Big Data and the birth of a Science of Humanities

The 1st International Conference on Humanities and Big Data in Ibero-America

Wenceslao Arroyo-Machado
ORCID: 0000-0001-9437-8757

Nicolas Robinson-Garcia
ORCID: 0000-0002-0585-7359

This study is part of the COMPARE project (REF: PID2020-117007RA-I00) which aims at studying diversity of profiles in academia.

We would like to thank DIALNET for providing access to their data

COMPARE
UGR PROJECT

MINISTERIO
DE CIENCIA
E INNOVACIÓN

AGENCIA
ESTATAL DE
INVESTIGACIÓN

Agenda

1. The challenges of the Humanities - Why are they always neglected by scientometrics?

2. Big Data and why it can be a game changer

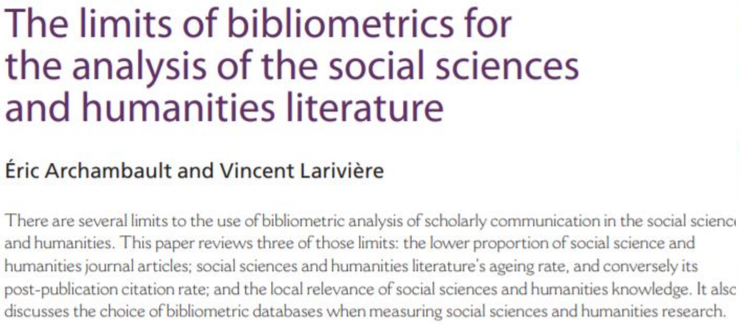
CASE STUDY 1: Humanities under the lens of Big Data

CASE STUDY 2: Profiling humanists with Archetypal Analysis

3. Towards a **Science of Humanities**

The challenges of the Humanities

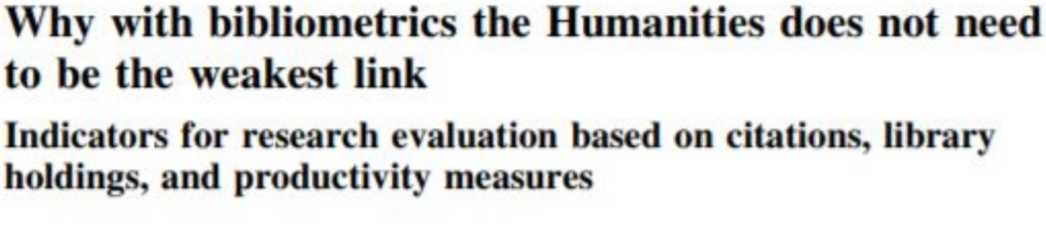
GARY NEILL FOR THE CHRONICLE REVIEW



The limits of bibliometrics for the analysis of the social sciences and humanities literature

Éric Archambault and Vincent Larivière

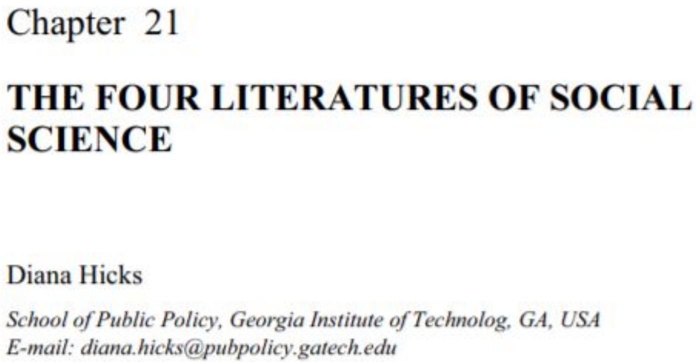
There are several limits to the use of bibliometric analysis of scholarly communication in the social scienc and humanities. This paper reviews three of those limits: the lower proportion of social science and humanities journal articles; social sciences and humanities literature's ageing rate, and conversely its post-publication citation rate; and the local relevance of social sciences and humanities knowledge. It alsc discusses the choice of bibliometric databases when measuring social sciences and humanities research.



Why with bibliometrics the Humanities does not need to be the weakest link

Indicators for research evaluation based on citations, library holdings, and productivity measures

A. J. M. Linmans

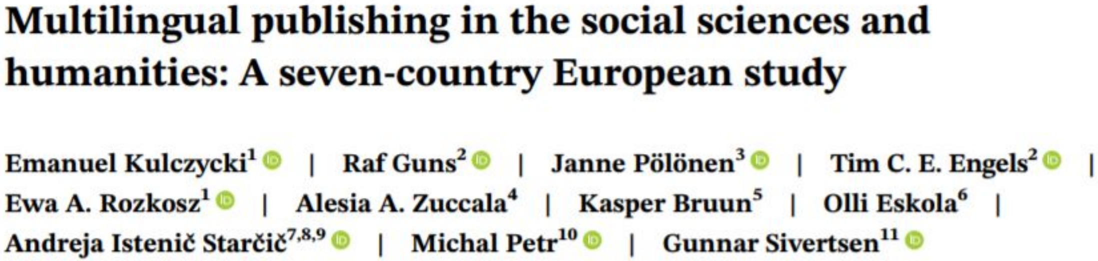


Chapter 21

THE FOUR LITERATURES OF SOCIAL SCIENCE

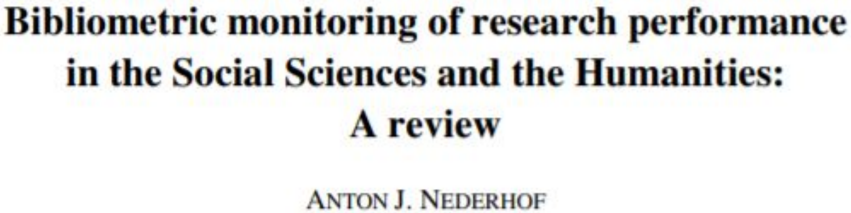
Diana Hicks

School of Public Policy, Georgia Institute of Technolog, GA, USA
E-mail: diana.hicks@pubpolicy.gatech.edu



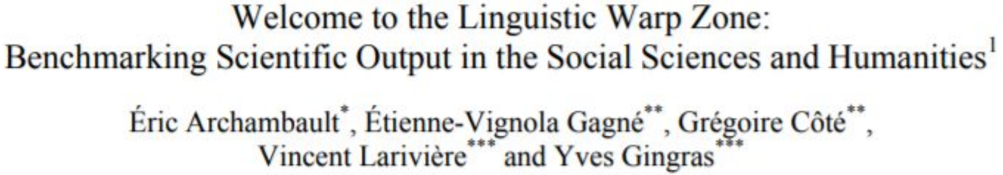
Multilingual publishing in the social sciences and humanities: A seven-country European study

Emanuel Kulczycki[1] | Raf Guns[2] | Janne Pölönen[3] | Tim C. E. Engels[2] | Ewa A. Rozkosz[1] | Alesia A. Zuccala[4] | Kasper Bruun[5] | Olli Eskola[6] | Andreja Istenič Starčič[7,8,9] | Michal Petr[10] | Gunnar Sivertsen[11]



Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review

ANTON J. NEDERHOF

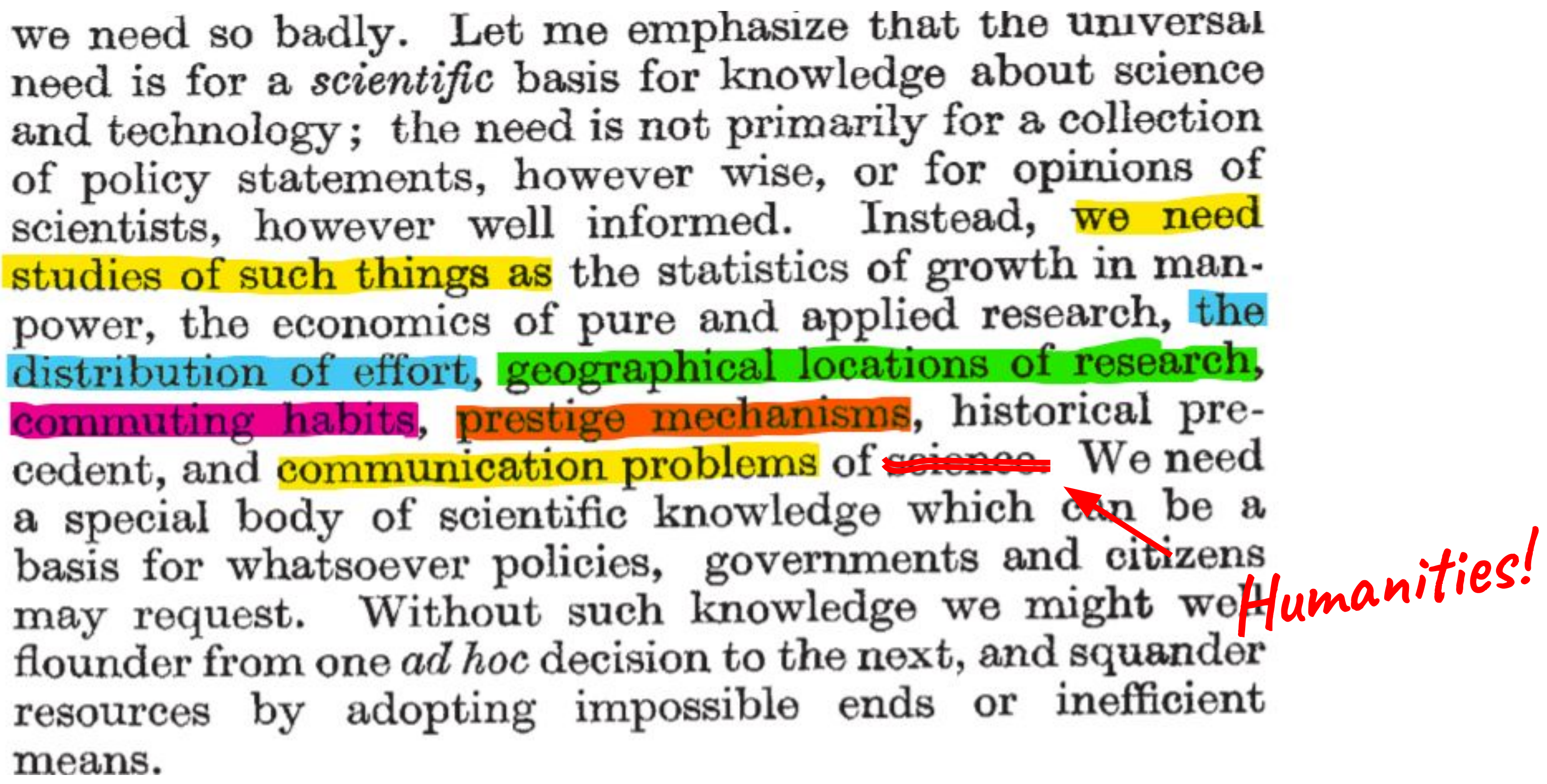


Welcome to the Linguistic Warp Zone: Benchmarking Scientific Output in the Social Sciences and Humanities[1]

Éric Archambault*, Étienne-Vignola Gagné**, Grégoire Côté**, Vincent Larivière*** and Yves Gingras***

> “ [I]t is much easier to make a measurement than to ascertain just what has been measured ”

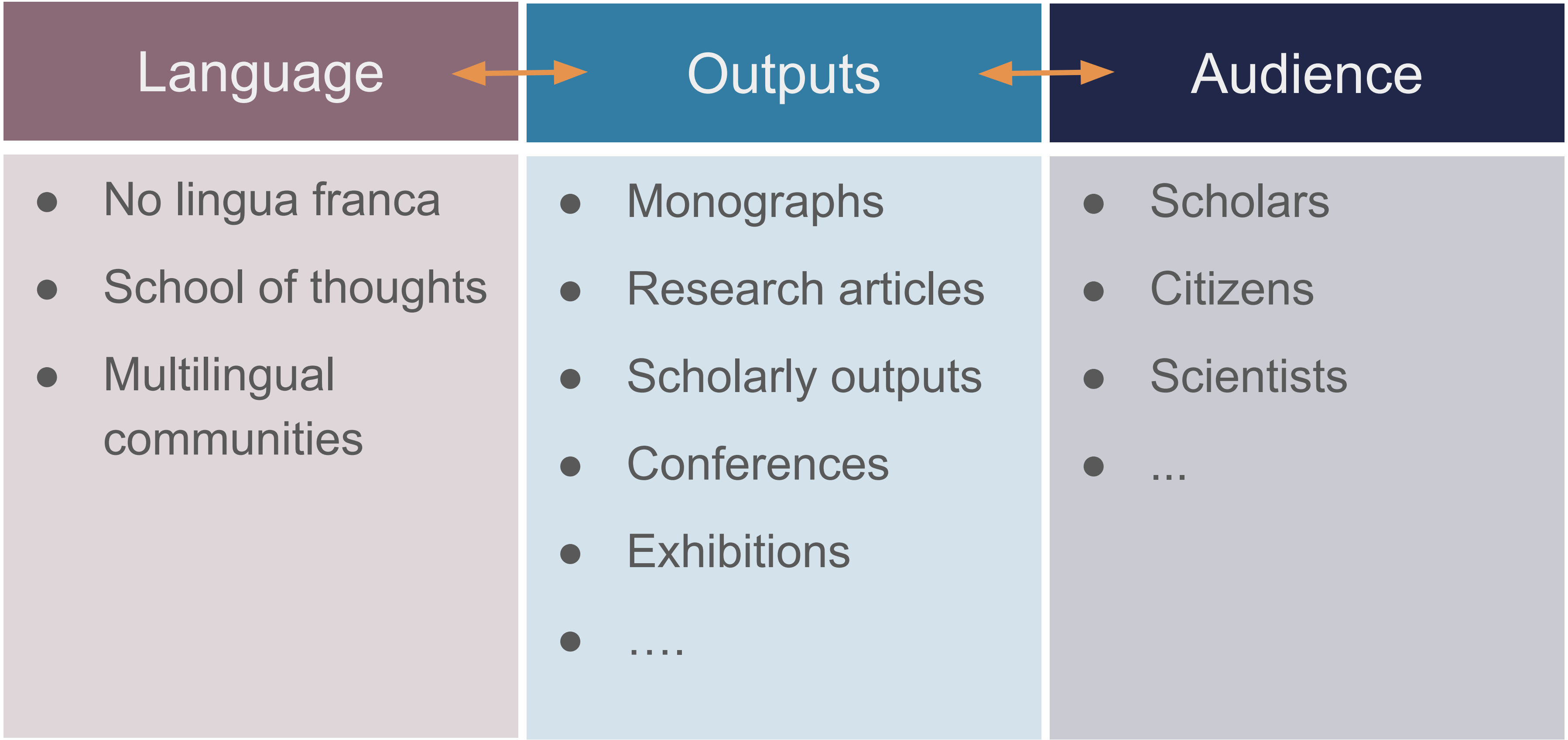
De Solla Price, Derek J. ‘Quantitative Measures of the Development of Science’. In *Archives Internationales d’Histoire Des Sciences*, 14:85–93. Amsterdam, 1951.



we need so badly. Let me emphasize that the universal need is for a *scientific* basis for knowledge about science and technology; the need is not primarily for a collection of policy statements, however wise, or for opinions of scientists, however well informed. Instead, we need studies of such things as the statistics of growth in manpower, the economics of pure and applied research, the distribution of effort, geographical locations of research, commuting habits, prestige mechanisms, historical precedent, and communication problems of ~~science.~~ We need a special body of scientific knowledge which can be a basis for whatsoever policies, governments and citizens may request. Without such knowledge we might well flounder from one *ad hoc* decision to the next, and squander resources by adopting impossible ends or inefficient means.

Humanities!

De Solla Price, Derek J. 'The Scientific Foundations of Science Policy'. *Nature* 206, no. 4981 (1965): 233–38.



Language

- No lingua franca
- School of thoughts
- Multilingual communities

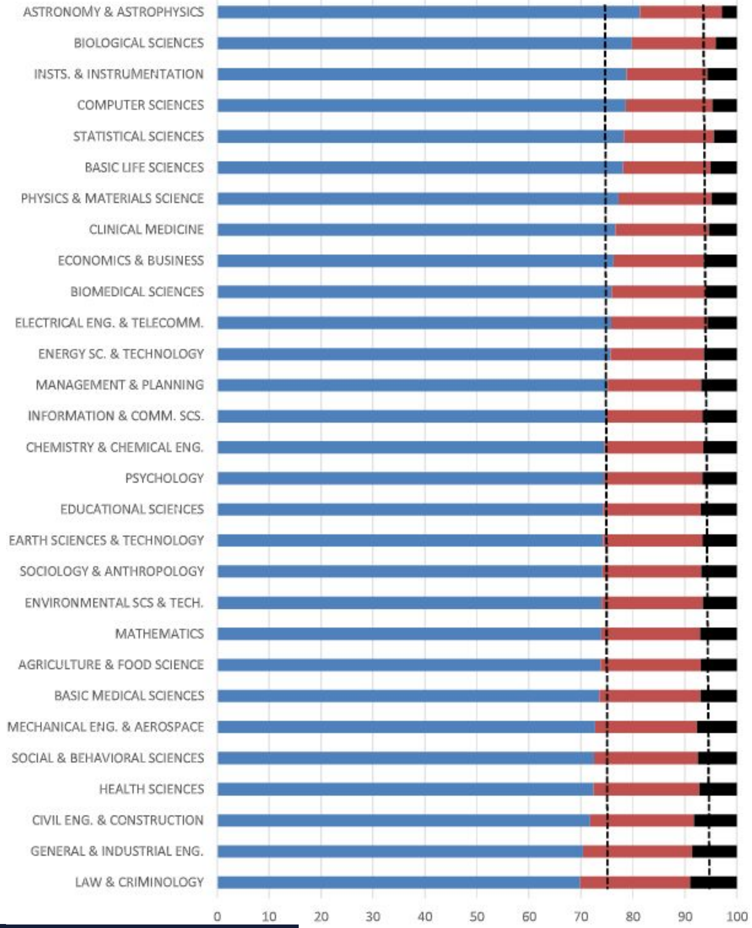
Outputs

- Monographs
- Research articles
- Scholarly outputs
- Conferences
- Exhibitions
- ….

Audience

- Scholars
- Citizens
- Scientists
- ...

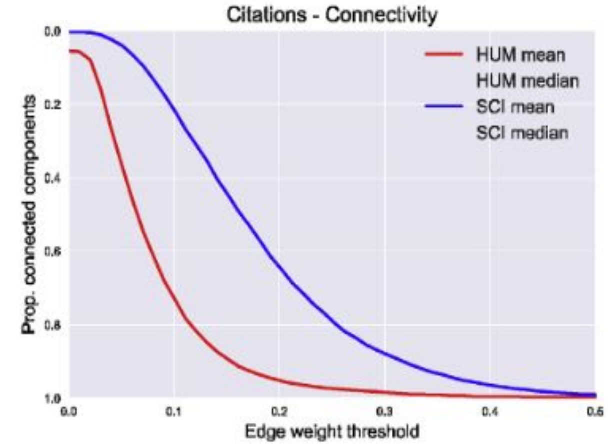
- Scientometrics is built on the notion that science is governed by highly skewed distributions.

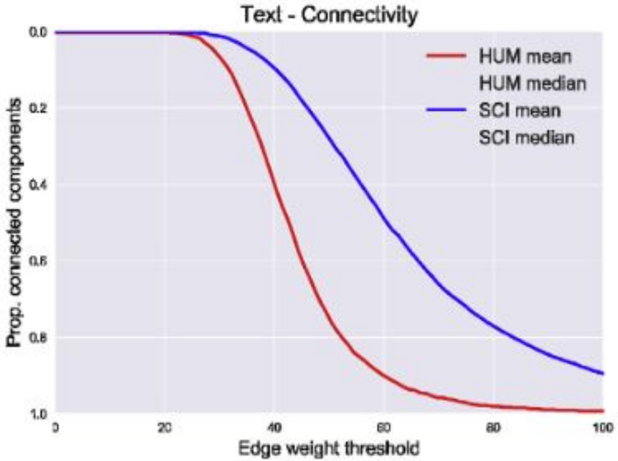


Ruiz-Castillo, Javier, and Rodrigo Costas. ‘Individual and Field Citation Distributions in 29 Broad Scientific Fields’. *Journal of Informetrics* 12, no. 3 (1 August 2018): 868–92.

- Scientometrics is built on the notion that science is governed by highly skewed distributions.

- Humanities fields show more scattered and disconnected communities than general sciences

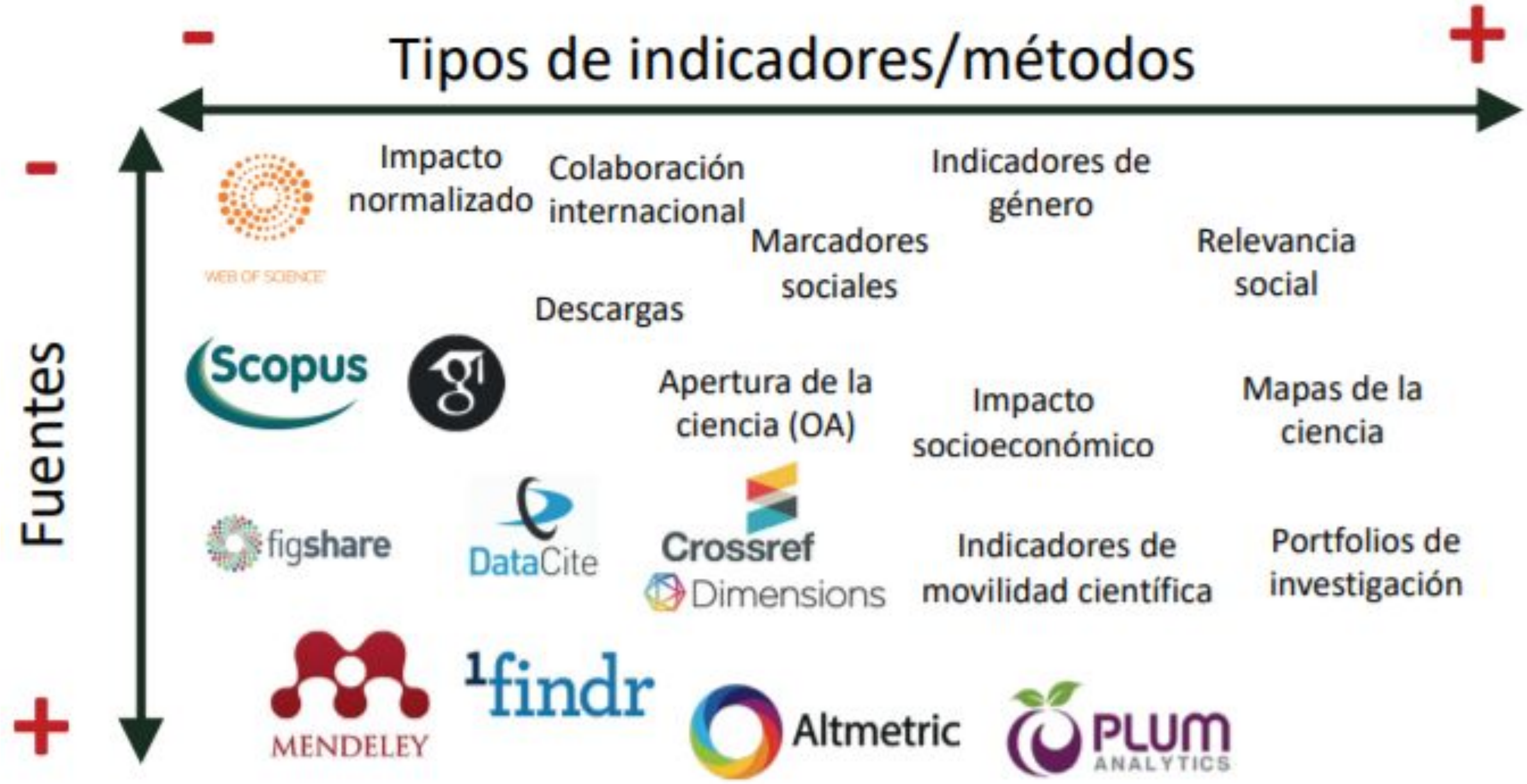




Colavizza, Giovanni, Thomas Franssen, and Thed van Leeuwen. ‘An Empirical Investigation of the Tribes and Their Territories: Are Research Specialisms Rural and Urban?’ *Journal of Informetrics* 13, no. 1 (1 February 2019): 105–17.

The opportunities of Big Data

Photo by h heyerlein on Unsplash



Tipos de indicadores/métodos

Fuentes

Impacto normalizado

Colaboración internacional

Indicadores de género

Marcadores sociales

Relevancia social

Descargas

Apertura de la ciencia (OA)

Impacto socioeconómico

Mapas de la ciencia

Indicadores de movilidad científica

Portfolios de investigación

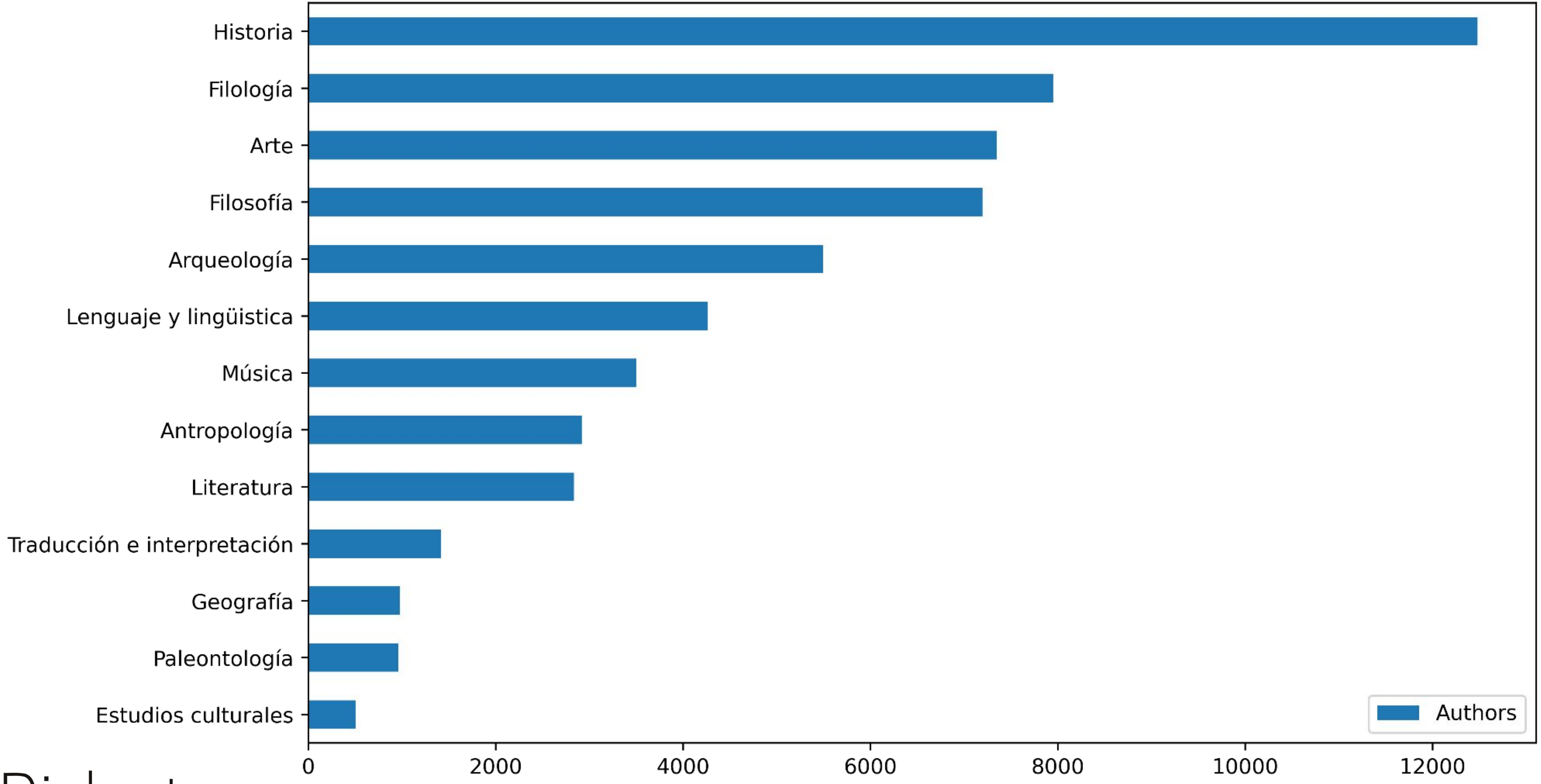
Robinson-García, Nicolás, Rafael Repiso, and Daniel Torres-Salinas. 'Perspectiva y retos de los profesionales de la evaluación científica y la bibliometría'. *El Profesional de la Información* 27, no. 3 (17 June 2018): 461–66.

DISCIPLINES	AUTHORS
History	12,480
Philology	7,951
Arts	7,351
Philosophy	7,197
Archaeology	5,495
Language & Linguistics	4,264
Music	3,497
Anthropology	2,920
Literature	2,834
Translation & Interpretation	1,416
Geography	981
Paleontology	961
Cultural Studies	504

57,851 scholars
752,423 publications

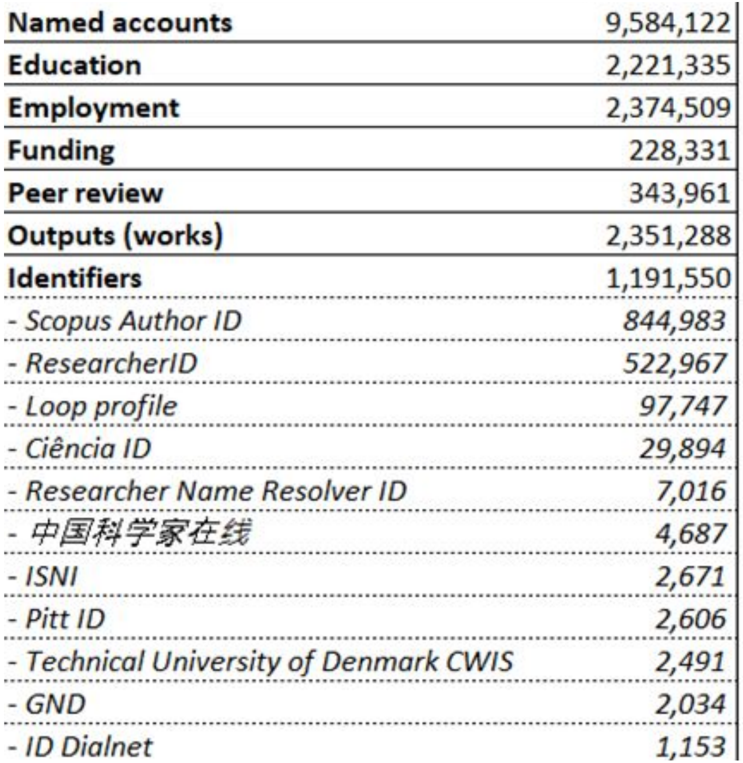
416,537 articles
213,088 chapters
89,195 books
33,603 proceedings

Dialnet

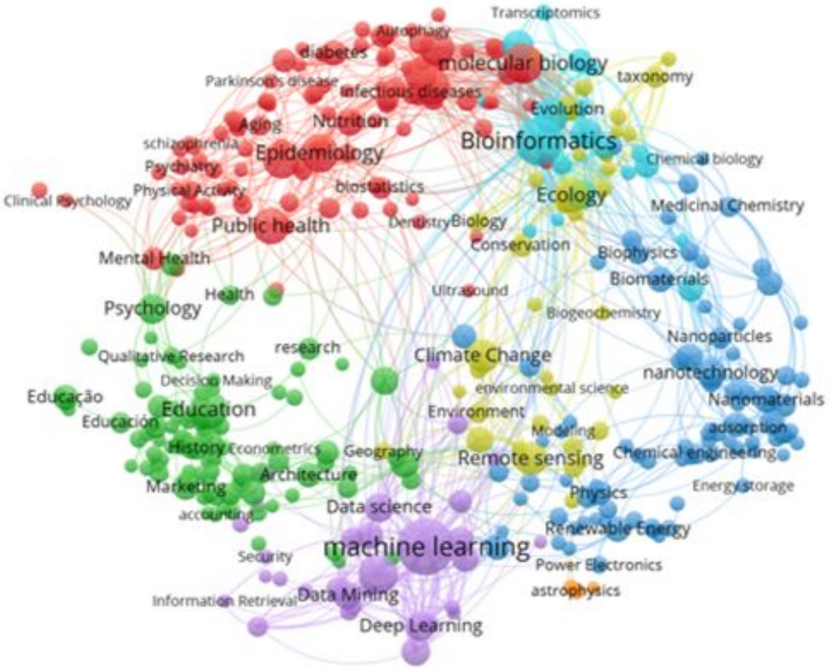


Dialnet

Distribution by fields of humanists



Named accounts	9,584,122
Education	2,221,335
Employment	2,374,509
Funding	228,331
Peer review	343,961
Outputs (works)	2,351,288
Identifiers	1,191,550
- Scopus Author ID	*844,983*
- ResearcherID	*522,967*
- Loop profile	*97,747*
- Ciência ID	*29,894*
- Researcher Name Resolver ID	*7,016*
- 中国科学家在线	*4,687*
- ISNI	*2,671*
- Pitt ID	*2,606*
- Technical University of Denmark CWIS	*2,491*
- GND	*2,034*
- ID Dialnet	*1,153*

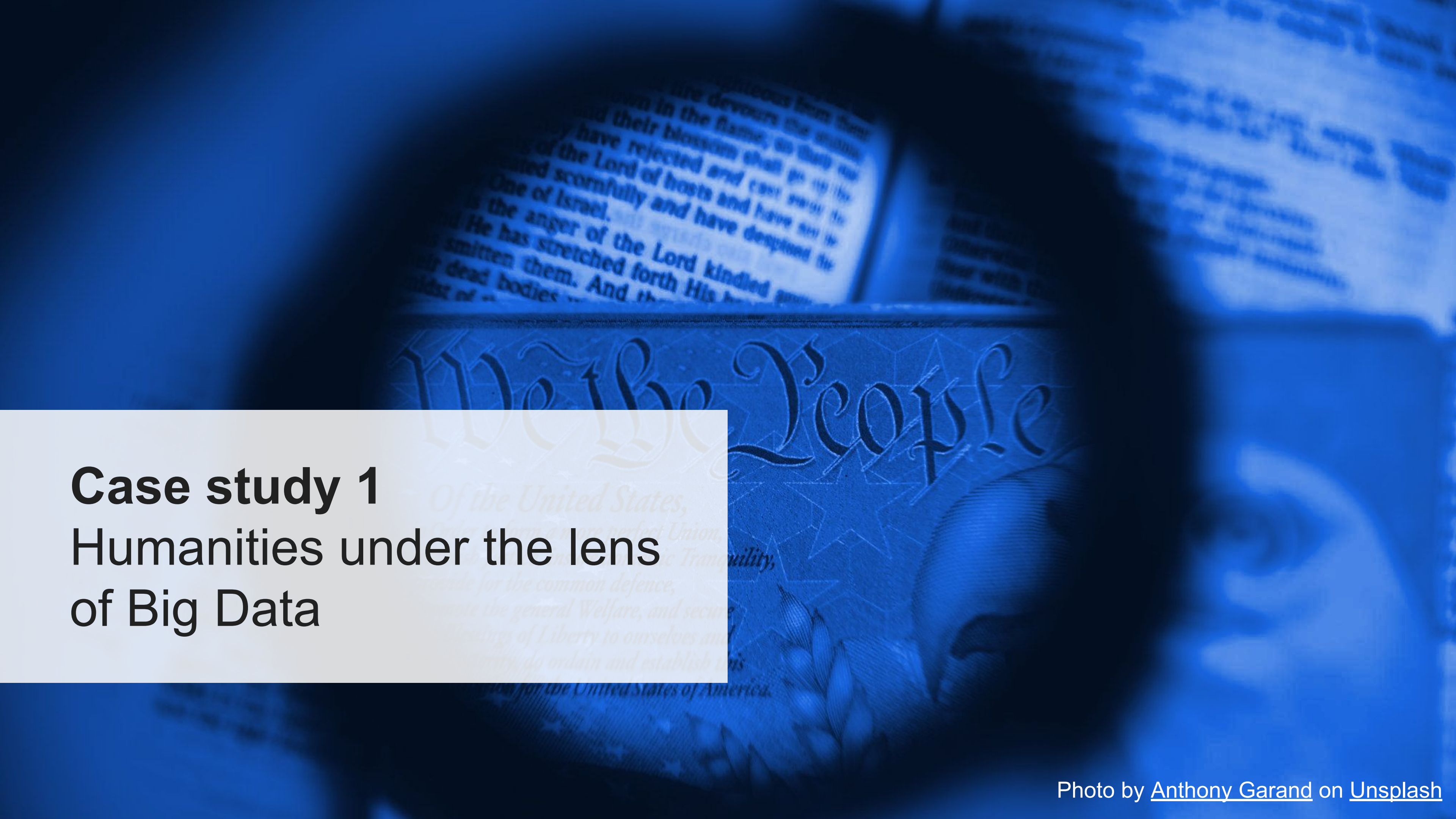


Costas, R., Corona, C., Robinson-Garcia, N. Could ORCID play a key role in meta-research? Discussing new analytical possibilities to study the dynamics of science and scientists. In A. Oancea, G. Derrick & N. Nuseibeh (eds.). *Handbook on Meta-Research*. Edward Elgar.

ORCID

stands for

Open Researcher and Contributor ID



Case study 1
Humanities under the lens of Big Data

Photo by Anthony Garand on Unsplash

Research questions

1. Can we identify humanists and their outputs beyond traditional or specialised scientific databases?
2. Can we identify and map research topics?
3. Do we observe their publication patterns at a macro level?

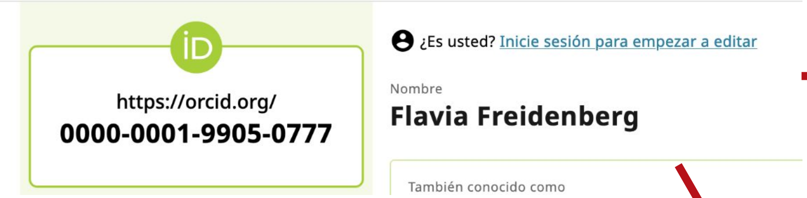
Rationale

Phase 1 Use of author keywords and departments to identify humanists in ORCID

Phase 2 Matching of author and publication records using NLP techniques between ORCID and Dialnet

Phase 3 Publication trend analysis

ORCID metadata



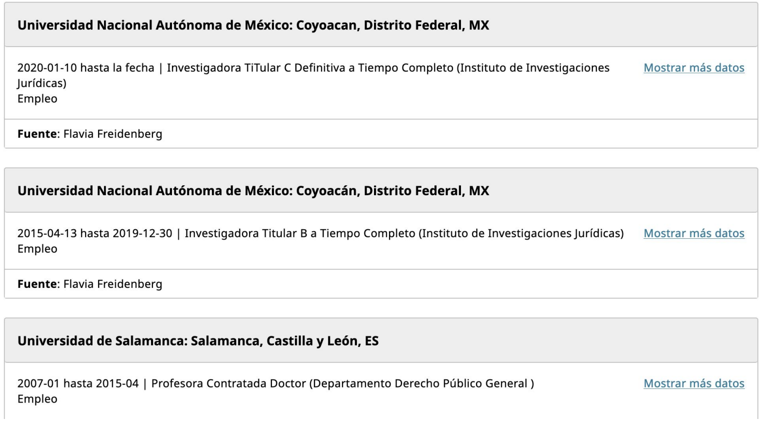
https://orcid.org/
0000-0001-9905-0777

¿Es usted? Inicie sesión para empezar a editar

Nombre
Flavia Freidenberg

También conocido como

Affiliation (1,943,623 records)



Universidad Nacional Autónoma de México: Coyoacan, Distrito Federal, MX

2020-01-10 hasta la fecha | Investigadora TiTular C Definitiva a Tiempo Completo (Instituto de Investigaciones Jurídicas)
Empleo

Mostrar más datos

Fuente: Flavia Freidenberg

Universidad Nacional Autónoma de México: Coyoacán, Distrito Federal, MX

2015-04-13 hasta 2019-12-30 | Investigadora Titular B a Tiempo Completo (Instituto de Investigaciones Jurídicas)
Empleo

Mostrar más datos

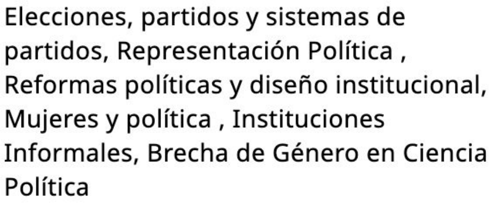
Fuente: Flavia Freidenberg

Universidad de Salamanca: Salamanca, Castilla y León, ES

2007-01 hasta 2015-04 | Profesora Contratada Doctor (Departamento Derecho Público General)
Empleo

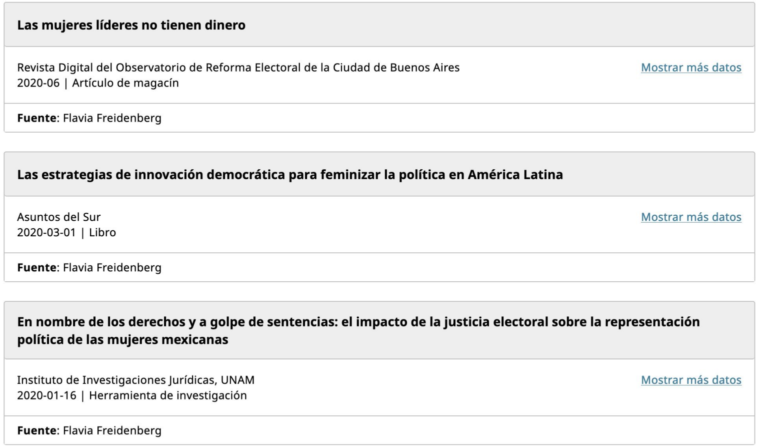
Mostrar más datos

Keywords (588,794 records)



Elecciones, partidos y sistemas de partidos, Representación Política , Reformas políticas y diseño institucional, Mujeres y política , Instituciones Informales, Brecha de Género en Ciencia Política

Works



Las mujeres líderes no tienen dinero

Revista Digital del Observatorio de Reforma Electoral de la Ciudad de Buenos Aires
2020-06 | Artículo de magacín

Mostrar más datos

Fuente: Flavia Freidenberg

Las estrategias de innovación democrática para feminizar la política en América Latina

Asuntos del Sur
2020-03-01 | Libro

Mostrar más datos

Fuente: Flavia Freidenberg

En nombre de los derechos y a golpe de sentencias: el impacto de la justicia electoral sobre la representación política de las mujeres mexicanas

Instituto de Investigaciones Jurídicas, UNAM
2020-01-16 | Herramienta de investigación

Mostrar más datos

Fuente: Flavia Freidenberg

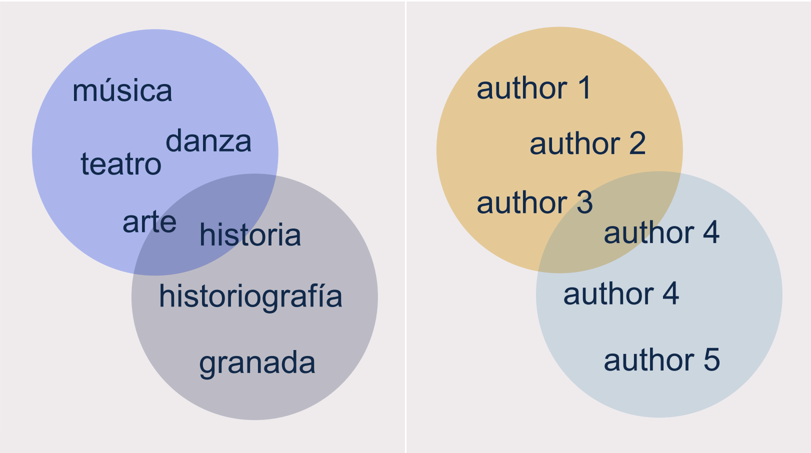
	NLP	Machine learning
Methods	Tokenization Bag of words POS tagging Word stemming	Clustering Classification Regression

NLP: NORMALIZATION

Simplify	Remove ambiguity
Dept. de FILOLOGÍA ↓ dept filologia	TORRES-SALINAS, D. ↕ Daniel Torres-Salinas
Noise reduction	**Merge**
~~Departamento de Humanidades y Educación~~	Dialnet + ORCID

Machine learning: DISCOVERING

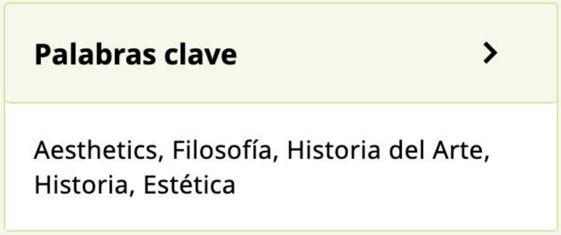
Community detection



Applications

Keyword clustering

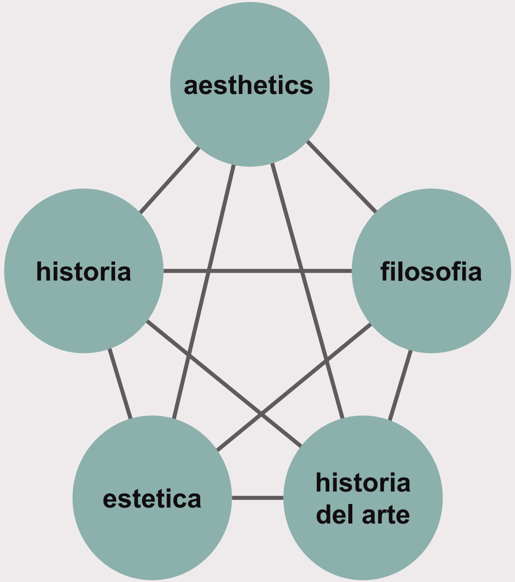
1. Preprocessing



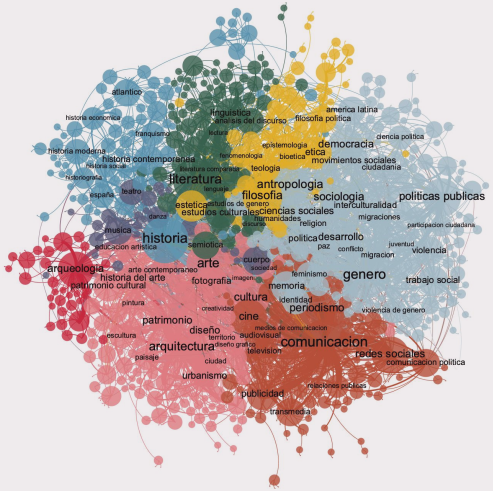
Palabras clave ›

Aesthetics, Filosofía, Historia del Arte, Historia, Estética

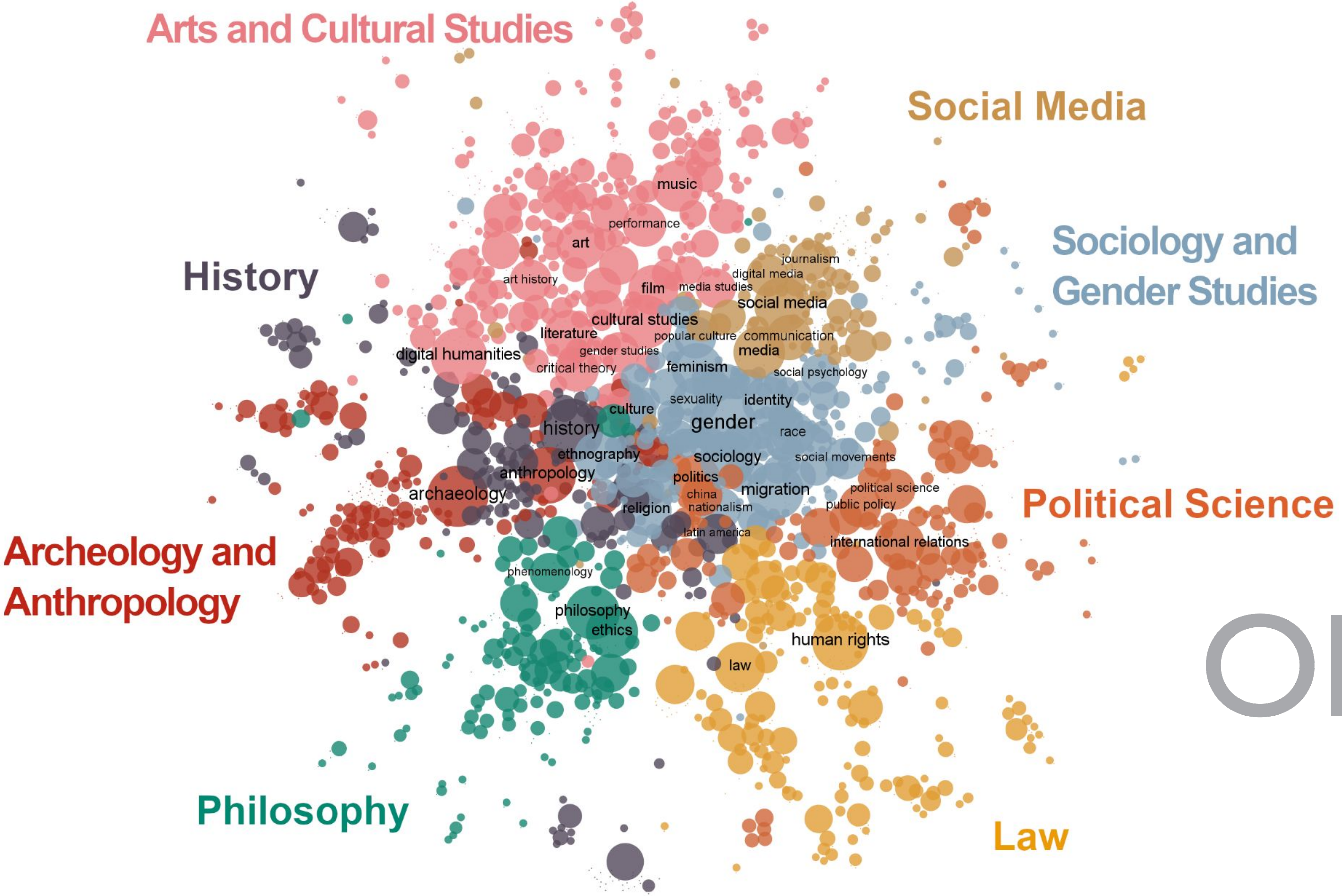
2. Network creation



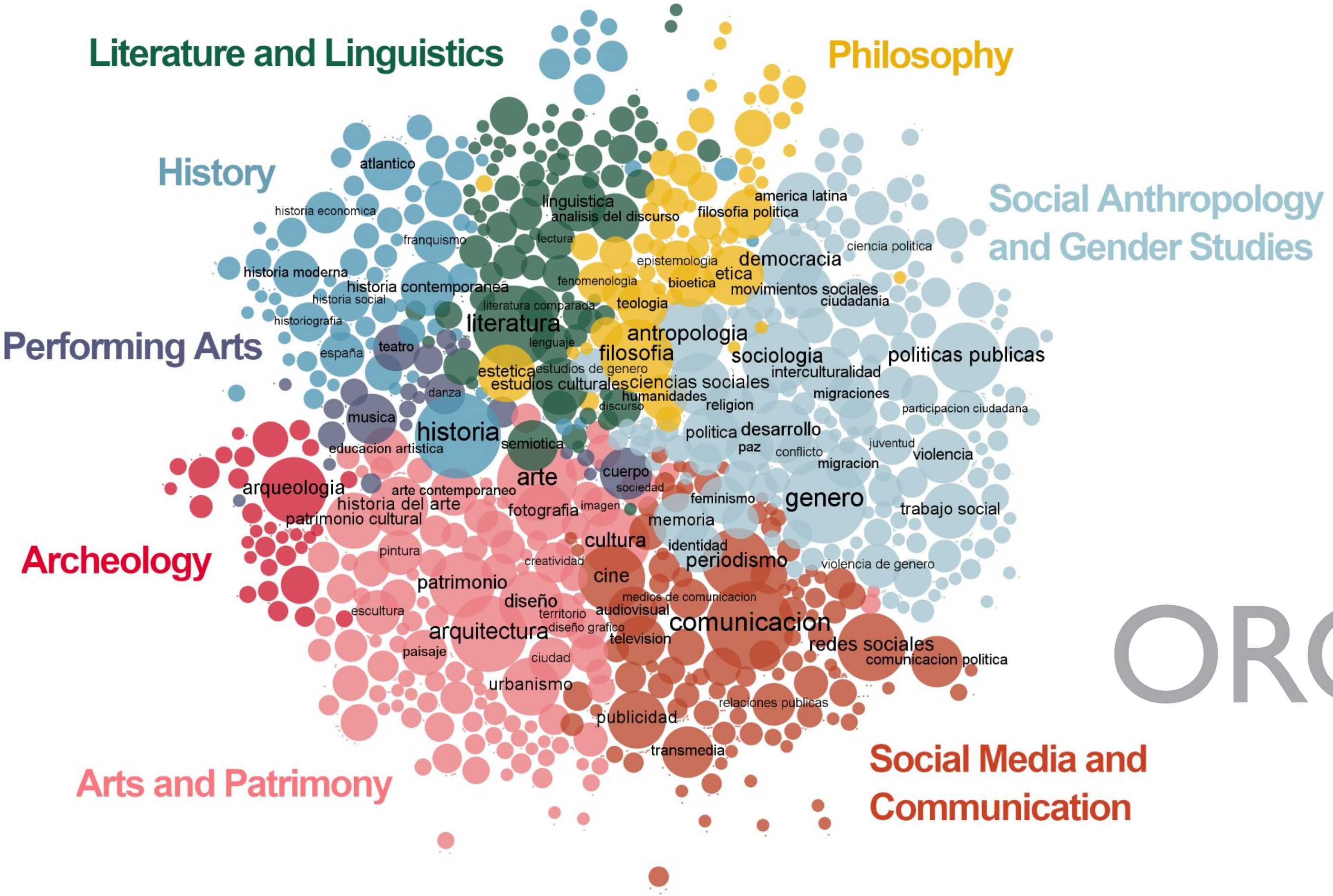
3. Community detection



Co-occurrence network of author keywords for English speaking humanists



Co-occurrence network of author keywords for Spanish speaking humanists

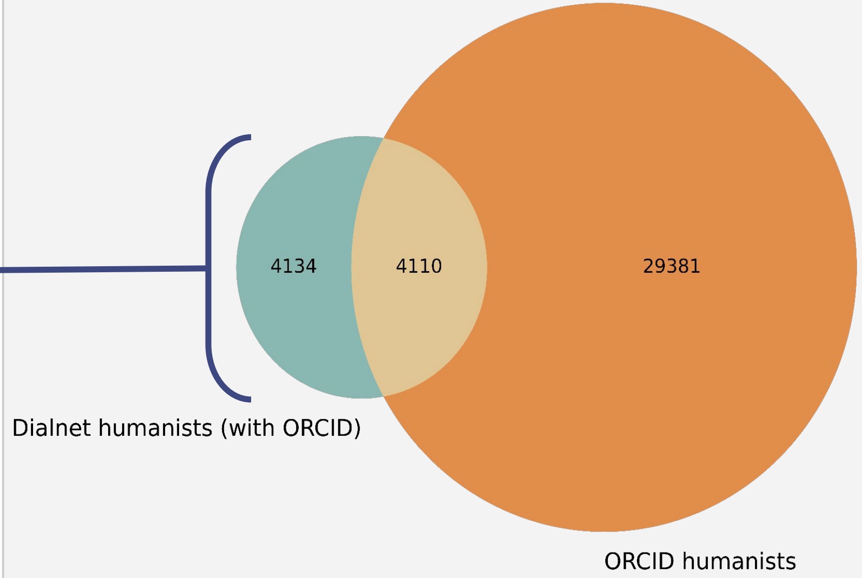


Overlap between data sources

Dialnet ∉ ORCID

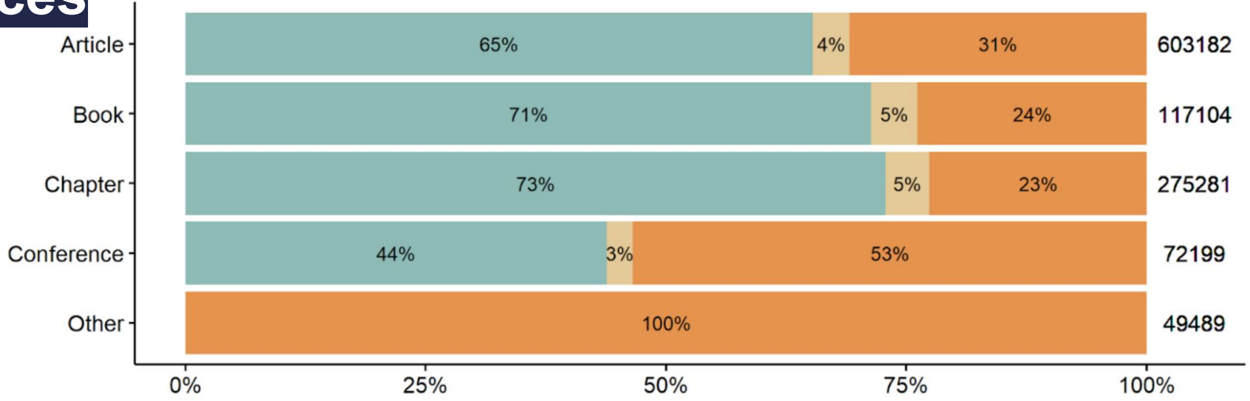
- Errors in Dialnet → Wrong field assignment
- Errors in ORCID → Private or deleted profile

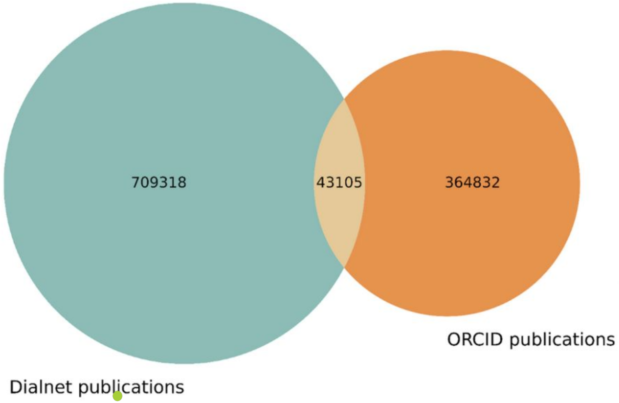
ONLY SCHOLARS WITH ORCID

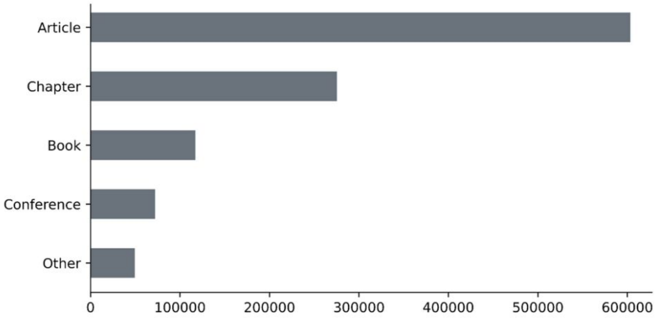


Dialnet + ORCID

Overlap of publications for common scholars in both data sources

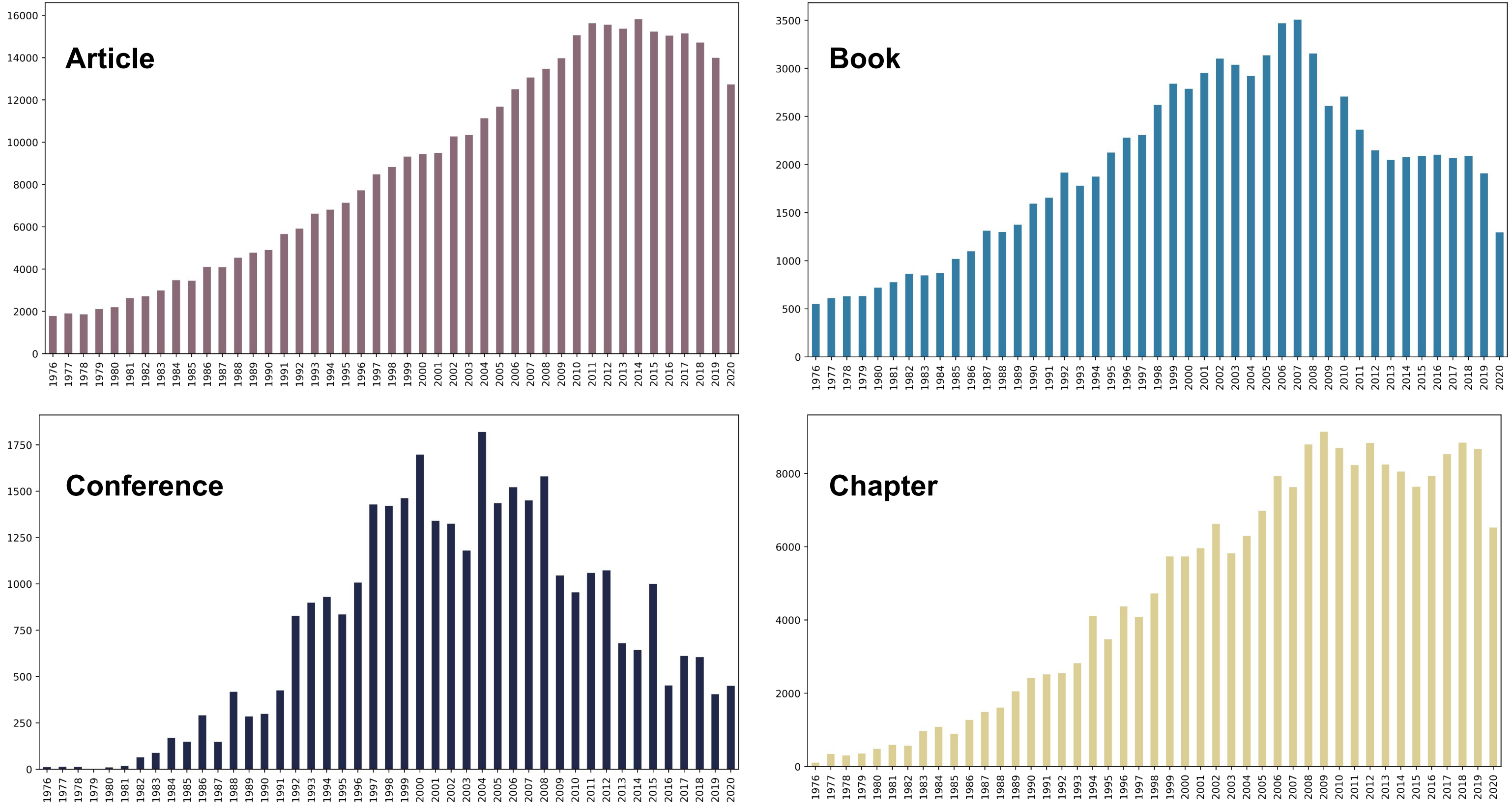






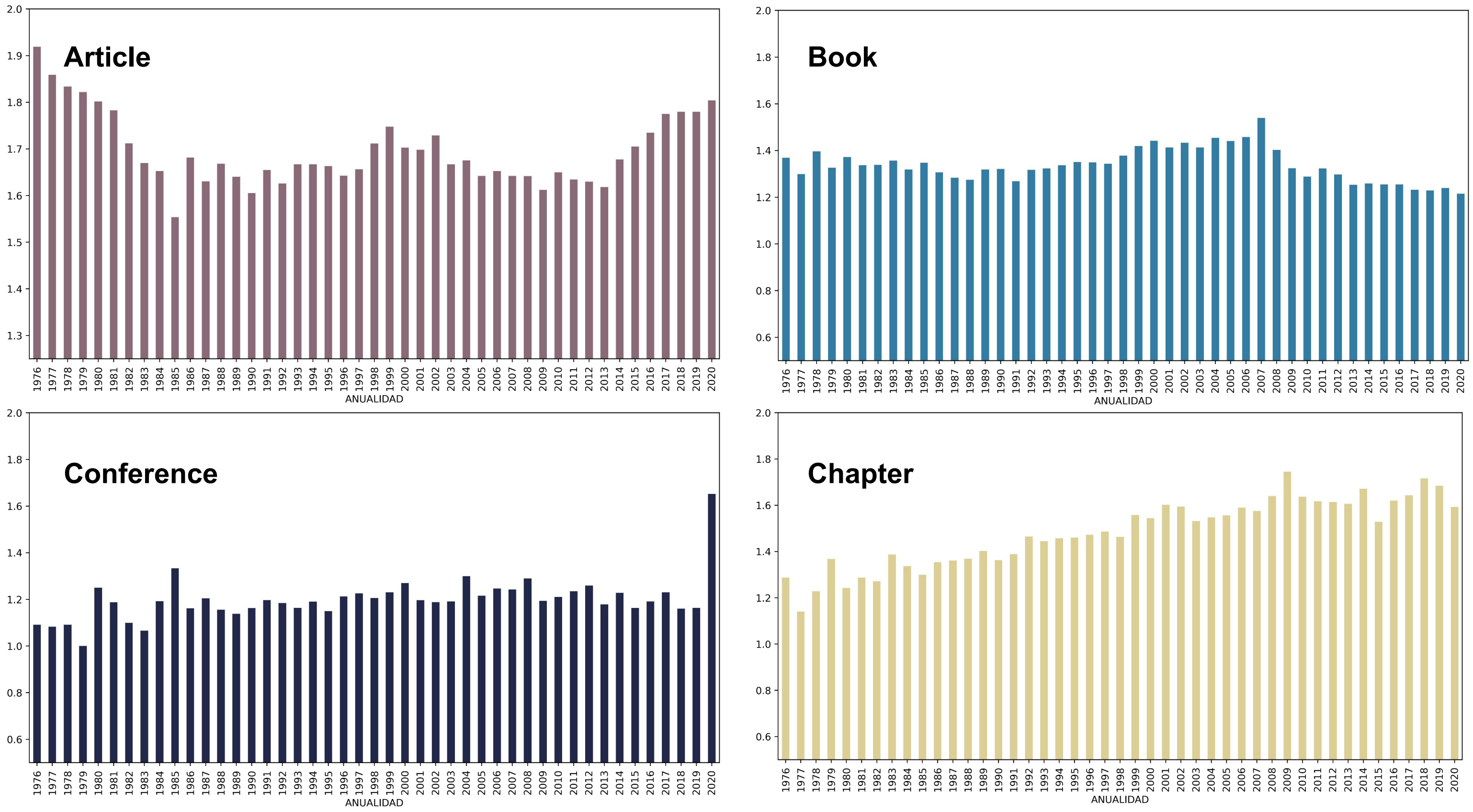
Dialnet + ORCID

Publication trends by document type (1976-2020)



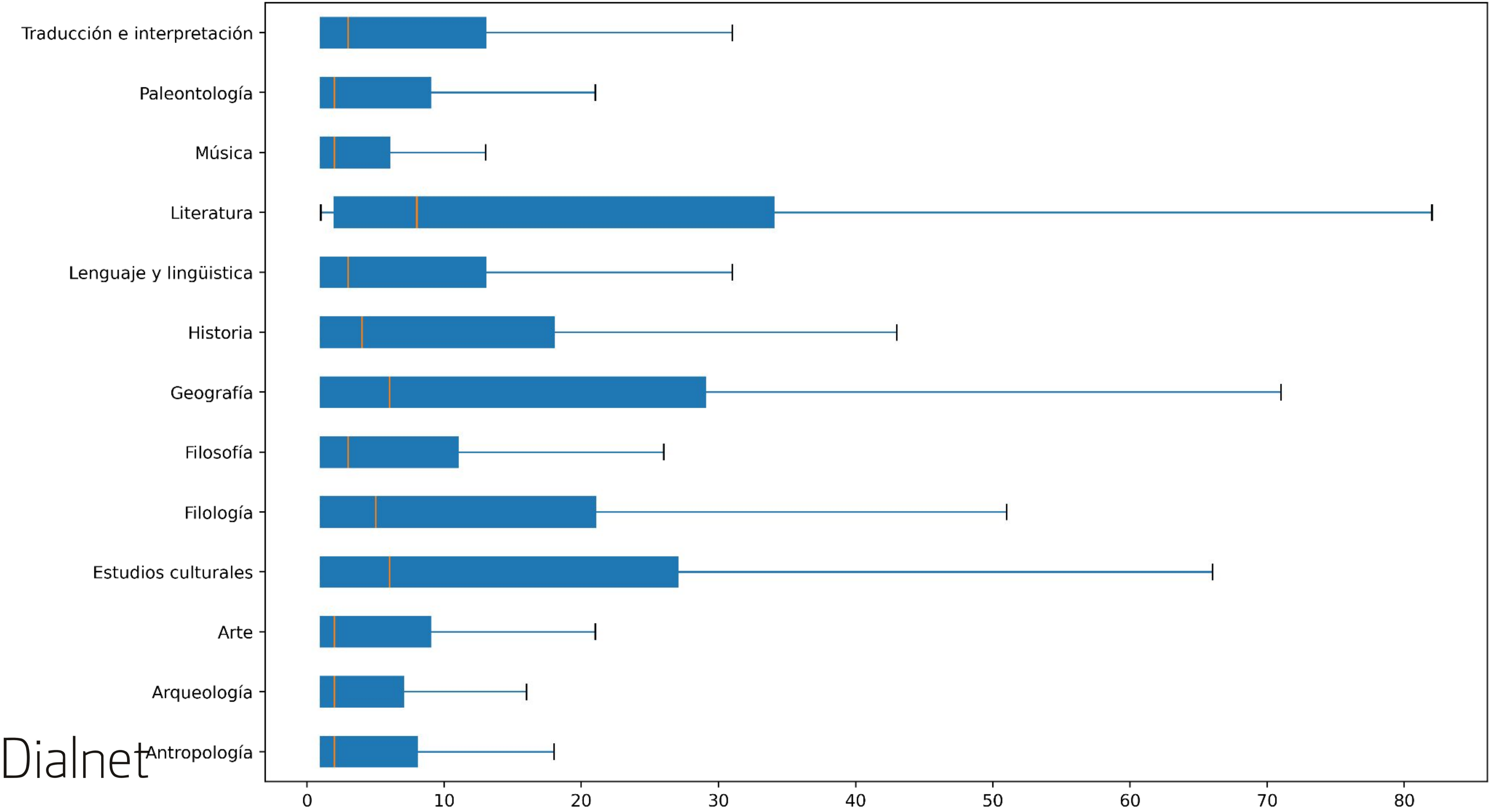
Dialnet

Yearly publication average by document type (1976-2020)



Dialnet

Average productivity by discipline



Dialnet

Limitations

- Humanists from very specific disciplines are not detected
- Publication typologies have to be normalized and reassigned
- Controlled vs non-controlled database
- Specific database limitations
  - Outdated or private ORCID records
  - Dialnet only covers local publications

Conclusions

- **Four literatures** are confirmed in Humanities!
- Research assessment is not limited to journal publications anymore…
- ...nor scientific publications
- Humanists are encouraged to make their research outputs more visible

CASE STUDY 2

Profiling humanists with Archetypal Analysis

Photo by Peter Gombos on Unsplash

Research questions

1. Can we identify different profiles of humanists based on their publication patterns?
2. How are humanists distributed among these profiles?
3. Do we observe differences by discipline?

Rationale

Phase 1

Select variables under analysis that will serve as profiling characteristics of individuals

Phase 2

Apply Archetypal Analysis to create profiles overall and by fields

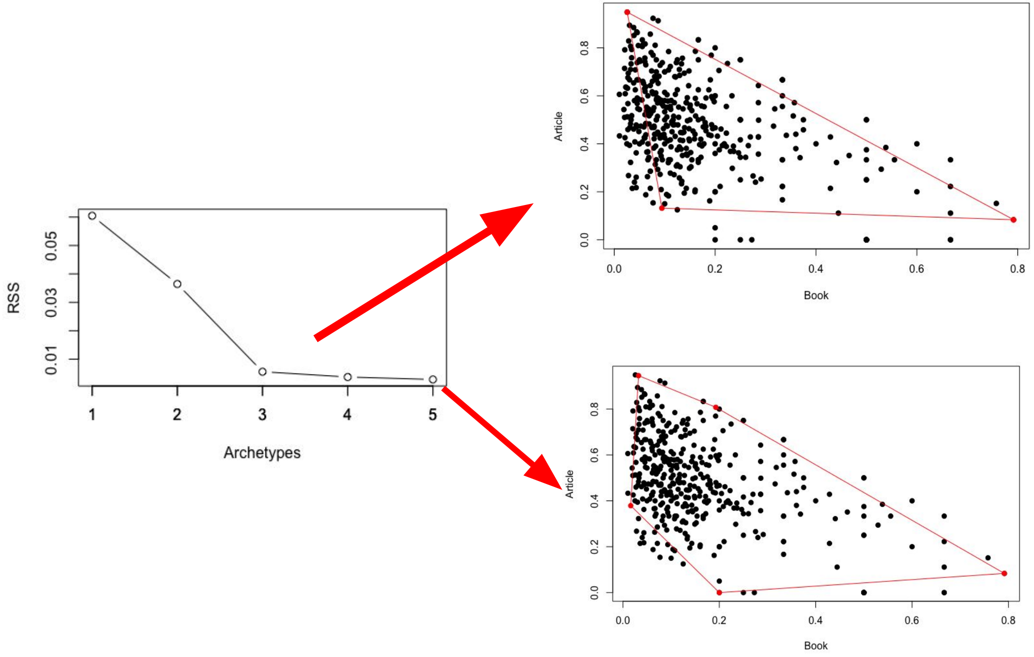
Phase 3

Analyze similarities of individuals to each profile

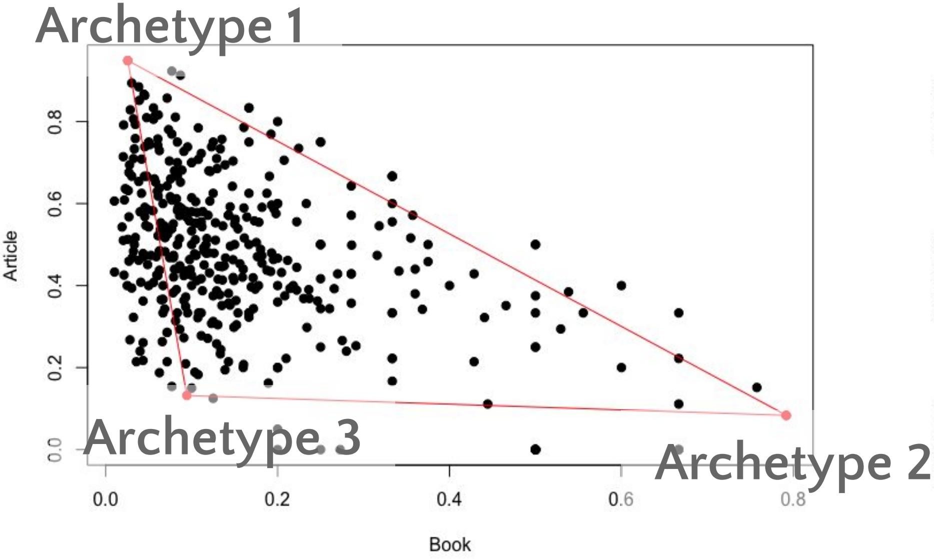
Variable	Definition	Source
Books	Share of edited and authored monographs	Dialnet; ORCID
Book chapters	Share or authored chapters	Dialnet; ORCID
Journal articles	Share of indexed and non-indexed journals publications	Dialnet; ORCID
Conference proceedings	Share of chapters identified as proceedings	Dialnet; ORCID
Non-scholarly docs.	Share of publications directed at non-scholars	ORCID
International output	Share of publications indexed in Scopus or Web of Science	Dialnet; ORCID; Scopus, WoS
Pubs	Total number of publications	Dialnet; ORCID

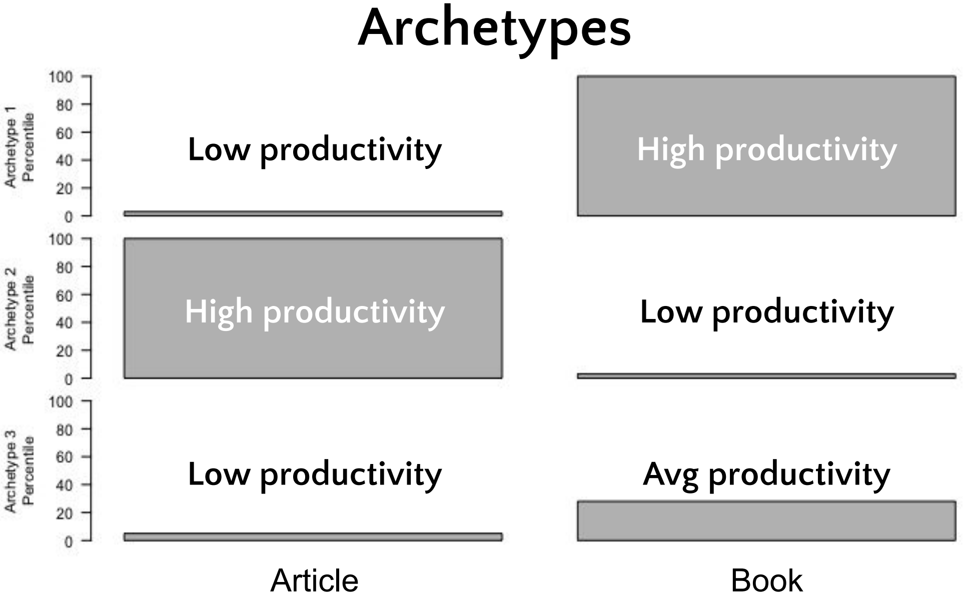
Archetypal Analysis

- Statistical data representation technique characterize multivariate data sets
- It identifies ***archetypes*** of individuals based on extreme combinations of two or more variables
- Individuals can then be characterized as ***pure or a mixture*** of archetypes

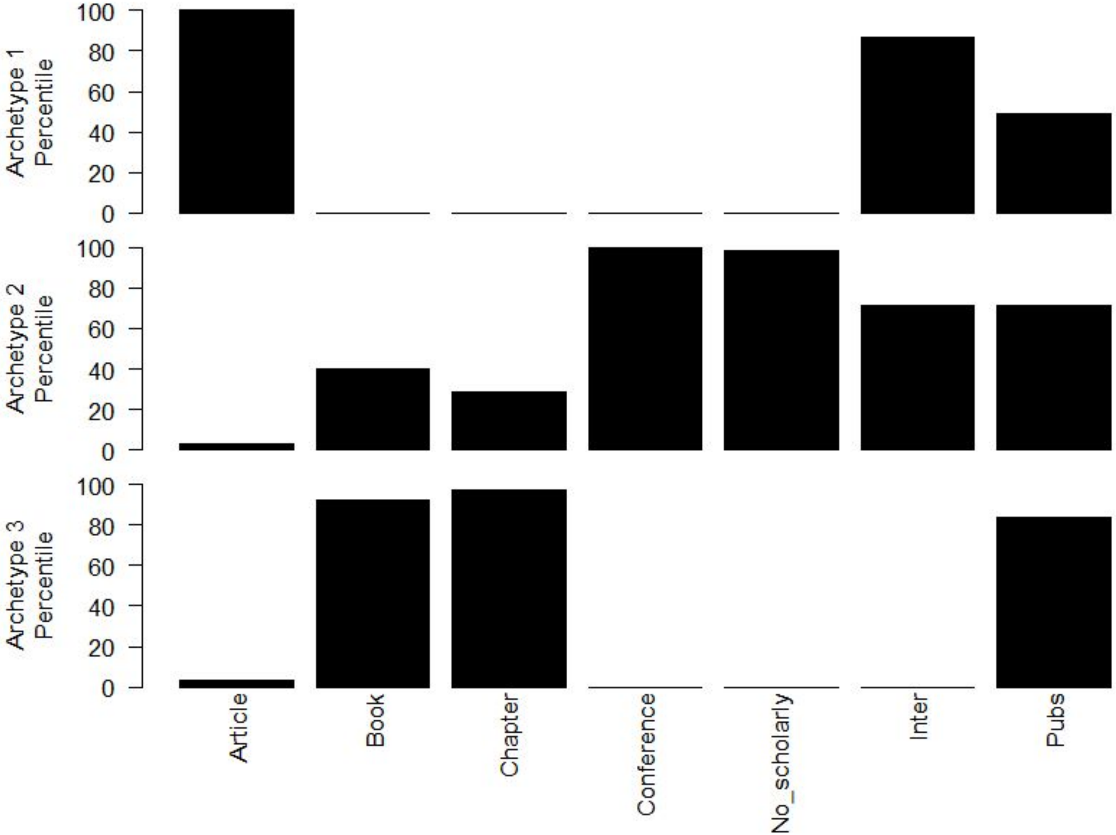


Archetypal Analysis

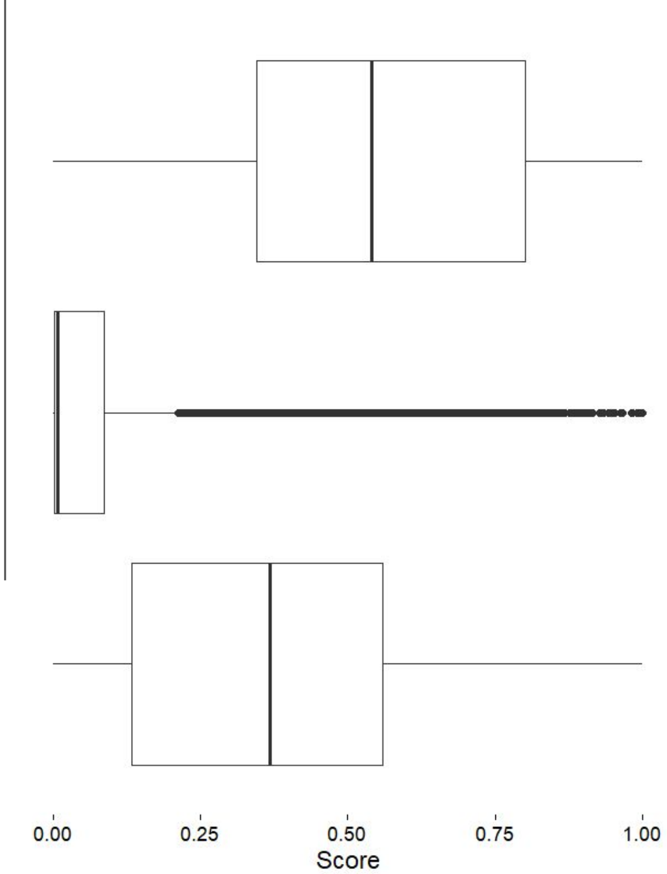




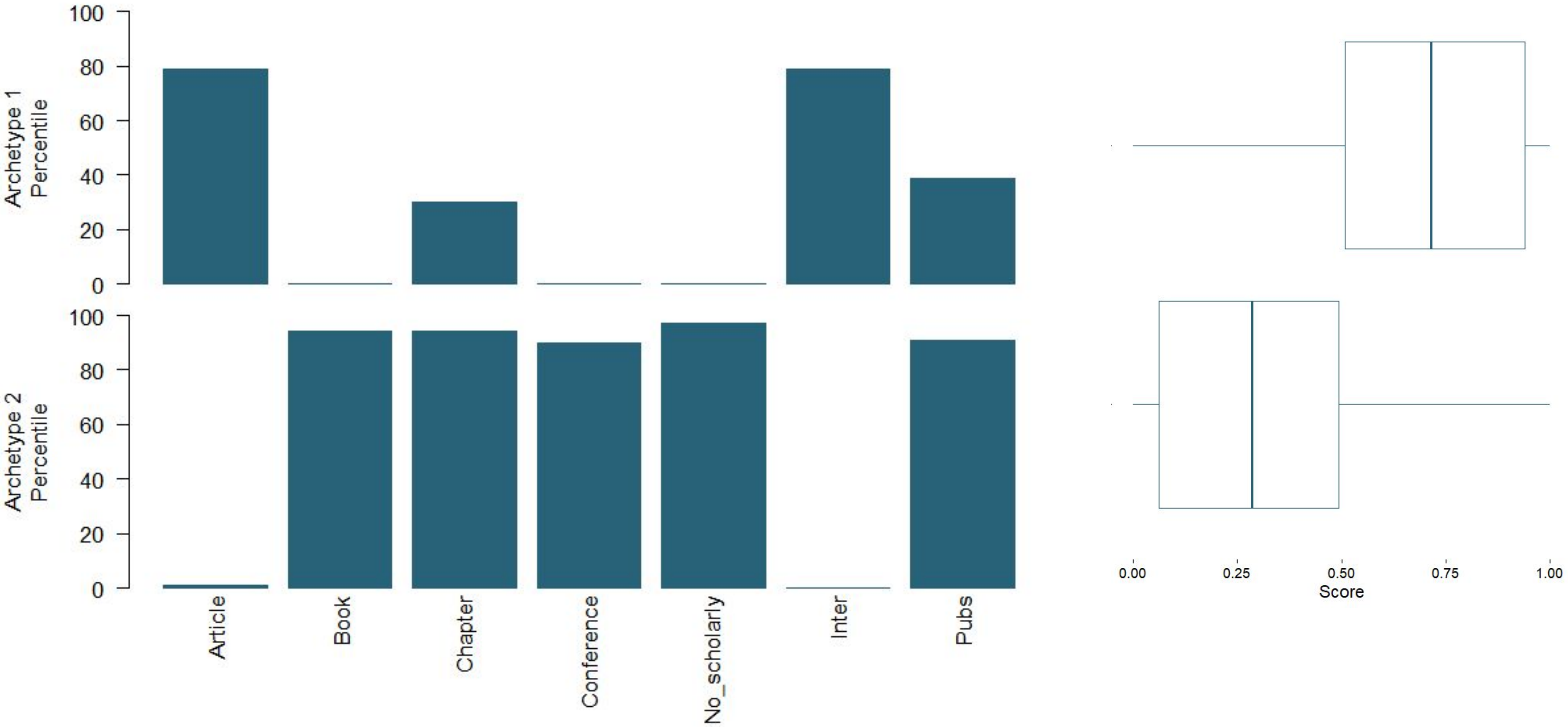
Characterization of the 3 profiles identified based on the average value of each of the selected variables



Distribution of humanists based on their similarity to each profile

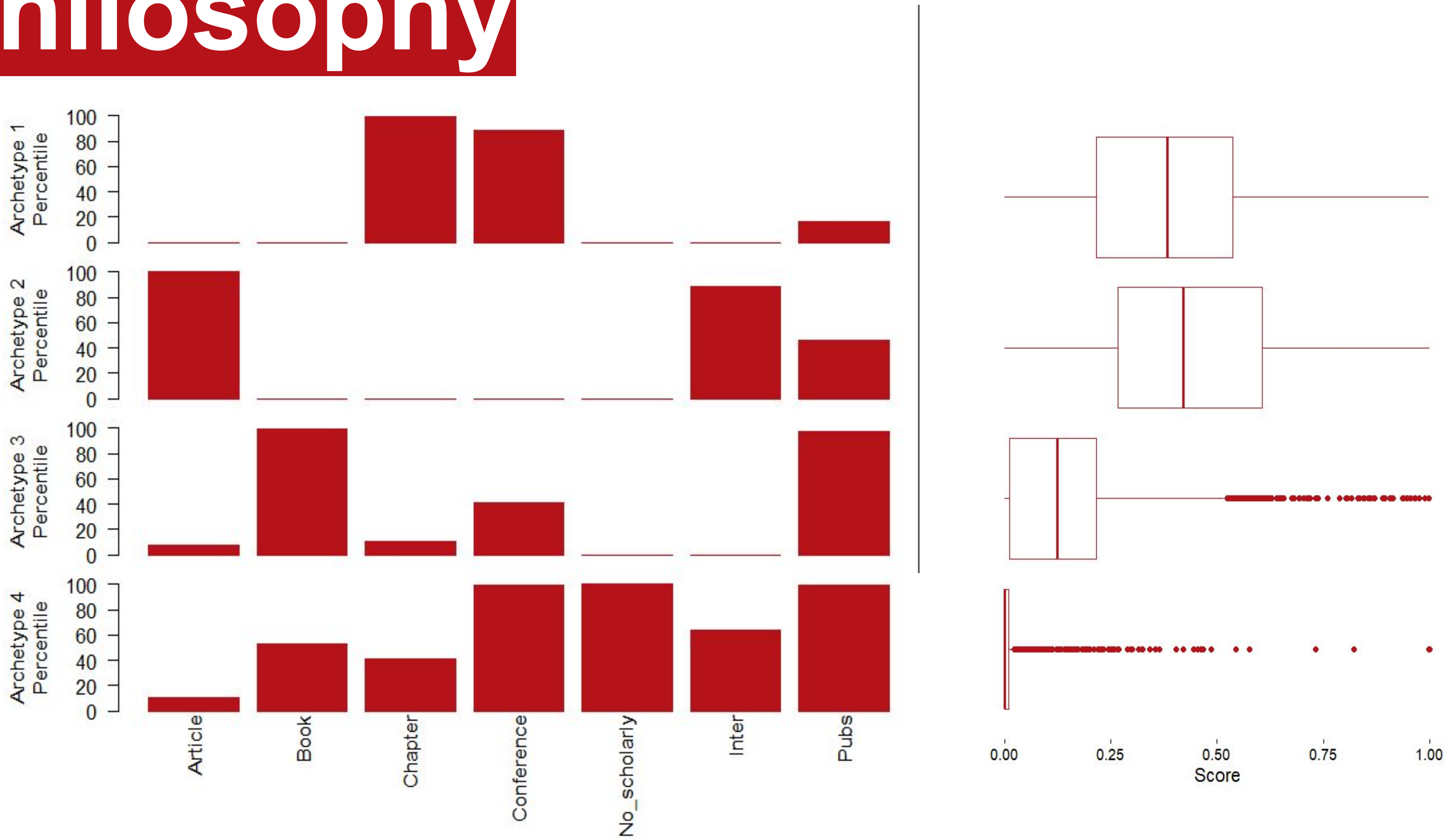


Philology

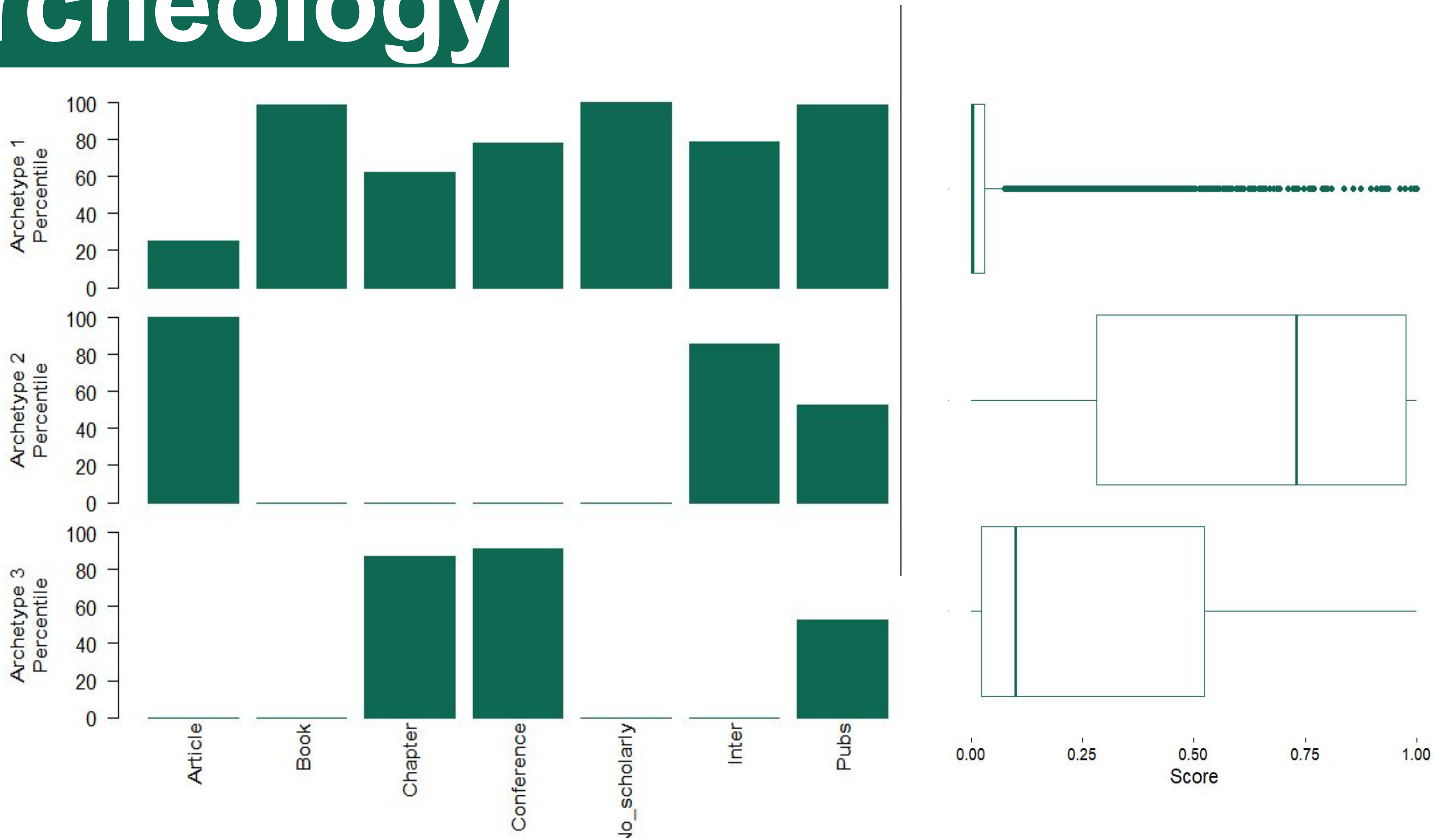


Archetype 1 Percentile
100
80
60
40
20
0
Archetype 2 Percentile
100
80
60
40
20
0
Article
Book
Chapter
Conference
No_scholarly
Inter
Pubs
0.00
0.25
0.50
0.75
1.00
Score

Philosophy


Archetype 1 Percentile
Archetype 2 Percentile
Archetype 3 Percentile
Archetype 4 Percentile
0
20
40
60
80
100
Article
Book
Chapter
Conference
No_scholarly
Inter
Pubs
0.00
0.25
0.50
0.75
1.00
Score

Archeology


Archetype 1
Percentile
Archetype 2
Percentile
Archetype 3
Percentile
100
80
60
40
20
0
Article
Book
Chapter
Conference
No_scholarly
Inter
Pubs
0.00
0.25
0.50
0.75
1.00
Score

Limitations

- Database limitations (as discussed in **Case Study 1**)
- What do we mean by non-scholarly?
- Even these data sources are biased towards what is easy to journal articles (identifiers, infrastructure)
- Validation interviewing experts in the field:

Do you see yourself reflected in those archetypes?

Are the variables selected the appropriate ones?

Conclusions

- Archetypal analysis helps us better understand differences on publication patterns between and within fields
- Research assessments must be adapted to scholars' production mode and not otherwise
- Impact must be defined operationally according to fields' core values

Conclusions

I think the training of future generations every year, I consider that far more important.

Biomedicine B

I am mostly doing outreach because I find it very relevant, but not everyone finds it very relevant.

Physics A

Robinson-Garcia, Nicolas, Rodrigo Costas, Tina Nane, and Thed N. van Leeuwen. 'Valuation Regimes in Academia: Researchers' Attitudes towards Their Diversity of Activities and Academic Performance'. SocArXiv, 10 November 2021. https://doi.org/10.31235/osf.io/ve7d3.

WE create our reality FROM our daily choices
The WORLD consists of billions of REALITIES

Towards a Science of Humanities

Photo by Elena Mozhvilo on Unsplash

Towards a research agenda

- Defining quality/impact/prestige in the Humanities
- Addressing multilingualism
- Sociological construction of knowledge (hot topics, influences, trends)
- Inter-generational differences
- …

Many thanks!

Questions, suggestions?

Nicolas Robinson-Garcia: elrobin@ugr.es
Wenceslao Arroyo-Machado: wences@ugr.es

DOI: 10.5281/zenodo.5793176