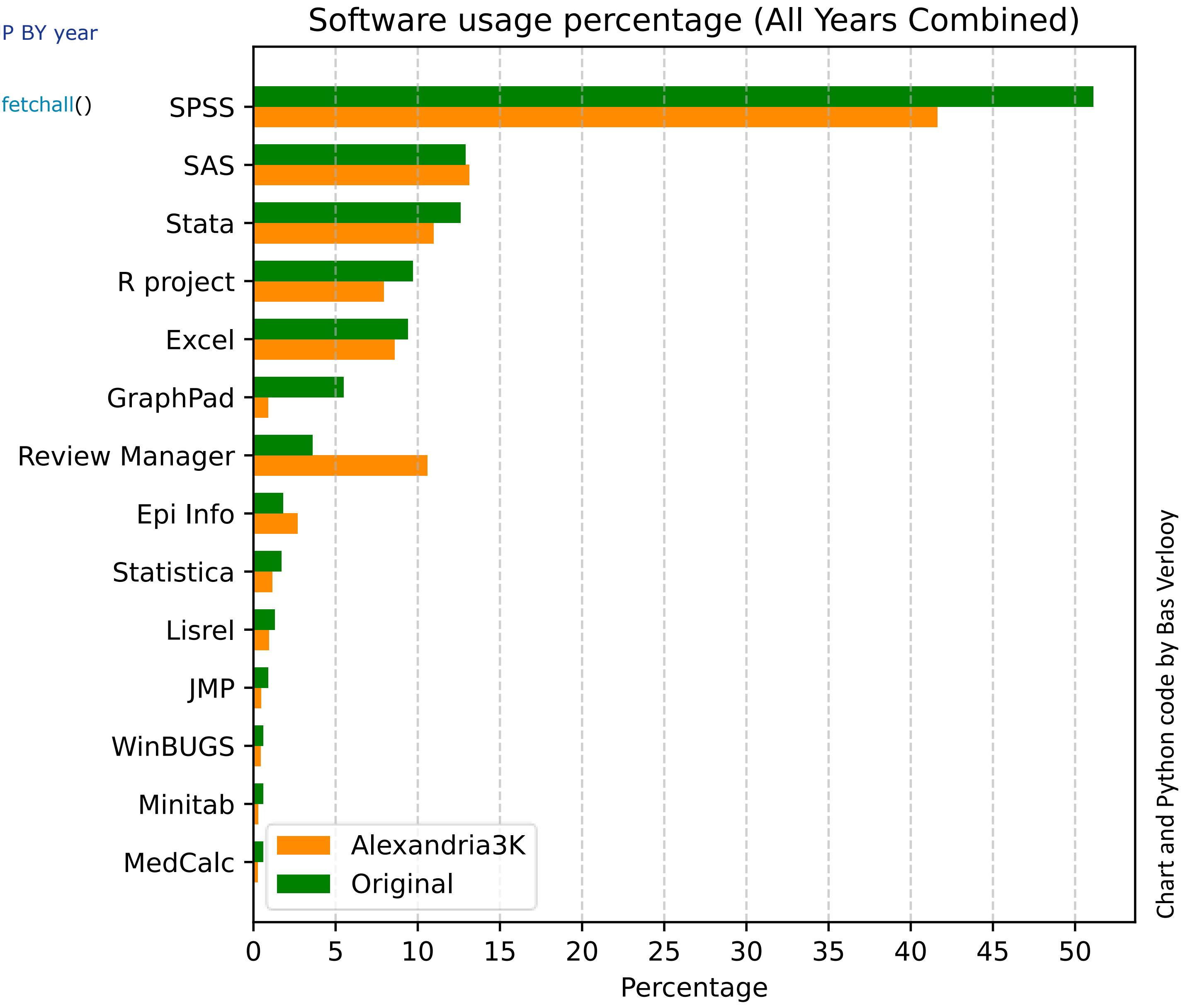
Introduction

With the evolution of open access in the publishing world, access to empirical research has never been more widespread than it is now. For most of the researchers, however, the key feature of their articles is the robustness and repeatability of their methods section particularly the design of the study and the type of statistical tests to employ. The emergency of statistical software has transformed the way scientists and researchers conducting their statistical analysis. Therefore, performing complex and at times erroneous statistical analysis manually has become thing of the past [1].

Statistical software has many useful applications for researchers in the healthcare sciences. Furthermore, the researchers conveniently read their data by representing their data as visual aids using charts and graphs [2]. It also helps the researchers to easily calculate their results using statistical tests by accounting for their variables either numerical, categorical, or both [2]. However, in the past few decades, statistical software usage went through different stages based on their development and applications [3]. Although some software are more dedicated to a specific field, the degree of usage of specific software may depend on the preference of the investigators or the type of study design that is selected in their research.

How to cite this article
Masuadi E, Mohamud M, Almutairi M, et al. (January 11, 2021) Trends in the Usage of Statistical Software and Their Associated Study Designs in Health Sciences Research: A Bibliometric Analysis. Cureus 13(1): e12639. DOI 10.7759/cureus.12639

```
def query_software(software):
    software_search = " OR ".join([f'"{s}"' for s in software])

    c.execute(
        f"""
        SELECT year, COUNT(DISTINCT(article_id)) FROM (
            SELECT article_id, year FROM fts_abstracts
            WHERE text MATCH '{software_search}' or title MATCH '{software_search}'
            GROUP BY article_id, year
        )
        GROUP BY year
        """
    )
    return c.fetchall()
```



Software usage percentage (All Years Combined)

Chart and Python code by Bas Verlooy

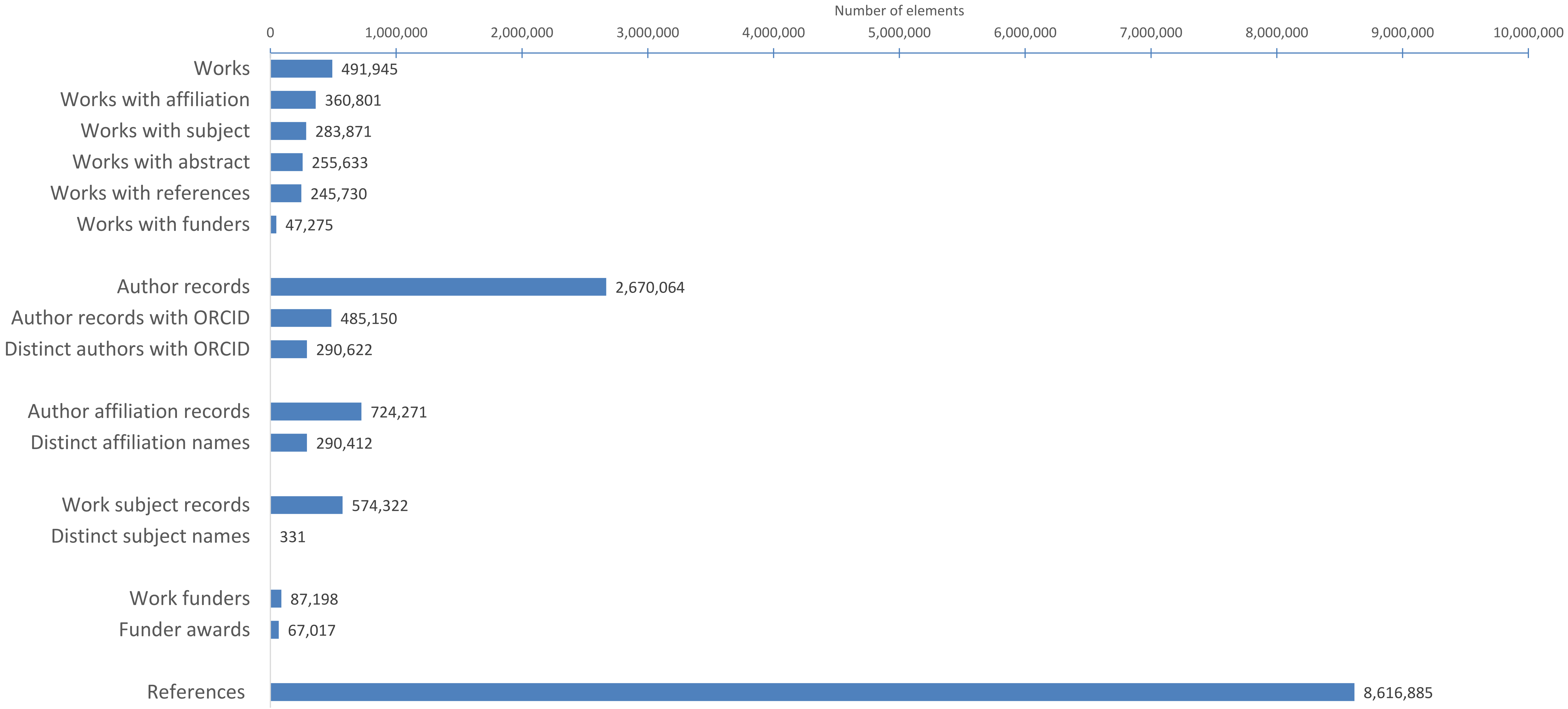
Covid Care in Color. (2021). In AMA Journal of Ethics (Vol. 23, Issue 4, pp. E349-361). DOI: 10.1001/amajethics.2021.349

A data set of COVID research

```
alexandria3k populate covid.db \
  crossref 'April 2022 Public Data File from Crossref’ \
  --row-selection "title like '%COVID%' OR abstract like '%COVID%' "
```

- 9:06:23 elapsed time
- 2.9 GB data, 3.6 GB fully indexed

COVID data set in numbers



COVID research topics

```
SELECT rank() OVER (ORDER BY count(*) DESC), count(*), name
  FROM work_subjects GROUP BY name;
```

Rank	Publications	Subject
1	70609	General Medicine
2	23070	Public Health, Environmental and Occupational Health
3	17254	Infectious Diseases
4	10404	Psychiatry and Mental health
5	9590	Education
18	6013	Computer Science Applications
20	5942	General Engineering
21	5940	Pulmonary and Respiratory Medicine
23	5908	Geography, Planning and Development
27	4991	Sociology and Political Science
28	4553	Critical Care and Intensive Care Medicine
32	4182	Epidemiology
36	4067	Virology
37	3898	Management, Monitoring, Policy and Law
40	3601	Economics and Econometrics
42	3208	Strategy and Management
58	2557	Law
62	2329	History
63	2251	Business and International Management
64	2196	Electrical and Electronic Engineering
76	1893	Cultural Studies
81	1734	Computer Networks and Communications
97	1549	Pollution
99	1519	Public Administration
111	1360	Tourism, Leisure and Hospitality Management
113	1339	General Business, Management and Accounting
119	1238	Industrial and Manufacturing Engineering
130	1032	Anthropology
131	996	Ecology, Evolution, Behavior and Systematics
140	912	Artificial Intelligence
141	909	Mechanical Engineering
142	899	Waste Management and Disposal
166	695	Ocean Engineering
169	657	Human-Computer Interaction
170	640	General Arts and Humanities
331	5	Podiatry

COVID research funding

Rank	Publications	Funding body
1	3506	National Natural Science Foundation of China
2	2316	National Institutes of Health
3	1022	National Science Foundation
4	914	Wellcome Trust
5	661	National Institute for Health Research
6	615	Medical Research Council
7	588	National Institute of Allergy and Infectious Diseases
8	541	Canadian Institutes of Health Research
9	520	Deutsche Forschungsgemeinschaft
10	503	Conselho Nacional de Desenvolvimento Científico e Tecnológico
11	495	Bill and Melinda Gates Foundation
12	483	National Research Foundation of Korea
13	481	Japan Society for the Promotion of Science
14	439	National Heart, Lung, and Blood Institute
15	430	National Key Research and Development Program of China
16	422	National Center for Advancing Translational Sciences
17	417	Instituto de Salud Carlos III
18	394	National Institute on Aging
19	382	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
20	365	National Cancer Institute

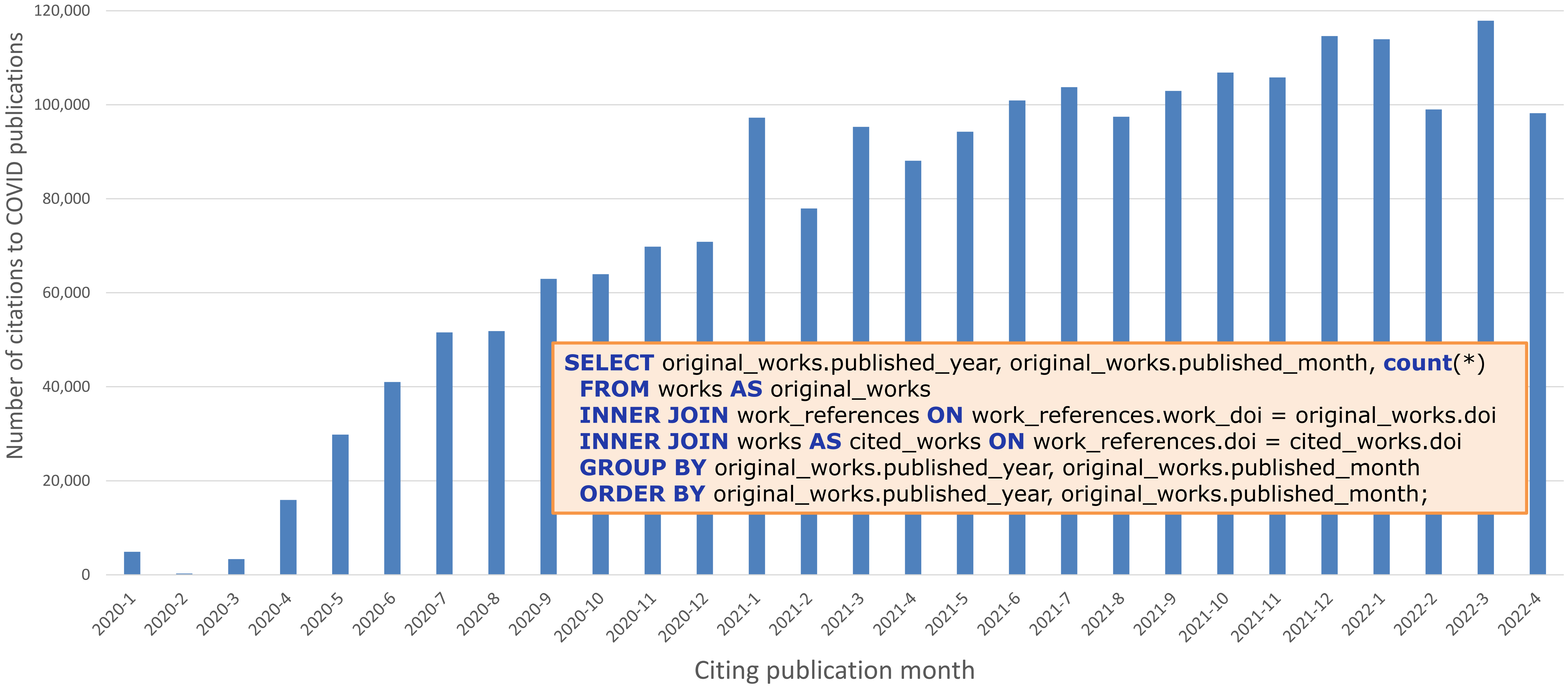
```sql
SELECT rank() OVER (
   ORDER BY count(*) DESC), count(*), name
 FROM work_funders GROUP BY name LIMIT 20;
```

Affiliations of COVID publications

Rank	Works	Affiliation (top parent)
1	1465	Government of the United States of America
2	925	University of California System
3	910	University of Toronto
4	824	University of London
5	660	University of Oxford
6	654	Istituti di Ricovero e Cura a Carattere Scientifico
7	632	Mount Sinai Health System
8	592	Tehran University of Medical Sciences
9	587	University of North Carolina System
10	501	University of Melbourne
11	437	The University of Texas System
12	434	National University of Singapore
13	428	University of Cambridge
14	425	French National Centre for Scientific Research
15	400	Yale University
16	371	UNSW Sydney
17	369	Government of India
17	369	Shahid Beheshti University of Medical Sciences
19	366	Raymond and Ruth Perelman School of Medicine at the University of Pennsylvania
20	361	Cornell University

```sql
-- Match works with identified authors' affiliations
WITH work_rors AS (
  -- Works and participating RORs
  SELECT DISTINCT work_id, ror_id
  FROM work_authors_rors
  LEFT JOIN work_authors
    ON work_authors_rors.work_author_id = work_authors.id
),

-- Count works by research organization (ROR)
ror_work_counts AS (
  SELECT ror_id, Count(*) AS number FROM work_rors GROUP BY ror_id
),

-- Add ROR names
ror_name_work_counts AS (
  SELECT name, number from ror_work_counts
  INNER JOIN research_organizations
    ON ror_work_counts.ror_id = research_organizations.id
),

-- Match works with unidentified author affiliations
unmatched_work_affiliations AS (
  SELECT DISTINCT work_id, author_affiliations.name FROM
    work_authors
    INNER JOIN author_affiliations
      ON work_authors.id = author_affiliations.author_id
    LEFT JOIN work_authors_rors
      ON work_authors_rors.work_author_id = work_authors.id
    WHERE work_authors_rors.ror_id is null
),

-- Count works by unidentified author affiliations
unmatched_affiliation_work_counts AS (
  SELECT name, Count(*) AS number FROM unmatched_work_affiliations
  GROUP BY name
),

-- Merge the two groups together
all_work_counts AS (
  SELECT * FROM ror_name_work_counts
  UNION
  SELECT * FROM unmatched_affiliation_work_counts
)

-- Output the top-20 affiliations according to number of published works
SELECT Rank() OVER (ORDER BY number DESC) AS rank, number, name
FROM all_work_counts
LIMIT 20;
```

Building on COVID knowledge



```sql
SELECT original_works.published_year, original_works.published_month, count(*)
  FROM works AS original_works
  INNER JOIN work_references ON work_references.work_doi = original_works.doi
  INNER JOIN works AS cited_works ON work_references.doi = cited_works.doi
  GROUP BY original_works.published_year, original_works.published_month
  ORDER BY original_works.published_year, original_works.published_month;
```

Extreme collaboration under COVID

Rank	Author records	Affiliation
1	2352	Writing Committee for the REMAP-CAP Investigators
2	1731	REMAP-CAP Writing Committee for the REMAP-CAP Investigators
3	734	for the Society of Critical Care Medicine Discovery Viral Infection and Respiratory Illness Universal Study (VIRUS): COVID-19 Registry Investigator Group
4	729	for the COVID-19 Phase 3 Prevention Trial Team
5	604	for the COVID-19 and Cancer Consortium
6	587	for the CORIMUNO-19 Collaborative Group
7	555	for the COVID-19 and Cancer Consortium (CCC19)
8	536	Shiraz University of Medical Sciences
9	412	for the PREP-IT Investigators
10	375	University of Oxford
11	369	for the RECOVERY-RS Collaborators
12	364	Universidade de São Paulo, Brazil
13	351	National Institute for Infectious Diseases “L. Spallanzani” IRCCS, Rome, Italy
14	336	ФКУЗ Российский научно-исследовательский противочумный институт «Микроб» Роспотребнадзора, Саратов, Российская Федерация
15	331	Tehran University of Medical Sciences
16	321	Hamad Medical Corporation
17	305	for the STOP-COVID Investigators
18	298	Fundação Oswaldo Cruz, Brazil
19	285	The WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group
20	276	for the Psoriasis Patient Registry for Outcomes, Therapy and Epidemiology of COVID-19 Infection (PsoProtect); the Secure Epidemiology of Coronavirus Under Research Exclusion for Inflammatory Bowel Disease (SECURE-IBD); and the COVID-19 Global Rheumatology Alliance (GRA)

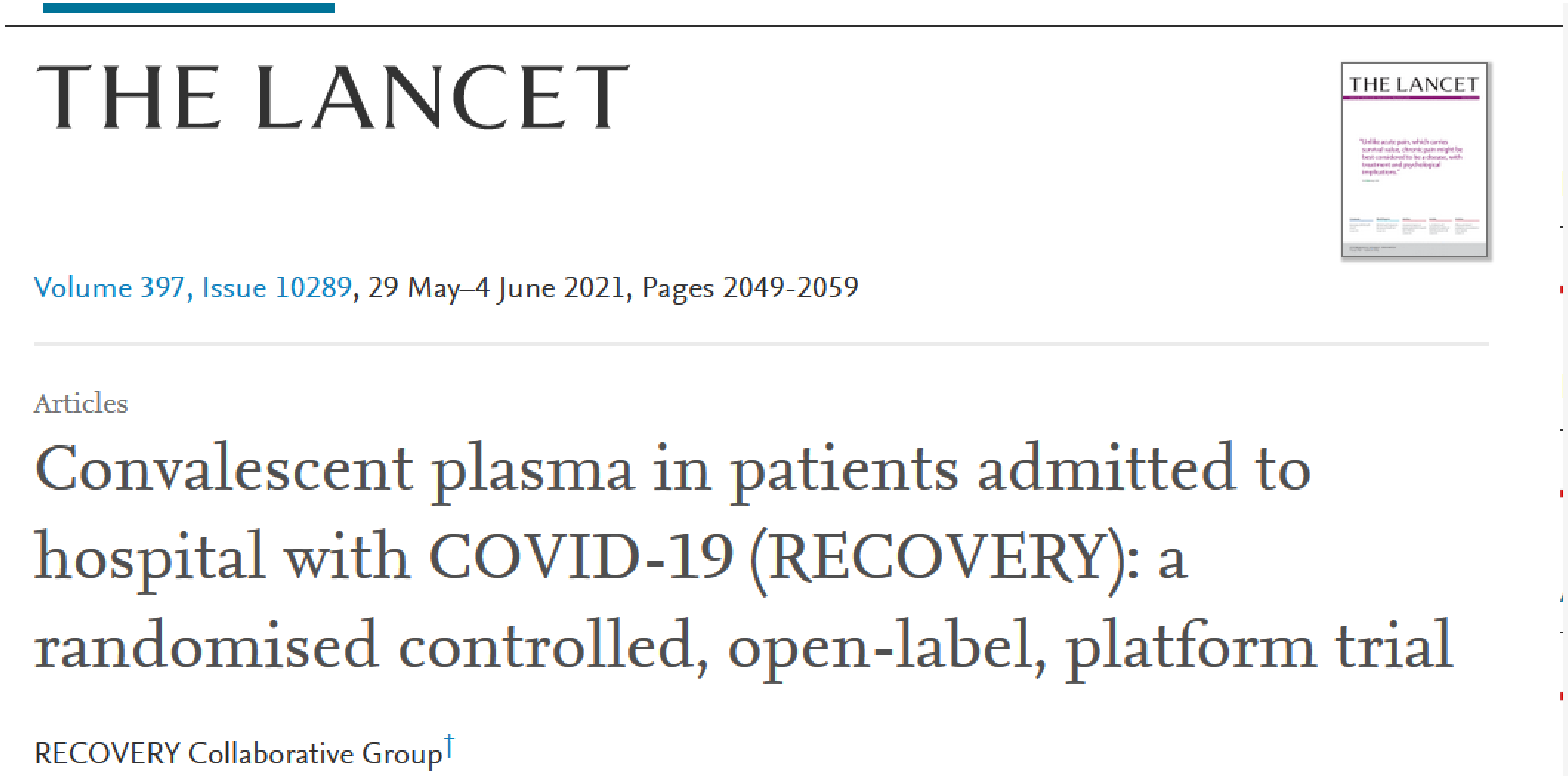
```
SELECT rank() OVER (ORDER BY count(*) DESC),
  count(*), name
  FROM author_affiliations GROUP BY name
  LIMIT 20;
```

Diving in

```sql
SELECT Avg(author_number), Max(author_number) FROM (
  SELECT Count(*) AS author_number FROM works
    LEFT JOIN work_authors ON works.doi = work_authors.work_doi
    GROUP BY works.doi
);
```

```
5.47      7194
```

The 7k author article



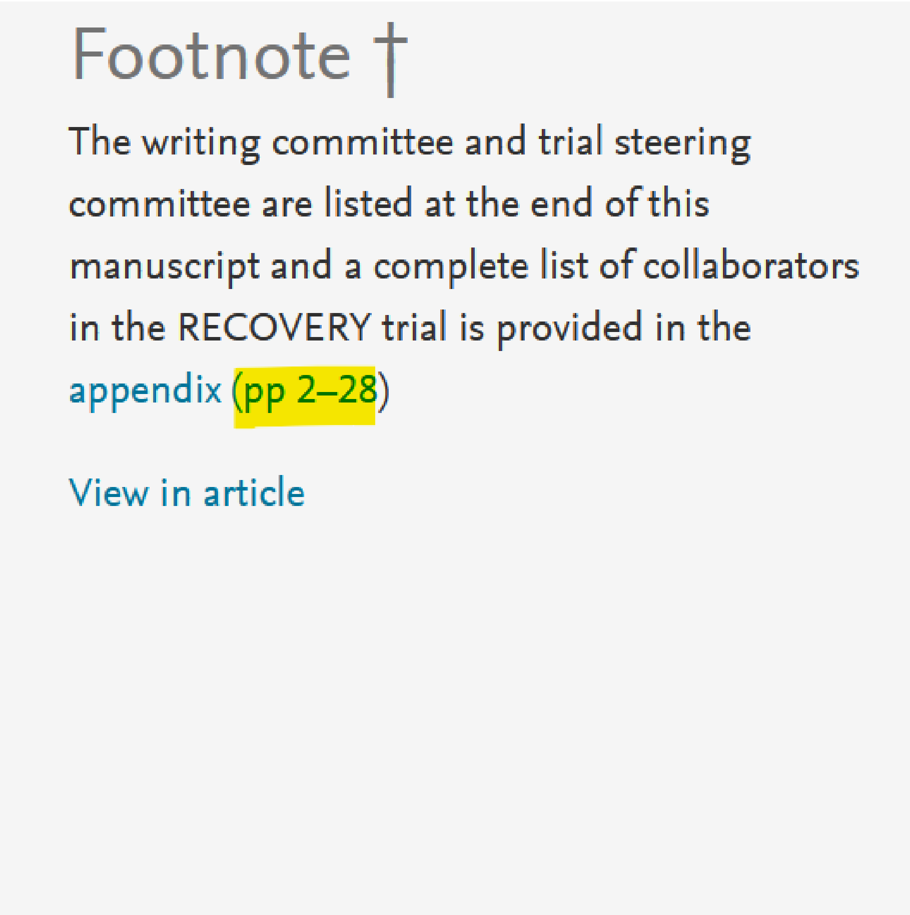
THE LANCET

Volume 397, Issue 10289, 29 May–4 June 2021, Pages 2049-2059

Articles

Convalescent plasma in patients admitted to hospital with COVID-19 (RECOVERY): a randomised controlled, open-label, platform trial

RECOVERY Collaborative Group[†]



Footnote †

The writing committee and trial steering committee are listed at the end of this manuscript and a complete list of collaborators in the RECOVERY trial is provided in the appendix (pp 2–28)

View in article

Not an isolated case

```
SELECT works.doi, Count(*) AS author_number FROM works
  LEFT JOIN work_authors
    ON works.doi = work_authors.work_doi
  GROUP BY works.doi
  ORDER BY Count(*) DESC
  LIMIT 20;
```

```
SELECT Count(*) FROM (
  SELECT Count(*) AS author_number FROM works
    LEFT JOIN work_authors
      ON works.doi = work_authors.work_doi
    GROUP BY works.doi
    HAVING author_number > 100
  );
```

457

DOI	Authors
10.1016/s0140-6736(21)00897-7	7,194
10.1016/s0140-6736(21)00676-0	6,349
10.1016/s0140-6736(22)00163-5	6,303
10.1016/s0140-6736(21)01825-0	6,215
10.1093/bjs/znab336	5,549
10.1016/s0140-6736(21)00149-5	5,370
10.1016/s1470-2045(21)00493-9	5,203
10.1093/bjs/znab183	4,870
10.1038/s41586-021-03767-x	3,903
10.1200/jco.20.01933	3,647
10.1093/bjs/znaa051	3,608
10.1001/jama.2021.18178	2,445
10.1007/s00134-021-06448-5	2,013
10.1001/jama.2022.2910	1,805
10.1007/s00439-021-02397-7	1,577
10.1016/s2352-3018(21)00151-x	1,574
10.1016/s2214-109x(21)00289-8	1,555
10.1503/cjs.021321	1,431
10.1093/bjs/znab307	1,295
10.1186/s12967-021-03094-9	1,295

The dreaded Journal Impact Factor

DALL·E 2022-11-21 13.28.42 - university academic suffering from the journal impact factor

$$\mathrm{IF}_y = \frac{\mathrm{Citations}_y}{\mathrm{Publications}_{y-1} + \mathrm{Publications}_{y-2}}.$$

Journal Impact Factor

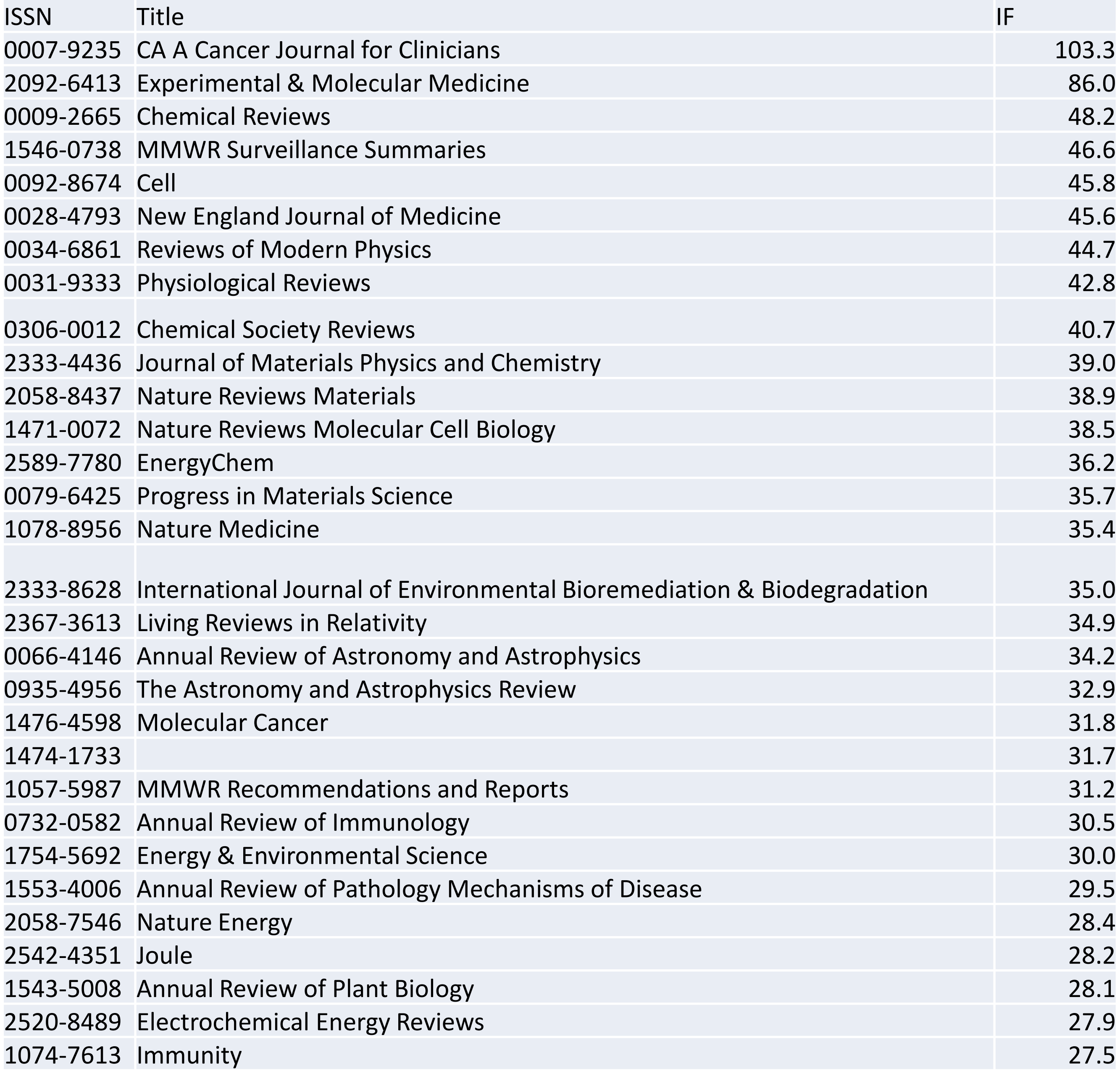
```
alexandria3k populate impact_data.db crossref 'April 2022 Public Data File from Crossref’
   --row-selection 'works.published_year BETWEEN 2019 AND 2021'
   --columns works.doi works.issn_print works.issn_electronic works.published_year \
       work_references.work_doi work_references.doi
alexandria3k populate impact_data.db journal-names
```

```
ATTACH 'impact_data.db' AS impact_data;

CREATE TABLE works_issn AS
  SELECT doi AS doi, published_year
    Coalesce(issn_print, issn_electronic) AS issn
  FROM impact_data.works
  WHERE issn is not null;

CREATE index works_issn_doi_idx ON works_issn(doi);

CREATE TABLE citations AS
  SELECT cited_work.issn, COUNT(*) AS citations_number
  FROM impact_data.work_references
  INNER JOIN works_issn AS published_work
    ON work_references.work_doi = published_work.doi
  INNER JOIN works_issn AS cited_work
    ON work_references.doi = cited_work.doi
  WHERE published_work.published_year = 2021
    AND cited_work.published_year BETWEEN 2019 AND 2020
  GROUP BY cited_work.issn;
```

```
CREATE TABLE publications AS
  SELECT issn, COUNT(*) AS publications_number FROM works_issn
  WHERE published_year BETWEEN 2019 AND 2020
  GROUP BY issn;

CREATE TABLE impact_factor AS
  SELECT publications.issn, citations_number, publications_number,
    Cast(Coalesce(citations_number, 0) AS FLOAT) / publications_number
      AS impact_factor
  FROM publications
  LEFT JOIN citations ON citations.issn = publications.issn
  WHERE publications_number > 0;
```

Results

```
SELECT issn, title, impact_factor
  FROM impact_factor
  LEFT JOIN journal_names
    ON impact_factor.issn = journal_names.issn_print
      OR impact_factor.issn = journal_names.issn_eprint
ORDER BY impact_factor DESC LIMIT 30;
```

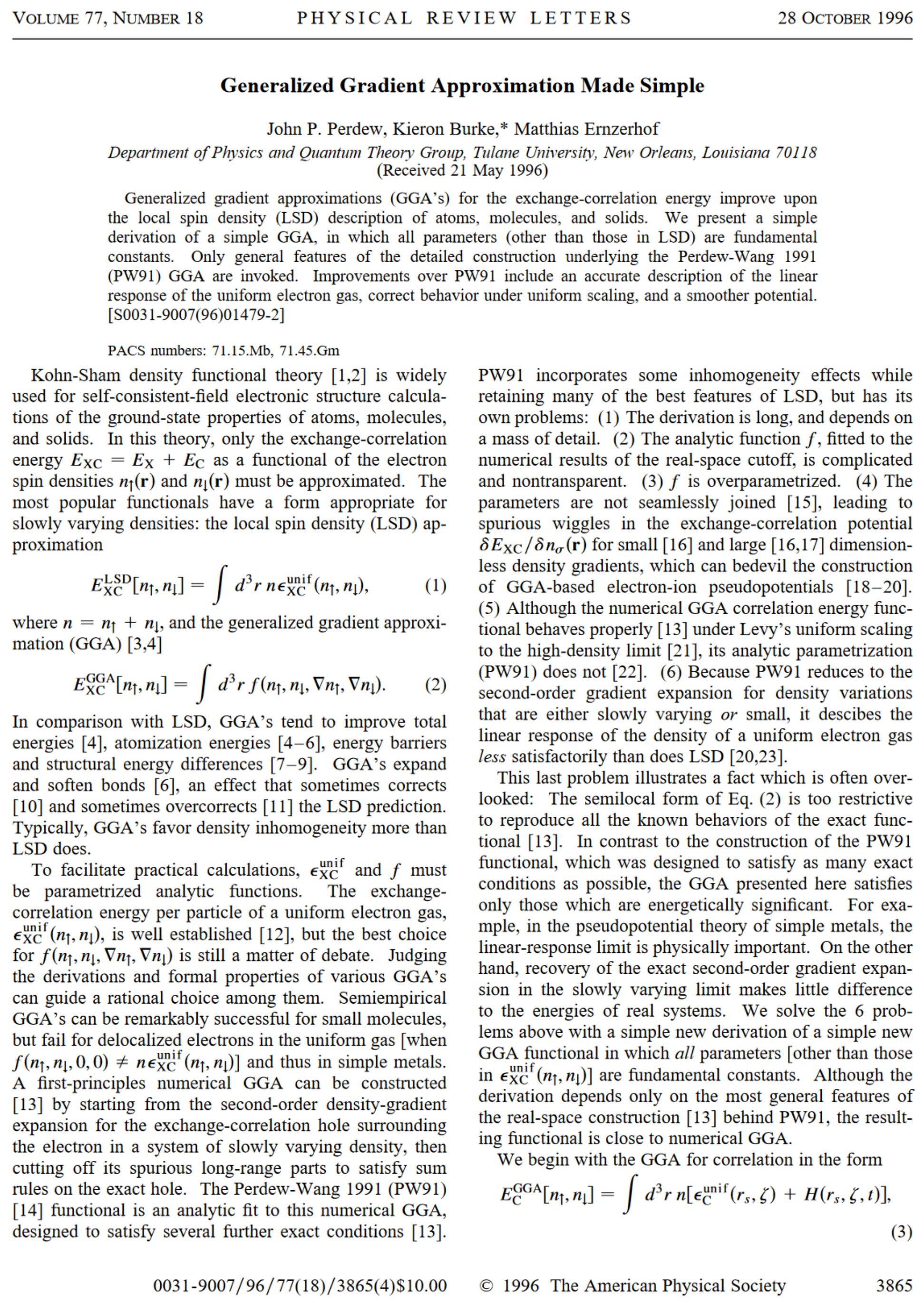


ISSN	Title	IF
0007-9235	CA A Cancer Journal for Clinicians	103.3
2092-6413	Experimental & Molecular Medicine	86.0
0009-2665	Chemical Reviews	48.2
1546-0738	MMWR Surveillance Summaries	46.6
0092-8674	Cell	45.8
0028-4793	New England Journal of Medicine	45.6
0034-6861	Reviews of Modern Physics	44.7
0031-9333	Physiological Reviews	42.8
0306-0012	Chemical Society Reviews	40.7
2333-4436	Journal of Materials Physics and Chemistry	39.0
2058-8437	Nature Reviews Materials	38.9
1471-0072	Nature Reviews Molecular Cell Biology	38.5
2589-7780	EnergyChem	36.2
0079-6425	Progress in Materials Science	35.7
1078-8956	Nature Medicine	35.4
2333-8628	International Journal of Environmental Bioremediation & Biodegradation	35.0
2367-3613	Living Reviews in Relativity	34.9
0066-4146	Annual Review of Astronomy and Astrophysics	34.2
0935-4956	The Astronomy and Astrophysics Review	32.9
1476-4598	Molecular Cancer	31.8
1474-1733		31.7
1057-5987	MMWR Recommendations and Reports	31.2
0732-0582	Annual Review of Immunology	30.5
1754-5692	Energy & Environmental Science	30.0
1553-4006	Annual Review of Pathology Mechanisms of Disease	29.5
2058-7546	Nature Energy	28.4
2542-4351	Joule	28.2
1543-5008	Annual Review of Plant Biology	28.1
2520-8489	Electrochemical Energy Reviews	27.9
1074-7613	Immunity	27.5

```
-- Most cited article in the period 2019-2021

SELECT doi, Count(*)
  FROM work_references
  GROUP BY doi
  ORDER BY count(*) DESC
  LIMIT 10;
```

31” elapsed time

39 715 citations



VOLUME 77, NUMBER 18 PHYSICAL REVIEW LETTERS 28 OCTOBER 1996

Generalized Gradient Approximation Made Simple

John P. Perdew, Kieron Burke,* Matthias Ernzerhof

Department of Physics and Quantum Theory Group, Tulane University, New Orleans, Louisiana 70118

(Received 21 May 1996)

Generalized gradient approximations (GGA’s) for the exchange-correlation energy improve upon the local spin density (LSD) description of atoms, molecules, and solids. We present a simple derivation of a simple GGA, in which all parameters (other than those in LSD) are fundamental constants. Only general features of the detailed construction underlying the Perdew-Wang 1991 (PW91) GGA are invoked. Improvements over PW91 include an accurate description of the linear response of the uniform electron gas, correct behavior under uniform scaling, and a smoother potential. [S0031-9007(96)01479-2]

PACS numbers: 71.15.Mb, 71.45.Gm

Kohn-Sham density functional theory [1,2] is widely used for self-consistent-field electronic structure calculations of the ground-state properties of atoms, molecules, and solids. In this theory, only the exchange-correlation energy $E_{\mathrm{XC}} = E_{\mathrm{X}} + E_{\mathrm{C}}$ as a functional of the electron spin densities $n_\uparrow(\mathbf{r})$ and $n_\downarrow(\mathbf{r})$ must be approximated. The most popular functionals have a form appropriate for slowly varying densities: the local spin density (LSD) approximation

$$E_{\mathrm{XC}}^{\mathrm{LSD}}[n_\uparrow, n_\downarrow] = \int d^3r\, n\epsilon_{\mathrm{XC}}^{\mathrm{unif}}(n_\uparrow, n_\downarrow), \qquad (1)$$

where $n = n_\uparrow + n_\downarrow$, and the generalized gradient approximation (GGA) [3,4]

$$E_{\mathrm{XC}}^{\mathrm{GGA}}[n_\uparrow, n_\downarrow] = \int d^3r\, f(n_\uparrow, n_\downarrow, \nabla n_\uparrow, \nabla n_\downarrow). \qquad (2)$$

In comparison with LSD, GGA’s tend to improve total energies [4], atomization energies [4–6], energy barriers and structural energy differences [7–9]. GGA’s expand and soften bonds [6], an effect that sometimes corrects [10] and sometimes overcorrects [11] the LSD prediction. Typically, GGA’s favor density inhomogeneity more than LSD does.

To facilitate practical calculations, $\epsilon_{\mathrm{XC}}^{\mathrm{unif}}$ and f must be parametrized analytic functions. The exchange-correlation energy per particle of a uniform electron gas, $\epsilon_{\mathrm{XC}}^{\mathrm{unif}}(n_\uparrow, n_\downarrow)$, is well established [12], but the best choice for $f(n_\uparrow, n_\downarrow, \nabla n_\uparrow, \nabla n_\downarrow)$ is still a matter of debate. Judging the derivations and formal properties of various GGA’s can guide a rational choice among them. Semiempirical GGA’s can be remarkably successful for small molecules, but fail for delocalized electrons in the uniform gas [when $f(n_\uparrow, n_\downarrow, 0, 0) \neq n\epsilon_{\mathrm{XC}}^{\mathrm{unif}}(n_\uparrow, n_\downarrow)$] and thus in simple metals. A first-principles numerical GGA can be constructed [13] by starting from the second-order density-gradient expansion for the exchange-correlation hole surrounding the electron in a system of slowly varying density, then cutting off its spurious long-range parts to satisfy sum rules on the exact hole. The Perdew-Wang 1991 (PW91) [14] functional is an analytic fit to this numerical GGA, designed to satisfy several further exact conditions [13]. PW91 incorporates some inhomogeneity effects while retaining many of the best features of LSD, but has its own problems: (1) The derivation is long, and depends on a mass of detail. (2) The analytic function f, fitted to the numerical results of the real-space cutoff, is complicated and nontransparent. (3) f is overparametrized. (4) The parameters are not seamlessly joined [15], leading to spurious wiggles in the exchange-correlation potential $\delta E_{\mathrm{XC}}/\delta n_\sigma(\mathbf{r})$ for small [16] and large [16,17] dimensionless density gradients, which can bedevil the construction of GGA-based electron-ion pseudopotentials [18–20]. (5) Although the numerical GGA correlation energy functional behaves properly [13] under Levy’s uniform scaling to the high-density limit [21], its analytic parametrization (PW91) does not [22]. (6) Because PW91 reduces to the second-order gradient expansion for density variations that are either slowly varying *or* small, it descibes the linear response of the density of a uniform electron gas *less* satisfactorily than does LSD [20,23].

This last problem illustrates a fact which is often overlooked: The semilocal form of Eq. (2) is too restrictive to reproduce all the known behaviors of the exact functional [13]. In contrast to the construction of the PW91 functional, which was designed to satisfy as many exact conditions as possible, the GGA presented here satisfies only those which are energetically significant. For example, in the pseudopotential theory of simple metals, the linear-response limit is physically important. On the other hand, recovery of the exact second-order gradient expansion in the slowly varying limit makes little difference to the energies of real systems. We solve the 6 problems above with a simple new derivation of a simple new GGA functional in which *all* parameters [other than those in $\epsilon_{\mathrm{XC}}^{\mathrm{unif}}(n_\uparrow, n_\downarrow)$] are fundamental constants. Although the derivation depends only on the most general features of the real-space construction [13] behind PW91, the resulting functional is close to numerical GGA.

We begin with the GGA for correlation in the form

$$E_{\mathrm{C}}^{\mathrm{GGA}}[n_\uparrow, n_\downarrow] = \int d^3r\, n[\epsilon_{\mathrm{C}}^{\mathrm{unif}}(r_s, \zeta) + H(r_s, \zeta, t)], \qquad (3)$$

0031-9007/96/77(18)/3865(4)$10.00 © 1996 The American Physical Society 3865

Really?

```
alexandria3k query crossref 'April 2022 Public Data File from Crossref' --partition \
 --query "SELECT title FROM work_references
   LEFT JOIN works
     ON work_references.work_doi = works.doi
   WHERE work_references.doi = '10.1103/physrevlett.77.3865'"
```

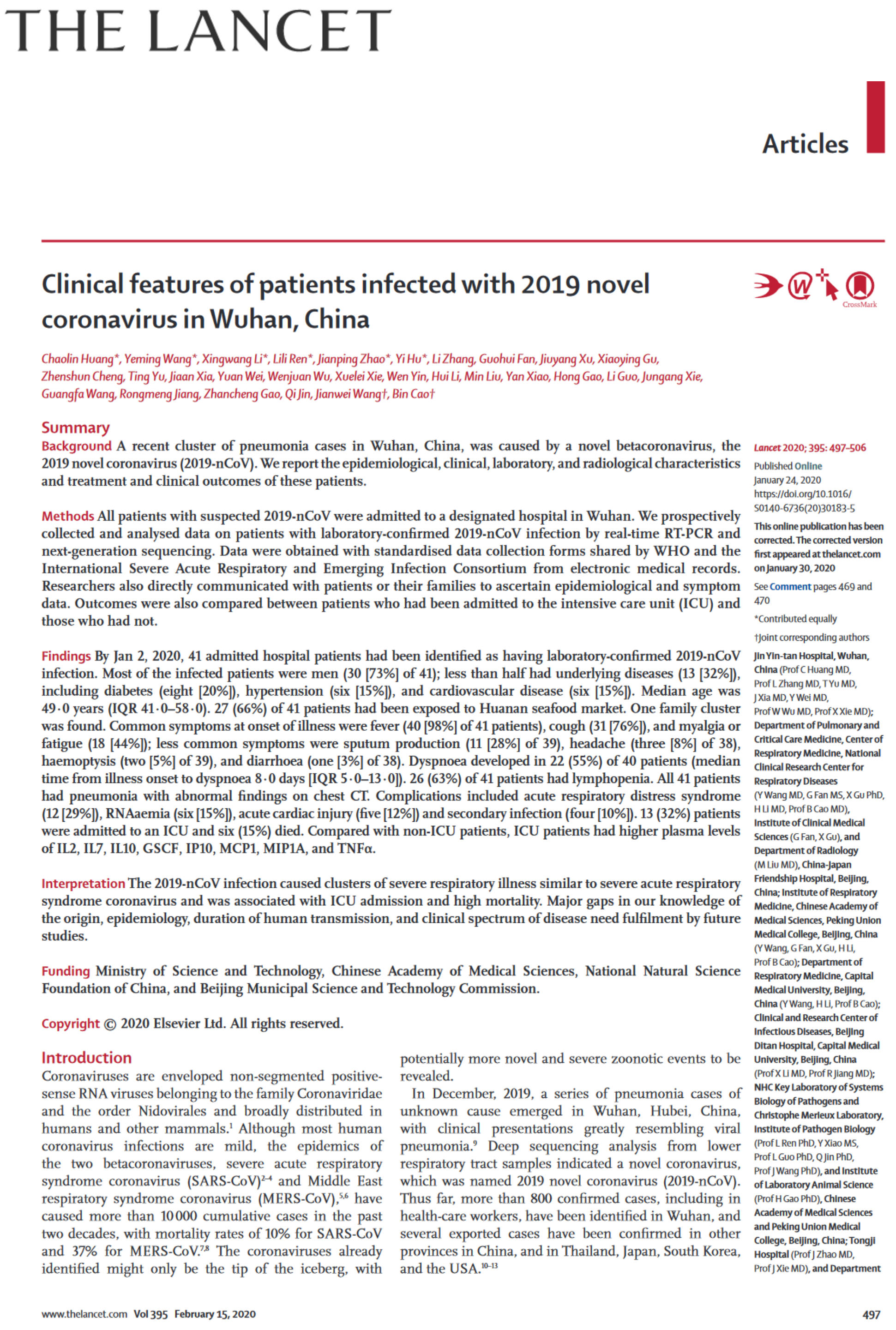
8 records per second

```
"Solid-liquid density and spin crossovers in (Mg, Fe)O system at deep mantle conditions"
Two-Dimensional BAs/InTe: A Promising Tandem Solar Cell with High Power Conversion Efficiency
Fatigue of graphene
Energetics of paramagnetic oxide clusters: the Fe(<scp>iii</scp>) oxyhydroxy Keggin ion
Stochastic many-body perturbation theory for Moiré states in twisted bilayer phosphorene
Dual-hybrid direct random phase approximation and second-order screened exchange with nonlocal van der Waals correlations for noncovalent interactions
Prediction on temperature dependent elastic constants of “soft” metal Al by AIMD and QHA
Triple VTe2/graphene/VTe2 heterostructures as perspective magnetic tunnel junctions
On the nature of homo- and hetero-dinuclear metal–metal quadruple bonds — Analysis of the bonding situation and benchmarking DFT against wave function methods
The extraordinary stability imparted to silver monolayers by chloride
Efficient Band Gap Prediction for Solids
Importance of Electronic Relaxation for Inter-Coulombic Decay in Aqueous Systems
Prediction of Reorganization Free Energies for Biological Electron Transfer: A Comparative Study of Ru-Modified Cytochromes and a 4-Helix Bundle Protein
…
```

```sql
-- Find the most cited articles in the period 2019-2021
-- published within that period

SELECT works.doi, Count(*)
  FROM work_references
  LEFT JOIN works ON work_references.doi = works.doi
  WHERE published_year BETWEEN 2019 AND 2021
  GROUP BY works.doi
  ORDER BY Count(*) DESC
  LIMIT 10;
```

48” elapsed time

21 424 citations



THE LANCET

Articles

Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China

Chaolin Huang*, Yeming Wang*, Xingwang Li*, Lili Ren*, Jianping Zhao*, Yi Hu*, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, Guangfa Wang, Rongmeng Jiang, Zhancheng Gao, Qi Jin, Jianwei Wang†, Bin Cao†

Summary

Background A recent cluster of pneumonia cases in Wuhan, China, was caused by a novel betacoronavirus, the 2019 novel coronavirus (2019-nCoV). We report the epidemiological, clinical, laboratory, and radiological characteristics and treatment and clinical outcomes of these patients.

Methods All patients with suspected 2019-nCoV were admitted to a designated hospital in Wuhan. We prospectively collected and analysed data on patients with laboratory-confirmed 2019-nCoV infection by real-time RT-PCR and next-generation sequencing. Data were obtained with standardised data collection forms shared by WHO and the International Severe Acute Respiratory and Emerging Infection Consortium from electronic medical records. Researchers also directly communicated with patients or their families to ascertain epidemiological and symptom data. Outcomes were also compared between patients who had been admitted to the intensive care unit (ICU) and those who had not.

Findings By Jan 2, 2020, 41 admitted hospital patients had been identified as having laboratory-confirmed 2019-nCoV infection. Most of the infected patients were men (30 [73%] of 41); less than half had underlying diseases (13 [32%]), including diabetes (eight [20%]), hypertension (six [15%]), and cardiovascular disease (six [15%]). Median age was 49·0 years (IQR 41·0–58·0). 27 (66%) of 41 patients had been exposed to Huanan seafood market. One family cluster was found. Common symptoms at onset of illness were fever (40 [98%] of 41 patients), cough (31 [76%]), and myalgia or fatigue (18 [44%]); less common symptoms were sputum production (11 [28%] of 39), headache (three [8%] of 38), haemoptysis (two [5%] of 39), and diarrhoea (one [3%] of 38). Dyspnoea developed in 22 (55%) of 40 patients (median time from illness onset to dyspnoea 8·0 days [IQR 5·0–13·0]). 26 (63%) of 41 patients had lymphopenia. All 41 patients had pneumonia with abnormal findings on chest CT. Complications included acute respiratory distress syndrome (12 [29%]), RNAaemia (six [15%]), acute cardiac injury (five [12%]) and secondary infection (four [10%]). 13 (32%) patients were admitted to an ICU and six (15%) died. Compared with non-ICU patients, ICU patients had higher plasma levels of IL2, IL7, IL10, GSCF, IP10, MCP1, MIP1A, and TNFα.

Interpretation The 2019-nCoV infection caused clusters of severe respiratory illness similar to severe acute respiratory syndrome coronavirus and was associated with ICU admission and high mortality. Major gaps in our knowledge of the origin, epidemiology, duration of human transmission, and clinical spectrum of disease need fulfilment by future studies.

Funding Ministry of Science and Technology, Chinese Academy of Medical Sciences, National Natural Science Foundation of China, and Beijing Municipal Science and Technology Commission.

Copyright © 2020 Elsevier Ltd. All rights reserved.

Introduction

Coronaviruses are enveloped non-segmented positive-sense RNA viruses belonging to the family Coronaviridae and the order Nidovirales and broadly distributed in humans and other mammals.[1] Although most human coronavirus infections are mild, the epidemics of the two betacoronaviruses, severe acute respiratory syndrome coronavirus (SARS-CoV)[2–4] and Middle East respiratory syndrome coronavirus (MERS-CoV),[5,6] have caused more than 10 000 cumulative cases in the past two decades, with mortality rates of 10% for SARS-CoV and 37% for MERS-CoV.[7,8] The coronaviruses already identified might only be the tip of the iceberg, with potentially more novel and severe zoonotic events to be revealed.

In December, 2019, a series of pneumonia cases of unknown cause emerged in Wuhan, Hubei, China, with clinical presentations greatly resembling viral pneumonia.[9] Deep sequencing analysis from lower respiratory tract samples indicated a novel coronavirus, which was named 2019 novel coronavirus (2019-nCoV). Thus far, more than 800 confirmed cases, including in health-care workers, have been identified in Wuhan, and several exported cases have been confirmed in other provinces in China, and in Thailand, Japan, South Korea, and the USA.[10–13]

Lancet 2020; 395: 497–506

Published Online January 24, 2020 https://doi.org/10.1016/S0140-6736(20)30183-5

This online publication has been corrected. The corrected version first appeared at thelancet.com on January 30, 2020

See Comment pages 469 and 470

*Contributed equally

†Joint corresponding authors

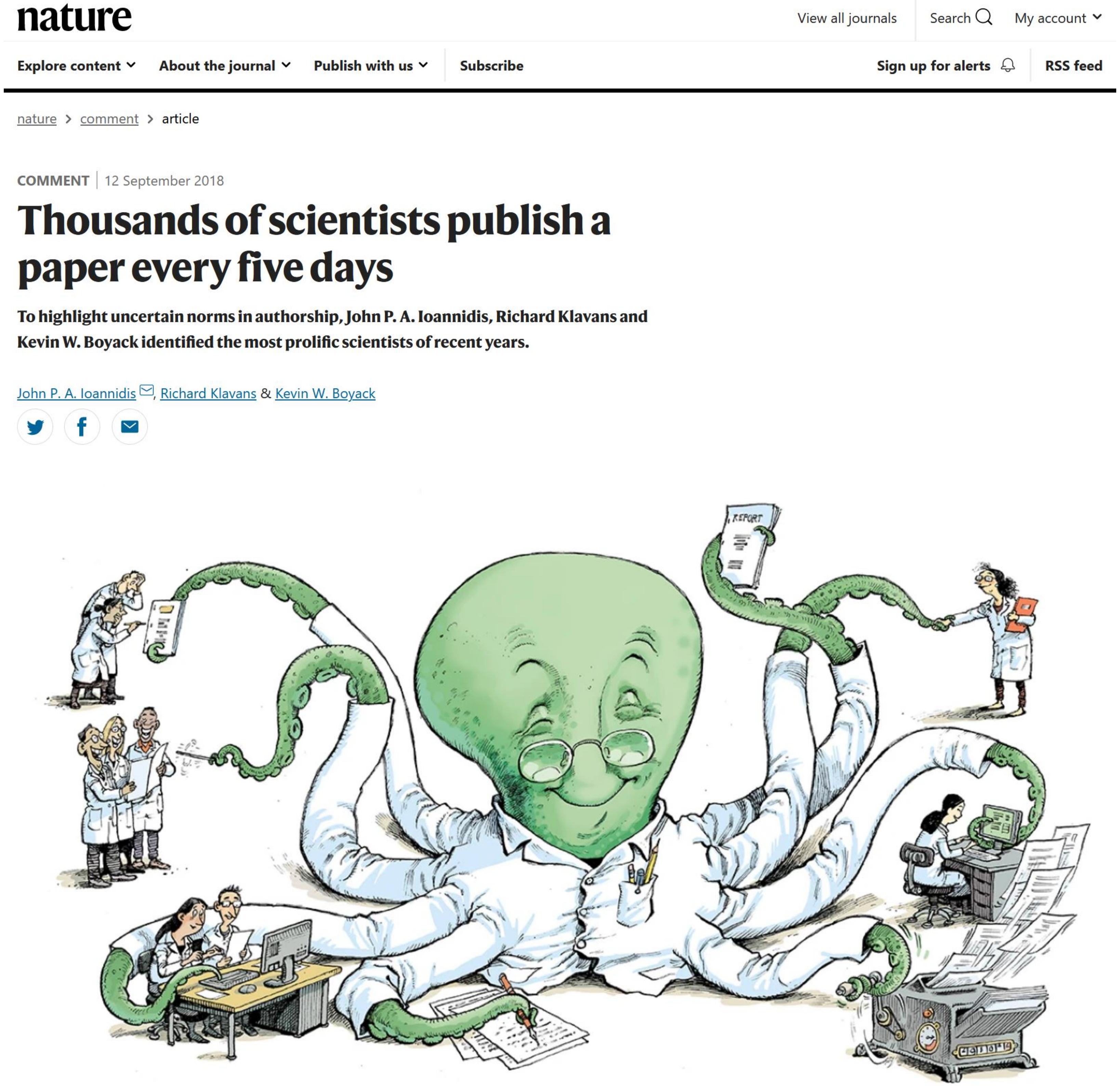
Jin Yin-tan Hospital, Wuhan, China (Prof C Huang MD, Prof L Zhang MD, T Yu MD, J Xia MD, Y Wei MD, Prof W Wu MD, Prof X Xie MD); Department of Pulmonary and Critical Care Medicine, Center of Respiratory Medicine, National Clinical Research Center for Respiratory Diseases (Y Wang MD, G Fan MS, X Gu PhD, H Li MD, Prof B Cao MD), Institute of Clinical Medical Sciences (G Fan, X Gu), and Department of Radiology (M Liu MD), China-Japan Friendship Hospital, Beijing, China; Institute of Respiratory Medicine, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China (Y Wang, G Fan, X Gu, H Li, Prof B Cao); Department of Respiratory Medicine, Capital Medical University, Beijing, China (Y Wang, H Li, Prof B Cao); Clinical and Research Center of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, Beijing, China (Prof X Li MD, Prof R Jiang MD); NHC Key Laboratory of Systems Biology of Pathogens and Christophe Merieux Laboratory, Institute of Pathogen Biology (Prof L Ren PhD, Y Xiao MS, Prof L Guo PhD, Q Jin PhD, Prof J Wang PhD), and Institute of Laboratory Animal Science (Prof H Gao PhD), Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China; Tongji Hospital (Prof J Zhao MD, Prof J Xie MD), and Department

www.thelancet.com Vol 395 February 15, 2020

497

Author h5-index

- Zhanhu Guo = 76 (15 papers / year)
- 12 authors > 60
- 100 > 38



COMMENT | 12 September 2018

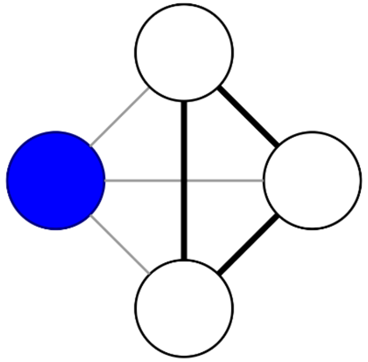
Thousands of scientists publish a paper every five days

To highlight uncertain norms in authorship, John P. A. Ioannidis, Richard Klavans and Kevin W. Boyack identified the most prolific scientists of recent years.

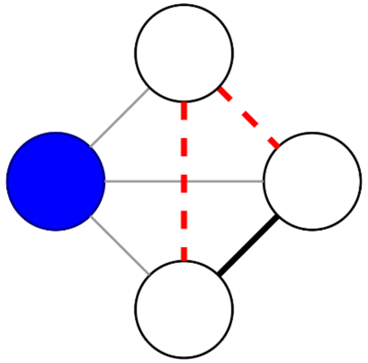
John P. A. Ioannidis, Richard Klavans & Kevin W. Boyack

How is this possible?

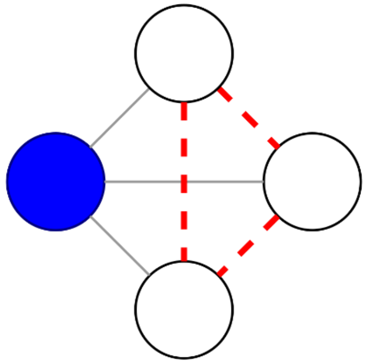
- Clustering coefficient of distance 2 citations
- **Significantly** different from other highly-cited papers
  - – For h5 > 50: median 0.05
  - – For random sample: median 0.03
  - – Mann-Whitney U test U_M=781, *p*-value 0.0006



c = 1

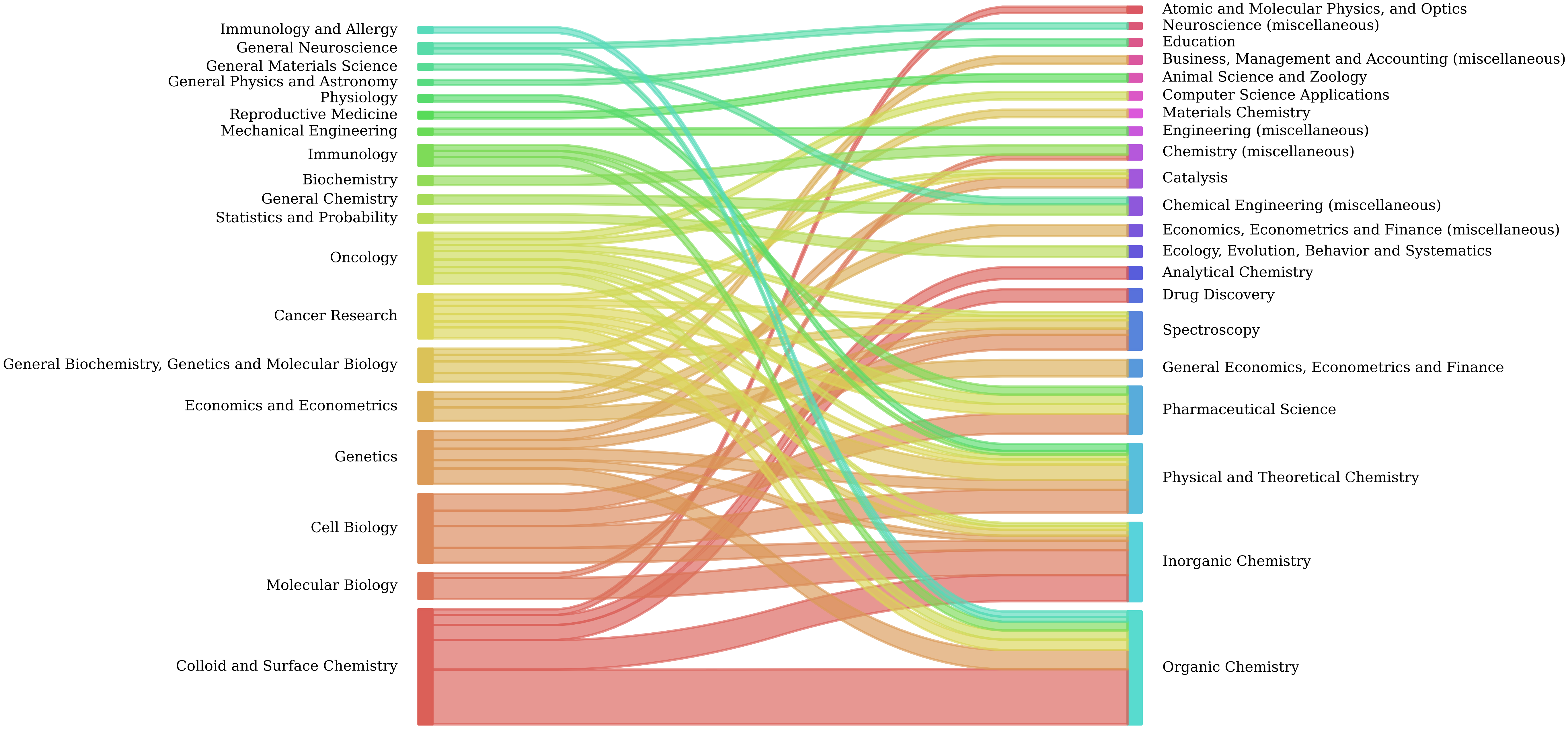


c = 1/3

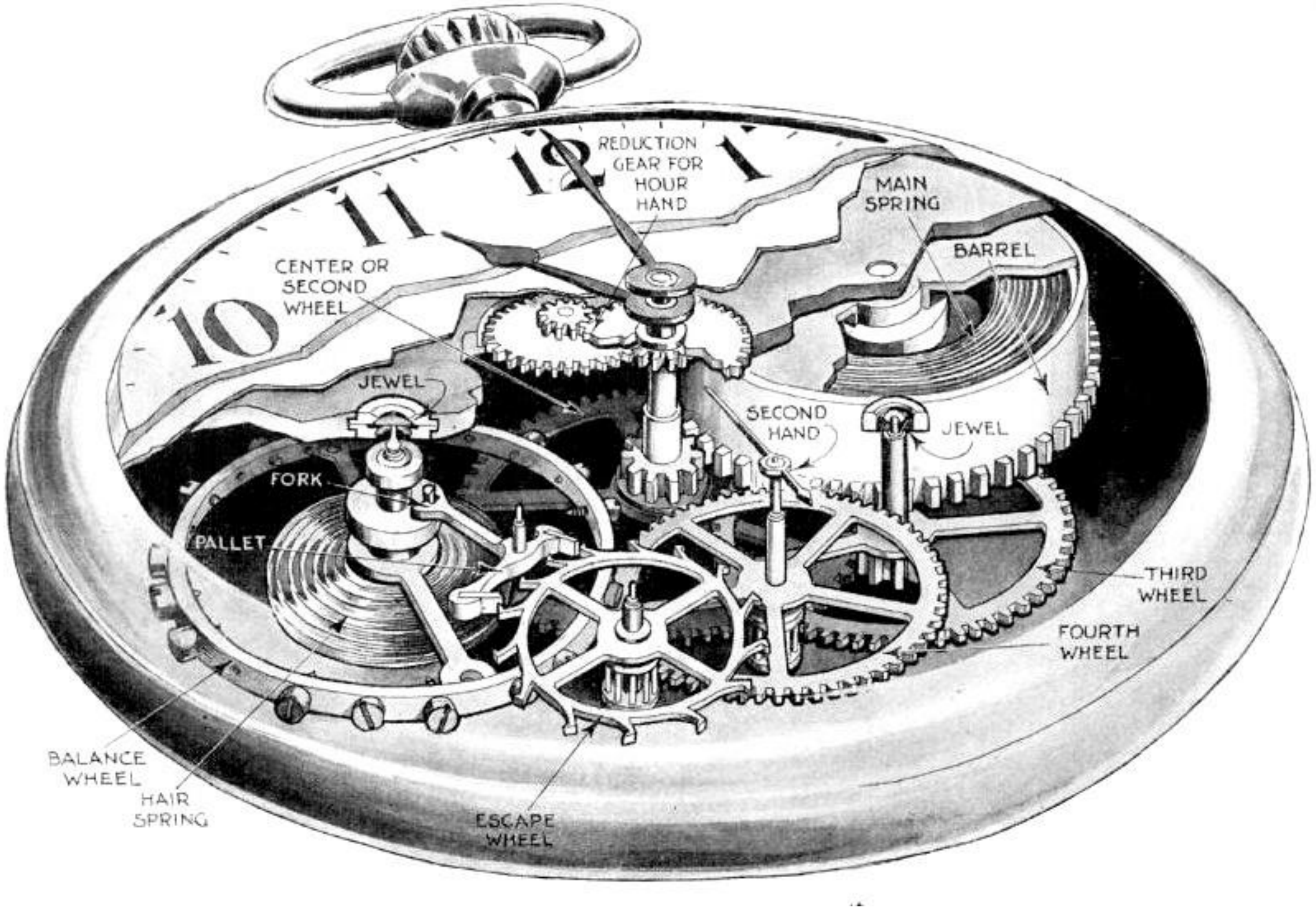


c = 0

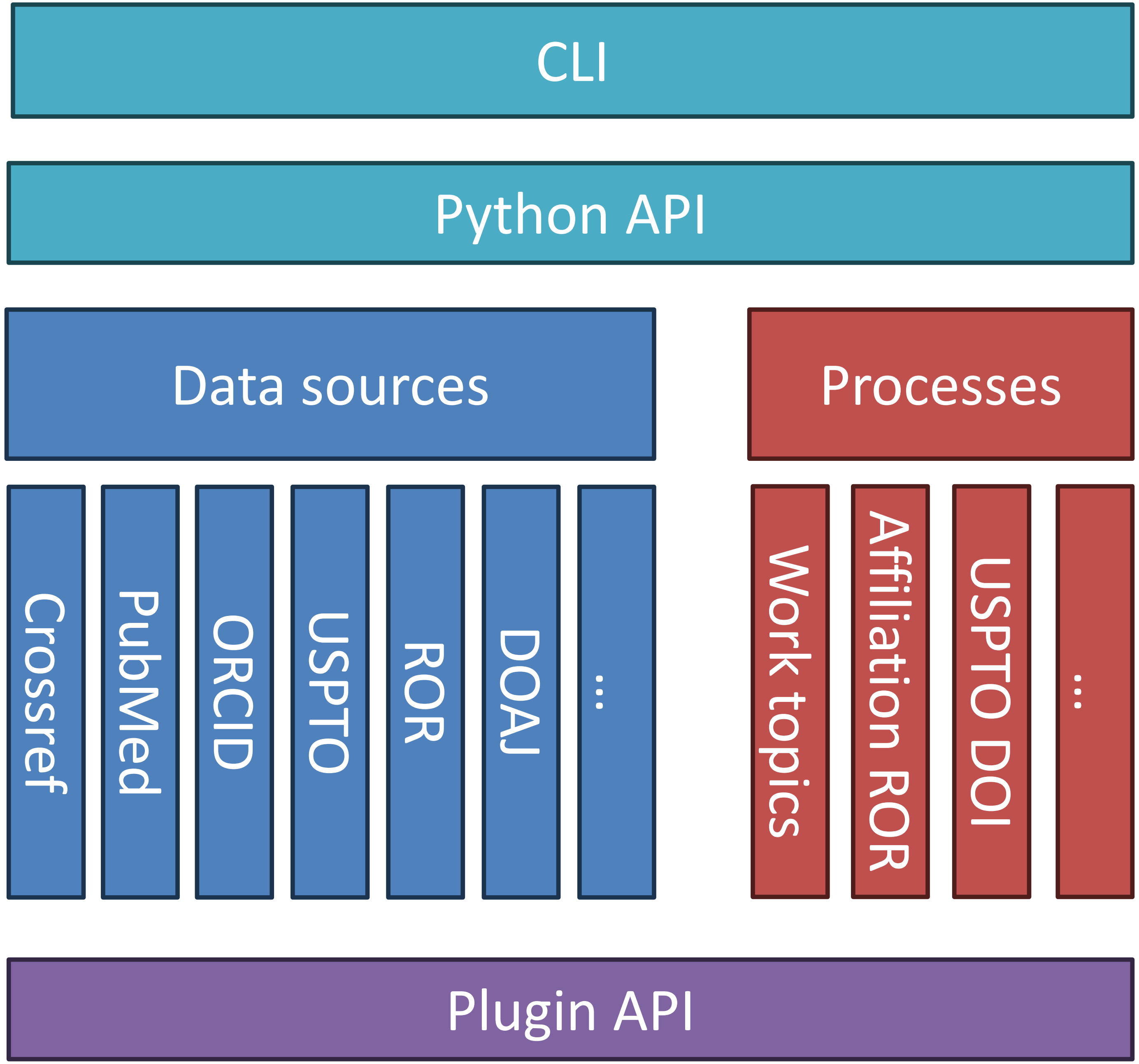
Scientific field dependencies



Implementation



Plugin-based architecture



Crossref key implementation ideas

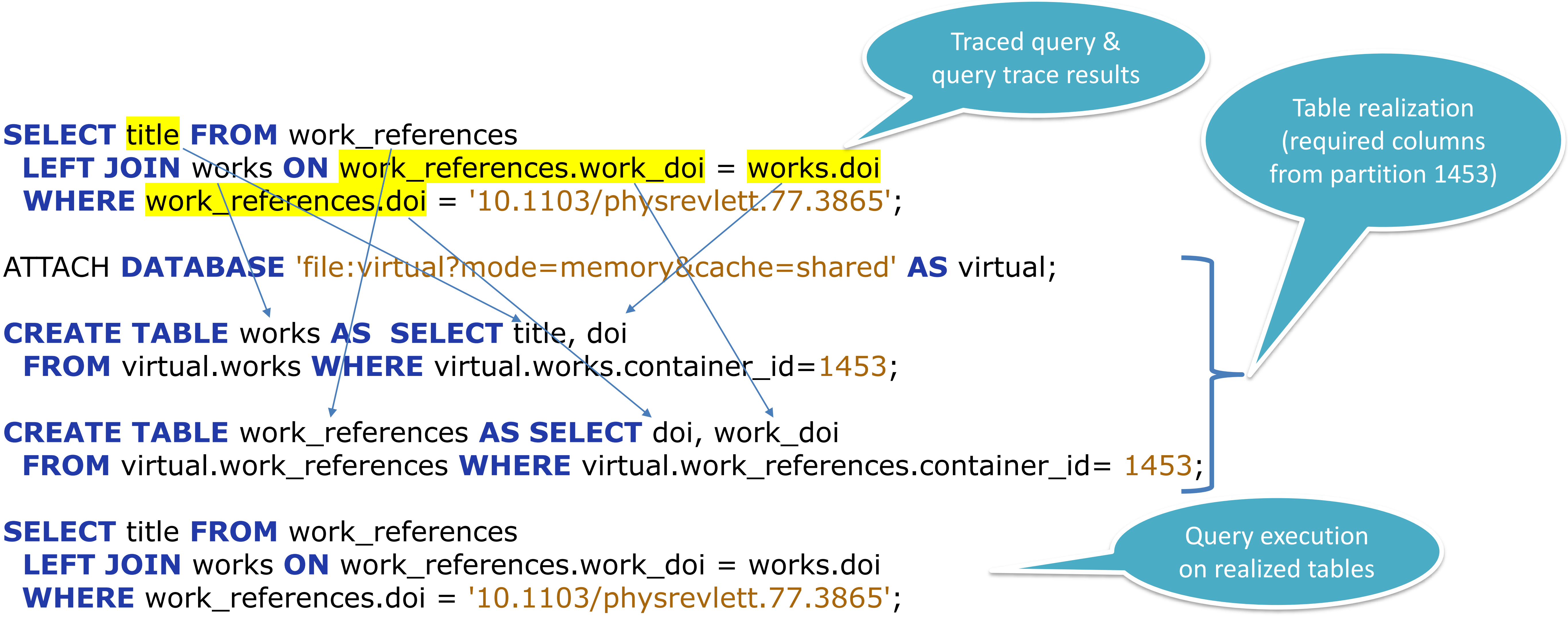
- SQLite + virtual tables
- Database partitioning, partition index
- Query tracing
- Realized vertical slices of partitions for queries
- PK, FK table with matched population query records

How to run Crossref query on 1 TB (simple case)

```
alexandria3k query crossref 'April 2022 Public Data File from Crossref' \
  --query "SELECT doi FROM work_references where doi is not null"
```

```
CREATE VIRTUAL TABLE work_references USING filesource();

SELECT doi FROM work_references where doi is not null;
```

How to run Crossref query on 1 TB



```
SELECT title FROM work_references
  LEFT JOIN works ON work_references.work_doi = works.doi
  WHERE work_references.doi = '10.1103/physrevlett.77.3865';

ATTACH DATABASE 'file:virtual?mode=memory&cache=shared' AS virtual;

CREATE TABLE works AS  SELECT title, doi
  FROM virtual.works WHERE virtual.works.container_id=1453;

CREATE TABLE work_references AS SELECT doi, work_doi
  FROM virtual.work_references WHERE virtual.work_references.container_id= 1453;

SELECT title FROM work_references
  LEFT JOIN works ON work_references.work_doi = works.doi
  WHERE work_references.doi = '10.1103/physrevlett.77.3865';
```

Traced query &
query trace results

Table realization
(required columns
from partition 1453)

Query execution
on realized tables

Crossref population: simple case

```sql
INSERT INTO populated.works
    SELECT works.title, works.doi FROM works
    WHERE works.container_id = 0;

INSERT INTO populated.work_authors
    SELECT work_authors.* FROM work_authors
    WHERE work_authors.container_id = 0;

INSERT INTO populated.works
    SELECT works.title, works.doi FROM works
    WHERE works.container_id = 1;

[…]
```

Conditional Crossref population 1/2

```
alexandria3k populate lis.db crossref … \
  --row-selection "work_subjects.name = 'Library and Information Sciences' "
  --columns works.title works.doi work_authors.orcid work_subjects.*
```

```
ATTACH DATABASE 'lis.db' AS populated;

SELECT DISTINCT 1 FROM works, work_authors, author_affiliations, …
  WHERE work_subjects.name = 'Library and Information Sciences';

CREATE TABLE populated.works(doi, container_id, title, …);
[…]

CREATE TEMP TABLE temp_works AS
  SELECT id, rowid FROM works WHERE container_id = 0;

CREATE TEMP TABLE temp_work_subjects AS
  SELECT rowid, name, work_id FROM work_subjects WHERE container_id = 0;

CREATE TEMP TABLE temp_work_authors AS
  SELECT rowid, work_id FROM work_authors WHERE container_id = 0;
[…]
```

Traced query & query trace results

Populated tables

Tables with PKs, FKs and query fields

Conditional Crossref population 2/2

```
CREATE TEMP TABLE temp_matched AS
  SELECT works.id, works.rowid
  FROM temp_works AS works
  LEFT JOIN temp_work_subjects AS work_subjects
    ON works.id = work_subjects.work_id
  WHERE (work_subjects.name = 'Library and Information Sciences');

INSERT INTO populated.work_authors
  SELECT work_authors.orcid FROM work_authors
  WHERE work_authors.container_id = 0
    AND EXISTS (SELECT 1
      FROM temp_matched AS temp_works
        LEFT JOIN temp_work_authors
          ON temp_works.id = temp_work_authors.work_id
            AND work_authors.rowid = temp_work_authors.rowid);

[…]
```

Key to all partition records matching the specified condition

Topologically ordered table JOINs

Populate tables with partition's data based on matched records

ORCID/USPTO key implementation ideas

- Stream-based
  - Web fetch
  - Decompress
  - Tar records
- Skip XML parsing where possible

Issues and limitations

- Low ORCID coverage:
  - Only 17/360 million author records
- Affiliations missing / appear in diverse forms
- Only 11% of Crossref records have an abstract
- Subjects cover only Scopus-indexed journals
- Difficulty of determining “citable items”

Way forward

- Help community to conduct studies
- Integrate more OA data
  - arXiv, DBLP, MESH, PLoS taxonomy, …
- Improve processes
  - Author & org disambiguation, topic classification, …
- Evangelize more and better data availability
  - ORCID
  - Publication metadata improvements

Thank you!

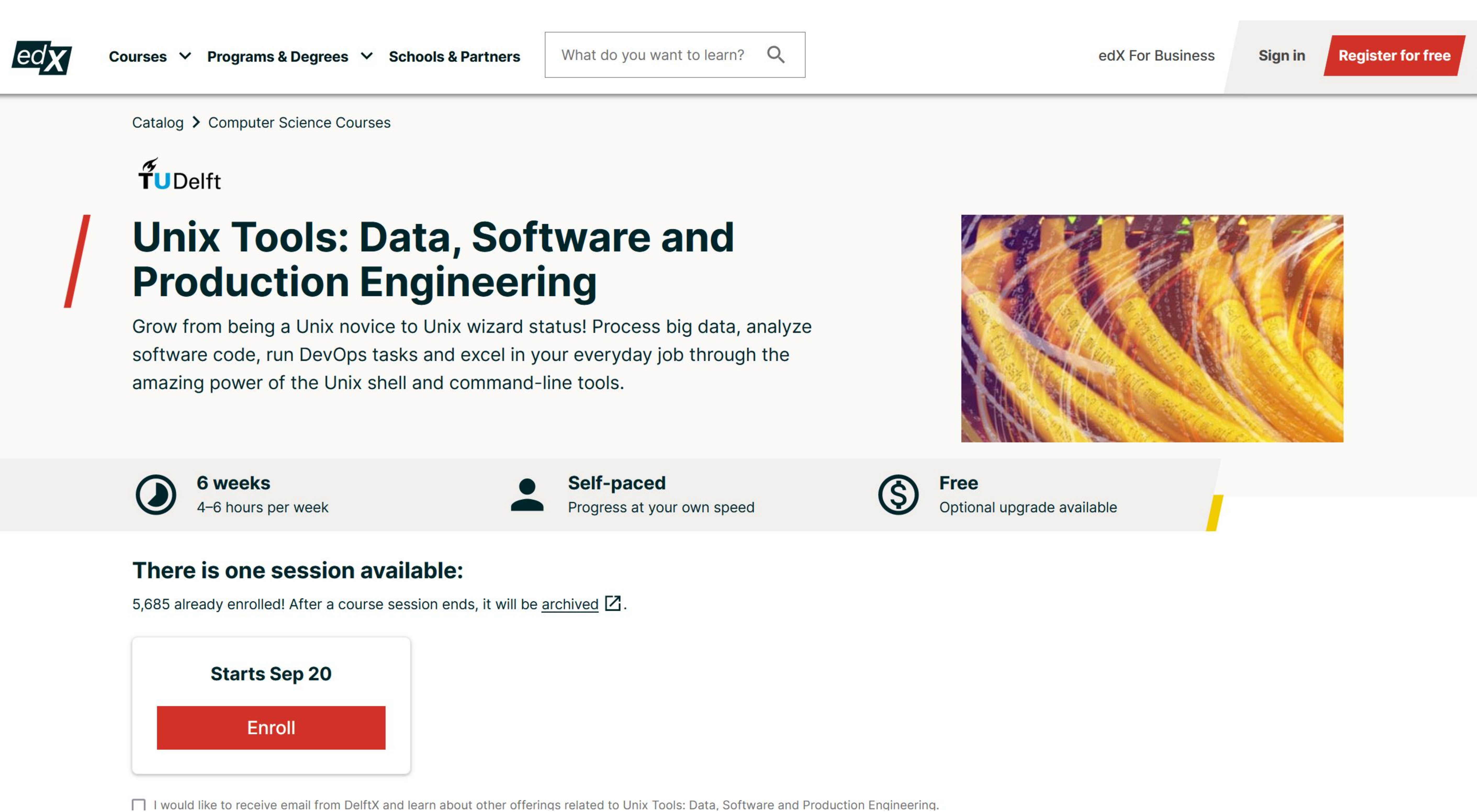
github.com/dspinellis/alexandria3k

🌐 www.spinellis.gr

𝕏 @CoolSWEng

@CoolSWEng@mastodon.acm.org

✉ dds@aueb.gr



Unix Tools: Data, Software and Production Engineering

Grow from being a Unix novice to Unix wizard status! Process big data, analyze software code, run DevOps tasks and excel in your everyday job through the amazing power of the Unix shell and command-line tools.

6 weeks
4–6 hours per week

Self-paced
Progress at your own speed

Free
Optional upgrade available

There is one session available:

5,685 already enrolled! After a course session ends, it will be archived.

Starts Sep 20

□ I would like to receive email from DelftX and learn about other offerings related to Unix Tools: Data, Software and Production Engineering.