https://doi.org/10.51408/issi2025_151

The New Alliance. Bringing Together Bibliometric and Library Science for a Responsible Assessment of Research in SSH

Andrea Bonaccorsi

andrea.bonaccorsi@unipi.it
DESTEC, School of Engineering, University of Pisa, Largo Lucio Lazzarino 2, 56126 Pisa (Italy)

Abstract

We suggest a new alliance between two previously unrelated disciplines, namely Bibliometrics and Library science, with the goal of building a complete representation of the scientific production of humanities, including books and book chapters, as well as journal articles in a multilingualist perspective. We report on recent advancements on technology for interoperability of library resources that will permit an automatic validation of author identity via Authority Control. We discuss how this perspective will contribute to a fair and responsible research assessment for SSH; with particular attention to humanities.

Introduction

Among the opponents to research assessment a prominent role has been played, and is still played, by many scientific communities in humanities. Authors in fields such as history, literary criticism, or philosophy find the use of bibliometrics deeply unsatisfactory for their fields. In turn, while advocating for the use of peer review, as opposed to bibliometrics, they complain that the lack of accepted methodologies make research assessment procedures unreliable. This opinion is shared by some (not all) communities in social sciences. These arguments are well grounded.

This paper is a report on the main principles to design a new system to represent research in SSH, including some preliminary testing of the feasibility. It also includes some visionary ideas on how to use new data in order to do responsible research assessment in SSH.

The paper is organized as follows. The next section discusses the challenges to responsible research in SSH and the need for new bibliometric tools. The following section introduces the main principles and techniques to design a new data collection system. The following section introduces ideas on how to use new data in responsible evaluation of SSH. The final section concludes.

Humanities as the hidden science

While many of the issues about SSH are valid for both humanities and social sciences, they are more severe in humanities. Let us frame the discussion for humanities, and then discuss the role of social sciences at a later stage.

It has been known since long time that researchers in humanities follow a pattern of publication that differ from natural sciences (and partially differ from social sciences) (Hicks, 1999; Nederhof, 2006; Kulczycki et al. 2018). They have peculiar information needs and practices (Stone, 1982; Watson and Boone, 1994; Wiberley, 2009; Benardou et al. 2010).

Researchers in humanities publish significantly more books than their colleagues in STEM and assign to books a higher scientific value. Existing commercial datasets (e.g. Web of Science, Scopus) do not adequately cover books and, importantly, book chapters. These bibliographic forms (that include Festschriften and proceedings of major conferences) are crucial channels for scientific communication in humanities. In addition, by design they ignore the largest part of scientific production of researchers in humanities, which takes place in national languages in non-indexed journals (Harzing et al. 2016; Federation of Finnish Learned Societies et al., 2019; Visser et al. 2021; Martín-Martin et al. 2021; Petr et al. 2021).

In STEM research is more often than not published in English to guarantee a wider circulation of the content, greater accessibility across the discipline, better ranking in search results, less opaque indexing criteria. In contrast, the language of origin is of particular importance to Humanities as it has a closer and more significant relationship with the culture in which the research is rooted. Research in humanities can face substantial obstacles if it is to avoid marginalization, particularly in very specialized areas and in non-English language research (Tsakonas, 2024). The lack of books and book chapters and the limited coverage of journals in statistics of research mean that the overall representation of humanities is enormously undervalued (Linmans, 2009). This situation has prevented the test of alternative (even conflicting) theories of citation to humanities. We just know too little. There are good reasons to believe that the patterns of STEM do not apply to humanities (Ardanuy, 2013; Hellqvist, 2010; Engels et al. 2012; Waltman, 2016; dos Santos et al. 2021). For example, citations in books are structurally different from citations in articles: they constitute a longer list, which includes more heterogeneous sources, often from a variety of disciplines, over extended time periods (Cullars, 1989; 1998). Consequently, citation analysis must be completely redefined in the case of humanities, avoiding practices such as citation count, H-index, or Impact Factor, which are common (although contested) practice in STEM.

The poor representation of humanities in data collection has deep and negative consequences in the public visibility and the impact on policymaking. This state of the art is deeply unsatisfactory.

Data on humanities is desperately missing. Researchers in philology, history, philosophy, archaeology, or literary criticism and history of art almost never show up in official statistics and in public discussions on research. They are hidden. Data on their scientific production, in particular books, book chapters and journal articles in national languages, is never on the table.

Official statistics at UNESCO, OECD or European Union level simply ignore an entire region of researchers. Nor they appear in university rankings or in the top 1% most cited authors, or the top 2% worldwide scientists. Since humanities never appear in official statistics, in the public arena of democratic societies it can be said, provocatively that they do not properly exist.

In turn, the lack of data makes it acceptable that all efforts to carry out "research on research" or even to build an ambitious "science of science" simply ignore humanities and a good part of social sciences.

In the last three decades of academic work, the word most frequently associated to "humanities" has been "crisis" (Guillory, 1993; Donoghue, 2008; Rancière, 2009; Small, 2013), even "permanent crisis". The decline in public esteem, reduction in student enrollment, cut in public funding, lack of research positions, "adjunctification" of academic careers were the most cited phenomena, in US as well as in Europe. On top of these, there is clearly a lack of self-reflection on the epistemological and methodological grounds of research in humanities. This is in sharp contrast with the high status of natural sciences. The conventional argument is that humanities do differ from natural sciences on epistemological bases. Humanities deal with indexicality, subjectivity, judgment, while natural sciences deal with regularities, objectivity, and explanation. Natural sciences produce reliable knowledge, while humanities produce opinions, implying there is no chance to put them on a par. Consequently, we currently have a fully developed science of science (see for example Wang and Barabasi, 2021) that addresses natural sciences with an ambition to move into social sciences, but we have no comparable science of science in humanities. A simple example will clarify the urgent need for going beyond the state of the art: in the multi-author article (Liu et al., 2023) that summarizes two decades of high level research in the "science of science", presumably an authoritative reference for scientific communities and policy makers alike, the word "humanities" appear only once. For those that study the way in which science is produced, *humanities do not exist*.

In recent years there have been several proposals to address the situation, mainly by leveraging on open access publications (Colavizza et al. 2023) and making use of state-of-the-art technologies for citation mining and extraction (Sula and Miller, 2014; Peroni et al. 2016; Lent et al. 2018; Colavizza et al. 2018; 2019). In particular, the proposal of a Humanities Citation Index by Colavizza et al. (2023) is remarkable. While these works have made large progress in the state of the art, several gaps still remain. First, we need to extract citations not only from open access journals but also from traditional journals, as well as (most difficult) from books and book chapters. They still form a large core that cannot be ignored. Second, we need to address the issue of Author identity, with the final goal of reconstructing the entirety of production of researchers in humanities.

Main challenges

How can we enter into a responsible framework for the assessment of SSH research? Before advancing some solutions let us review some of the most intriguing and open issues.

Book and book chapters

A well known issue is the determination of the perimeter of the book production. Let us note that this problem has been solved since long time in bibliometrics using the technique of journal indexing. On the basis of a set of formal and published criteria the indexing organizations (Incites/Clarivate, SciVal/Elsevier or others) make a decision that defines the perimeter of analysis. On the basis of the perimeter, all kind of normalization and standardization practices can take place. *Without this technique, bibliometrics would not exist*. Now let us take note of the size of the problem with respect to books.

According to a 2023 press release from Scopus the total number of active peer-review journals is 27,950, of which 6,126 open access journals. Once the inclusion of a journal in the indexed perimeter is decided, the flow of articles (hence of authors) is automatically acquired.

It is interesting to note that the total number of books, which are not included in this flow of data, is at least one order of magnitude larger.

The organization holding the code number for books (ISBN) declares the number of books in its database as 42 million. But the definition of books and the practice of book recording from ISBN is controversial. According to an estimate by Google Books published in 2010 the number of books since the invention of print is 129,864,880. Given that UNESCO estimates that the total number of books per year is approximately 2,2 million, an updated number is 158,464,880 as for 2023. It seems that *the very definition of books is difficult to fix*, as it includes digital publishing, self-publishing, variations and repetitions that inflate the total number. A peculiar problem is also the size of Orphan books, or books for which the author is unknown or cannot be contacted. According to a study, the number of Orphan books is estimated in the order of 25 million (JISC 2009).

Given the uncertainties in the ontology of the object, as well as the large size of the universe, the goal of defining a perimeter of books, as it happens in standard bibliometrics for indexed journals, seems difficult to achieve. Shall we give up any hope? Perhaps no. Let us define the problem from a different attack point.

Authors

How many authors do exist in SSH? This question might be more manageable than the question on the number of books. To address this issue we might start with some order of magnitude from existing sources, searching for the total number of publications.

The AI-backed Dimensions, searched under the heading Human society, delivers 5,629,704 publications, of which 722,352 are book chapters, 80,917 are monographs, and 46,713 edited books. Another query on Language, Communication and Culture delivers 3,109,899 publications, of which 528,027 book chapters, 59,693 monographs, and 34,985 edited books.

Another fast-and-furious query on Open Aire delivers the following numbers: Humanities and the Arts 1,645,280 publications, Education 1,113, 256; History and archeology 721,008; Languages and literature 498,166; Philosophy, ethics and religion 316,847; Arts 111,774.

ERIH PLUS, the European Reference Index for the Humanities and Social Sciences, supported by the Norwegian Directorate for Higher Education and Skills includes more than 10,000 journals in SSH, while the number of individual authors is not declared.

The European Alliance for Social Sciences and Humanities include organizations in which 100,000 active researchers publish in SSH.

We do not know which is the share of SSH authors on the total at world level. However, there are some estimates on the total number of researchers at world level, based on UNESCO, OECD, European Commission and national data (in particular, US data). These estimate converge around a baseline number of 8 million currently active researchers worldwide (Burke et al. 2021; Ayan, Hakk and Ginther 2023). On this basis it is realistic to assume that SSH active researchers should be in the range between one and two million. If these are the numbers, the goal of building up a publicly available census of SSH authors is not out of reach, given the level of AI technologies available.

A plausible strategy might be the following. First, collect all national repositories that include only SSH authors whose scientific activity is validated. According to the survey by Sile et al. (2018) there are several European countries in which such repositories are publicly available (Kulczycki et al. 2018; 2020). This collection might create the backbone of the exercise.

Second, download authors from publicly available datasets, including Dimensions, OpenAire, Open Citations, and various repositories of open access journals. Several repositories of Open Access journals are available. The OpenDOAR directory (https://v2.sherpa.ac.uk/opendoar) already makes it possible to access to thousands of repositories across all countries. Regional federations of repositories, for example in Latin America, aggregate national and institutional repositories (e.g. https://www.lareferencia.info/es and https://www.redalyc.org/). The Directory of Open Access Books (www.doabooks.org) gives access to >80.000 books.

Third, compile an integrated list of authors with the associated metadata by integrating all publicly available sources.

Will this list be valid? Of course no. Further work should be done for the validation of authors. This problem is *not the same of disambiguation of authors in scientific journals*. The definition of author in scientific journals is very simple: any person that submits an article and gets published is ipso facto an author. The definition of the perimeter of indexed journals solves once and for all the issue of who is an author. What is left to journal publishers is the problem of disambiguation of journal authors, an issue which is largely solved by the mandatory inclusion of ORCID ID.

The largest author identification system is ORCID. The number of active records is 9,2 million in 2024, used in the same year by 2.3 billion external items (ORCID 2024) . ORCID is designed for the need of the research community and the publishing industry as a general purpose tool to reduce or mitigate the well known issue of name disambiguation. It has become the general standard, as a large number of journal editors, publishers and evaluation agencies started to ask the ORCID ID as a mandatory information for authors.

This is not the same for books and book chapters, since not all authors of books have an ORCID ID and not all authors qualify as authors of a scientific publication. Scientific publications are a subset, often a small one, of book publishing. In addition, the integration of repositories will create issues of duplication and disambiguation. In the absence of a mandatory ORCID ID procedure, we must find another solution.

Author validation

Is there a way to establish the identity of authors without a mandatory code such as ORCID? My suggestion is that an alternative is available *in a domain of expertise that has been traditionally separated from bibliometrics*, i.e. Library science, or Information science. After extensive study of the problem and consultation with key actors, it is possible to conclude that the key is the integration between the world of libraries and the world of bibliometric datasets. There is no other way to integrate book and journal metadata in order to build up a complete representation of the scientific production of scholars in humanities. *This has never been done before.* It is a truly new alliance.

With this approach an original combination of two disciplines, previously separated, will be achieved. Library science has developed accurate methods for the disambiguation and validation of authorship, but has no interest for the aggregation of data; on the contrary bibliometrics has constructed a large array of indicators but no adequate coverage of books and book chapters, as well as of non-indexed journals, which are extremely relevant in humanities.

In this scientific and intellectual domain the issue of how to achieve a unique author identification has been crucial for decades. One can say that among the distinguished skills of authors and practitioners in libraries the correct identification of authors has traditionally been prominent, together with the methods of cataloguing.

Libraries have a robust and well tested method for the unique identification of authors, called *Authority Control*. It is defined as follows (Clack 1990, 1): "Authority control is a technical process executed on a library catalog to provide structure. Uniqueness, standardization, and linkages are the foundation of authority control".

Authority control of a library catalog is maintained through an authority file that contains the terms used as access points in the catalog. The access points that determine the structure of the catalog may be real entry headings on bibliographic records or cross references. In library catalogs the entry headings under control generally consist of personal and corporate names, uniform titles, series, and subjects.

Libraries have developed Authority Files by using over time various generations of standards and software solutions. Historically, the main problem has been the lack of interoperability of definitions and software tools. The problems are under way of solution through collaborative projects such as Share VDE (https://wiki.share-vde.org/wiki/Main_Page). This is an international library driven initiative that adopts the entity-oriented bibliographic data model BIBFRAME proposed by the US Library of Congress and the Library Reference Model defined by IFLA with the goal of making accessible bibliographic records in the Linked Open Data format (Angjeli et al. 2014; Bennett et al. 2017; Koskas, 2022; Bianchini and Sardo, 2022). Within the Share VDE project several national libraries and university libraries are currently collaborating for bringing into practice a new level of cooperation based on interoperability and openness to sustain discovery of knowledge (Possemato, 2022).

Among them it is important to mention the US Library of Congress, which has the largest global collection of books in all fields. All living authors who have published at least one book are registered.

An important implication of this collaborative effort is that it is possible (and financially plausible) *to design a software procedure for the automatic control of the Authority File* in any language and for any name of author, managing all cases of ambiguity. This is the first foundation block of the new alliance, creating a linkage between bibliometrics and library science.

Contrary to the bibliometrics based on journal indexing, the new bibliometrics will be centered around authors, whatever the entry point in the data collection system.

Citations and abstracts

At this point we might have collected an official census of authors in SSH associated with the metadata regarding books, book chapters, and articles.

The next step is to extract citation data. This exercise is largely practiced in journals but almost unknown for books and book chapters. The are two reasons: (a) citations appear in books with a variety of formats, that are not standardized (e.g. full reference in the text, full reference in the footnote, author and date in the text etc.); (b) citations include many errors, since they are self-made by authors, with limited room for an automatic control by book editors and publishers (particularly in the absence of a DOI number).

This problem is nowadays largely solvable with dedicated software that is able to automatically recognize the textual entity within the text, *using AI techniques such as Named Entity Recognition (NER)* and its more recent developments. More difficult is the problem of errors, for which limited experience is available so far. Given these hard problems, how do we address the issue of citations from books and book chapters?

The Initiative for Open Citations (www.i4oc.org) and the Initiative for Open Abstracts (www.i4oa.org) have asked publishers to deliver citations and abstracts to CrossRef, together their metadata for indexing purposes, with mixed success. It is our contention that some of the existing institutions or publishers will in the near future develop a full scale initiative to extract automatic citation data and abstract data without infringing the copyrights of publishers. There is a huge value in this enterprise, the cost of which is currently largely reduced after the advent of Large Language Models.

Let us continue my suggestions in a scenario in which fully validated citation data will be available for all authors in SSH, both citations to other works (including books) and citation from other works (including books). Abstracts will also be available in this scenario.

Acknowledgments

Let me add another desideratum. Once the software solutions for the extraction of metadata has been put in place, another opportunity will be available. Most books include a section, usually in the initial chapters (e.g. Preface, Introduction, Foreword and the like), in which authors offer a list of names of colleagues and friends who are thanked for their collaboration with the work. While the literary style of the list is usually variable, from rigidly professional to personal and informal, the list of names offers a rich source of information. We anticipate a *new bibliometrics based on acknowledgments.*

Deceased authors

One intriguing issue in the structure of citations in books and book chapters is the *large share of citations to deceased authors.* This practice is largely different from the one in STEM, in which the life of citations is largely skewed towards recent authors (with a higher probability of citing living authors).

Citing deceased authors is a crucial practice for SSH, particularly in humanities, since the very object of study is located in the past. The epistemological role of these citations in humanities should not be underestimated (Grafton, 1999).

This however creates a serious bibliometric problem, since the computation of any citation index will be largely biased by unobserved differences in the share of deceased authors in the reference list.

Nor the problem can be addressed by discriminating the authors in the reference list using some author ID, for example ORCID. The problem with ORCID is that it is based on the principle of individual control, i.e. only authors themselves can apply for an ID and update or modify the information associated to the identity. This means that it will not be possible to build up an ORCID number for deceased authors. If we ask the FAQ system of ORCID about the ID of deceased authors the reply is the following: "Is it possible to register an ORCID iD for a deceased person?" "No. Our policy is that an ORCID iD can only be created by the individual themselves, not by any other person. This is because a core principle of ORCID is individual control. You may wish to contact ISNI (International Standard Name Identifier), as their mission is "to assign to the public name(s) of a researcher, inventor, writer, artist, performer, publisher, etc. a persistent unique identifying number"; they take a library authority approach to this, rather than a researcher-controlled one as we do".

ISNI, in turn, has 16.1 million identities for 14.3 million individual persons, of which 1.2 million are researchers (a significantly lower number than ORCID). ISNI keeps a record of deceased authors, but fails to disambiguate correctly. If we look for the record of Michael Polanyi, ISNI does not recognize that the author of *Science, faith and society* (Polanyi, 1946) is the same author of *The logic of liberty* (Polanyi, 1951). We therefore cannot rely, for different reasons, neither on ORCID nor on ISNI, irrespective of their respective values and contributions.

Within the proposed new alliance it is possible to refer again to the Authority Control methodology. Authority Files, as opposed to ORCID files, include the dates of publication of all works by the same author, *even if he/she deceased*. As opposed to ISNI identities, there are no errors or ambiguity. Using some conventions on the latest dates of publication we might identify deceased authors with reasonable approximation.

An automatic procedure might therefore *classify all citations in two categories of active vs. deceased authors* and calculate the citation indexes separately. The classification might be updated dynamically at regular intervals to take into account changes in the proportion between the two categories.

Academic publishers

While the Authority Control made possible by the library system will eliminate ambiguity on author identities, it will not per se discriminate with respect to the scientific content of the publication. This issue might be complicated by the circumstance that many academic authors do publish academic works alongside popular science publications, or collection of newspaper articles and book reviews.
While the general issue might remain controversial, a practical solution might be to refer to the list of academic publishers established by the Spanish CSIC (Gimenez-Toledo et al. 2019).

Affiliation

This information will be generally available in the metadata from journals and books, but several problems must be addressed. A combination of methods should be used here: official registers and AI.
First, it is extremely likely that the metadata will include definitions of the affiliation that are not standardized. It will be possible to use the available standard definitions of affiliations, such as ISNI (www.isni.org), ETER (European Tertiary Education Register) for European higher education institutions (https://eter-project.com/) and the ORGREG register for Public Research Organizations (https://www.risis2.eu/registers-orgreg/). Non-European affiliations will be checked against UNESCO datasets (https://www.whed.net/home.php).
Second, it is possible that in some cases the metadata on affiliation will be missing. In this case an AI-backed procedure will search for affiliation data of the identified author associated to dates and might produce an estimate of the affiliation for the missing publication.

The strenght of the new alliance

The strenght of the proposal lies in the alliance between bibliometrics and library science. The automatic validation of authors using Authority Files will ensure that all data, whatever the source of collection, will land into a validated database.
In turn, the classification of cited authors by age (in particular, the discrimination between living or recent authors and deceased authors) *will allow the deployment of the bibliometric toolbox* with respect to standardization and normalization of data.
This will create an incentive for publishers to deliver their metadata (including citations and abstracts) on a regular basis, in order to fill the census with their own data. Remaining outside the platform will be too costly. The idea needs someone who makes the initial investment and opens the way.

From the new alliance to responsible research assessment

The new alliance between bibliometrics and library science might deliver solutions that made it possible to improve the quality of research assessment and address the issues of transparency, diversity and fairness. Let us articulate this proposition.

It is fair to say that the dominant methodology for research assessment in SSH is the peer review. We know from a large literature, however, that peer review is not the golden age of research assessment. It has its own methodological weaknesses and is subject to biases of various types.

We need to go beyond the notion of informed peer review, whereby the individual peer review is assisted by a few simple bibliometric data such as citation count or citation weight. Let us consider a scenario in which quantitative bibliometrics will deliver qualitative insights that support and complement human evaluation.

In other words it is possible to anticipate a scenario in which

- a census of validated authors in SSH is established
- for each of the works of validated authors we have metadata
- metadata include citations, ackowledgments and abstracts
- data is available based on formats that allow large scale processing.

In this scenario we might give full justice to the humanities by addressing, first of all, the controversial issue of productivity. It is often assumed that research in humanities is less cumulative and less convergent than in STEM, hence less productive (Cole 1983; 1994; Clauset et al. 2015). The issue is controversial (Hedges 1987; Fawcett and Higginson 2012; Fanelli and Glänzel, 2013). A few years ago *Nature* made the claim that humanities, or soft science, should be preserved and protected (Nature 2015), but the issue of relative productivity has never been addressed systematically.

Are researchers in humanities less productive? No analysis of productivity can be done without the definition of the perimeter of the overall scientific production. To the extent that data collection is successful we might address several open (and contested) issues. Does the scientific production of researchers in humanities follow the same skewed distribution that we find in natural sciences? Is it subject to the Matthew effect? Does it decline with age or academic age? Is it associated to academic position, affiliation, type of institution? What is the typical life cycle of scientific production? On all these issues the current evidence is limited and scattered. Recent research has shown that researchers in humanities do not differ from STEM in the shape and asymmetry of the distribution of scientific production, following the so called Matthew effect (Bonaccorsi et al. 2017). A related issue is whether researchers in humanities adopt team production and authorship. Are researchers adopting the team-based inquiry approach of their colleagues in STEM? In which disciplines do we find a larger average (and median) number of co-authors? Does the size of team has an influence on the degree of novelty produced (Wu et al. 2019)?

In this scenario a whole range of Natural Language Processing techniques can be introduced, tested and validated as a support to human judgment. They might be a powerful support to responsible assessment of research. They include embeddings and variable length embeddings, network dynamics, knowledge graph, sentiment analysis, citation networks, citation clustering and many others (Chen et al. 2009; Guevara et al. 2016; Kozlovski et al 2018; Chinazzi et al. 2019; Tshitoyan et al. 2019; Miao et al. 2022; Peng et al. 2021).

Scientific texts are an optimal field for data analysis, because researchers speak a controlled language that is, by design, aimed at being critically evaluated. Recent technologies in NLP and pre-trained LLM systems allow a fine-grained analysis of the content of scientific publications, with unprecedented sophistication.

Thus for each of the main (and controversial) issues in the epistemology of humanities it will be possible implement one or more AI-based technique: word embeddings to examine the novelty of knowledge produced by humanities; Knowledge Graphs to examine the explanatory nature of statements and the cause-effect relations; Topic Modeling and citation clustering to study the formation of scientific consensus, the persistence of paradigmatic pluralism and the management of controversies; citation networks and field tracking to investigate into the cumulativeness of knowledge; again embeddings, but also information density and linguistic complexity to explore the level of interdisciplinarity.

Topic modeling (as in Bonaccorsi et al. 2022) and word embeddings (as in Melluso et al. 2024a; 2024b) might be applied to the collection of books and articles described above. Recently developed methodologies in the full text processing of publications, such as information density (Bernstein, 1964; Bischhof and Eppler, 2010; Evans and Aceves, 2016; Hamilton et al. 2016; Aceves and Evans, 2023) and linguistic complexity (Lu et al. 2019a; 2019b) allow a granular analysis of the structure of argumentation. The extent to which they can be replicated on abstracts is to be explored.

Conclusions

This paper suggests a new alliance between bibliometrics and library science in order to build up a responsible assessment for SSH. The evaluation of research in these fields requires the full scale consideration of books, book chapters, and journal articles in a multilingualist perspective.

My proposal is complementary to the institutional efforts, undertaken by the European Union, to establish Open Science, through the creation of a European Quality Standard for Institutional Open Access Publishing (EQSIP) (e.g. https://diamasproject.eu), the technical improvements of open journal platforms for the Diamond OA (www.craft-oa.eu) and the exploration of open metric data such as OpenCitations (https://opencitations.net) and Scholexplorer (https://scholexplorer.openaire.eu). With respect to these efforts one of the major limitations is that books and book chapters have very limited coverage in open access and even, as addressed by the Palomera project (https://operas-eu.org/projects/palomera/) in open access funding.

This paper argues that the technological resources to undertake the enterprise of a new alliance are available.

References

Aceves, P., Evans, J.A. (2023) Human languages with greater information density increase communication speed, but decrease conversation breadth. Pre-print.

Angjeli, A., MacEwan, A., Boulet, V. (2014) ISNT and VIAF. Transforming ways of trustfully consolidating identities. IFLA WLIC 2014 Conference. Lyon, 1-19.

Ardanuy, J (2013) Sixty years of citation analysis studies in the humanities (1951–2010). Journal of the American Society of Information Science and Technology, 64(8), 1751–1755.

Ayan, D.E., Hakk, L.L., Ginther, D.K. (2023) How many people in the world do research and development? Global Policy. DOI: 10.1111/1758-5899.13182

Benardou, A., Constantopoulos, P., Dallas, C., Gavrilis, D. (2010) Understanding the information requirements of arts and humanities scholarship. International Journal of Digital Curation 5(1), 18–33.

Bennett, R., Helgel-Dittrich, C., O'Neill, E.T.O., Tillett, B.B. (2017) VIAF (Virtual International Authority File): Linking the Deutche Nationalbibliothek and Library of Congress Name Authority Files. International Cataloguing and Bibliographic Control, XXXVI, 1, 12-18.

Bernstein, B. (1964) Elaborated and restricted codes. Their social origins and some consequences. American Anthropologist, 66, 55-69.

Bianchini, C., Sardo, L. (2022) Wikidata: A new perspective towards universal bibliographic control. In G. Bergamin and M. Guerrini (eds.) Bibliographic control in the digital ecosystem. Florence, AIB-EUM-Firenze University Press.

Bischhof, N., Eppler, M.J. (2010) Clarity in knowledge communication. Proceedings of the Tenth International Knowledge Management Conference Iknow, vol. 10, 162-174.

Bonaccorsi A. (2018) Towards an Epistemic Approach to Evaluation in SSH. In Bonaccorsi A. (ed.) (2018) The evaluation of research in Social Sciences and Humanities. Lessons from the Italian experience. New York, Springer International Publishing.

Bonaccorsi A., Daraio C., Fantoni S., Folli V., Leonetti M., Ruocco G. (2017) Do Social Sciences and Humanities behave like life and hard sciences? Scientometrics, 112, 607-653.

Bonaccorsi, A. (2023) An epistemic approach to research assessment in the social sciences. In Tim C.E. Engels, Emanuel Kulczycki (eds.) Handbook on research assessment in the Social Sciences. Cheltenham, Edward Elgar.

Bonaccorsi, A. (2025) The knowledge of humanities. A comparative epistemology of historiography, literary criticism, history of art, and history of architecture. Turnhout, Brepols (forthcoming).

Bonaccorsi, A., Melluso, N., Massucci, A. (2022) Exploring the antecedents of interdisciplinarity at the European Research Council: a topic modeling approach. Scientometrics. https://doi.org/10.1007/s11192-022-04368-9

Burke A, Finamore J, Foley D, Jankowski J, Moris F; National Center for Science and Engineering Statistics (NCSES) (2021). Measuring R&D Workers Using NCSES Statistics. NSF 21-335. Alexandria, VA: National Science Foundation. Available at https://ncses.nsf.gov/pubs/nsf21335/.

Chen, C., Chen, Y., Hou, H., Liu, Z., Pellegrino, D. (2009) Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics, 3, 191-209.

Chinazzi, M., Gonçalves, B., Zhang, Q., Vespignani, A. (2019) Mapping the physics research space. A machine learning approach. APJ Data Science, 8, 33.

Clack, D.H. (1990) Authority Control: Principles, Applications, and Instructions. Chicago, American Library Association.

Clauset A., Arbersman, S., Larremore, D.B. (2015) Systematic inequality and hierarchy in faculty hiring networks. Science Advances 1, e1400005.1

Colavizza, G., Peroni, S., Romanello, M. (2023) The case for the Humanities Citation Index (HuCI): A citation index by the humanities, for the humanities. International Journal on Digital Libraries, 24, 191-204.

Colavizza, G., Romanello, M., Babetto, M., Barbay, V., Bolli L., Ferronato, S., Kaplan, F. (2018) Linked Books: Towards A Collaborative Citation Index for the Arts and Humanities. Proceedings of the Red de Humanidades Digitales, A. C., Mexico City, 178–181.

Colavizza, G., Romanello, M. (2019) Citation mining of humanities journals: the progress to date and the challenges ahead. Journal of European Periodical Studies, 4, 36–53.

Cole, S. (1983) The hierarchy of the sciences? American Journal of Sociology, 89, 111-139.

Cole, S. (1994) Why sociology doesn't make progress like the natural sciences. Sociological Forum, 9, 133-154.

Cullars, J. M. (1989). Citation characteristics of French and German literary monographs. Library Quarterly, 59 (305–325).

Cullars, J. M. (1998). Citation characteristics of English-language monographs in philosophy. Library and Information Science Research, 20(1), 41–68.

Donoghue, F. (2008) The last professors. The corporate university and the fate of the humanities. New York, Fordham University Press.

dos Santos, E.A., Peroni, S., Mucheroni, M.L.(2021) Citing and referencing habits in medicine and social sciences journals in 2019. Journal of Documentation. 77(6), 1321–1342.

Engels, T. C., Ossenblok, T. L., & Spruyt, E. H. (2012). Changing publication patterns in the social sciences and humanities, 2000–2009. Scientometrics, 93(2), 373–390.

Evans, J.A, Aceves, P. (2016) Machine translation. Mining text for social theory. Annual Review of Sociology, 42(1), 21-50.

Fanelli, D., Glänzel, W. (2013) Bibliometric evidence for a hierarchy of sciences. PLoS ONE 8, e66938.

Fawcett, T.W., Higginson, A.D. (2012) Heavy use of equations impedes communication among biologists. PNAS, 109, 11735-11779.

Federation Of Finnish Learned Societies, Information, T.C.F.P., Publishing, T.F.A.F.S., Universities Norway, ENRESSH (2019) Helsinki initiative on multilingualism in scholarly communication. Figshare.

Fodor, J.A. (1974) Special sciences (or: the disunity of science as a working hypothesis). Synthese, 28(2), 97-115.

Fortunato, S., Bergstrom, C.T., Borner, K., Evans, J.A., Helbing, D., Milojevic, S., Petersen, A.M. et al. (2018) Science of science. Science, 359(6379), eaao0185.

Giménez-Toledo, E. Sivertsen, G., Mañana-Rodriguez, J. (2019) International Register of Academic Book Publishers (IRAP): overview, current state and future challenges. In Daraio et al. (eds.) Proceedings of the S&T Indicators Conference. Rome, Efesto Publishers.

Grafton, A. (1999) The footnote: A curious history. Cambridge, Mass. Harvard University Press.

Guevara, M.R., Hartman, D., Aristarán, M., Mendoza, M., Hidalgo, C.A. (2016) The research space: Using career paths to predict the evolution of the research output of individuals, institutions, and nations. Scientometrics, 109, 1695-1709.

Guillory, J. (1993) Cultural capital. The problem of literary canon formation. Chicago, University of Chicago Press.

Hamilton, W.L., Leskovec, J., Jurafsky, D. (2016) Diachronic word embeddings reveal statistical laws of semantic change. arXiv/org/abs/1605.09096.

Harzing, A.-W., Alakangas, S. (2016) Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics 106(2), 787–804.

Hedges, L.V. (1987) How hard is hard science, how soft is soft science? The empirical cumulativeness of research. American Psychologist, 42, 443-455.

Helbing, D. (2012) Accelerating scientific discovery by formulating grand scientific challenges. The European Physical Journal Special Topics, 214(1), 41-48.

Hellqvist, B. (2010) Referencing in the humanities and its implications for citation analysis. Journal of the American Society of Information Science and Technology, 61(2), 310–318

Hicks, D. (1999) The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. Scientometrics 44(2), 193–215.

Koskas, M. (2022) Universal bibliographic control today: Preliminary remarks. In G. Bergamin and M. Guerrini (eds.) Bibliographic control in the digital ecosystem. Florence, AIB-EUM-Firenze University Press.

Kozlowski, A.C., Taddy, M., Evans, J.A. (2018) The geometry of culture. Analyzing meaning through word embedding. American Sociological Review, 84, 905-949.

Kulczycki, E., Engels, T.C.E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Isteni-Stari, A., Zuccala, A. (2018) Publication patterns in the social sciences and humanities: evidence from eight European countries. Scientometrics, 116(1), 463–486.

Kulczycki, E., Guns, R., Pölönen J., … Sivertsen, G. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. Journal of the Association for Information Science and Technology. 1–15. DOI:https://doi.org/10.1002/asi.24336

JISC (2009) In from the Cold An assessment of the scope of ‘Orphan Works’ and its impact on the delivery of services to the public. Available at https://web.archive.org/web/20091118103903/http://sca.jiscinvolve.org/files/2009/06/sca_colltrust_orphan_works_v1-final.pdf#

Lent, H., Hahn-Powell,G., Haug-Baltzell, A.,Davey, S., Surdeanu, M., Lyons, E. (2018) Science citation knowledge extractor. Frontiers of Research Metrics and Analysis, 3, 35.

Linmans, A.J.M. (2009) 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. Scientometrics, 83(2), 337–354.

Liu, L., Jones, B.F., Uzzi, B., Wang, D. (2023) Data, measurement and empirical methods in the science of science. Nature Human Behaviour, 7, 1046-1058.

Lu, C. et al. (2019a) Analyzing linguistic complexity and scientific impact. Journal of Informetrics, 13, 817-829.

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schaars, M., Zhang, C. (2019b) Examining scientific writing styles from the perspective of linguistic complexity. Journal of the American Society of Information Science and Technology, 70(5), 462-475.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., Delgado López-Cózar, E. (2021) Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. Scientometrics, 126(1), 871–906.

Melluso, N., Bonaccorsi, A. (2024b) Novelty and interdisciplinarity. Mimeo.

Miao, L., Murray, D., Jung, W.S., Larivière, V., Sugimoto, C.R., Ann, Y.Y. (2022) The latent structure of global scientific development. Nature Human Behavior, 6, 1206-1217.

Nature (2005) In praise of soft science. Nature Editorial, 435, 1003.

Nederhof, A.J. (2006) Bibliometric monitoring of research performance in the social sciences and the humanities: a review. Scientometrics, 66(1), 81–100.

Peng, H., Ke, Q., Budek, C., Romero, D.M., Ann, Y.Y. (2021) Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. Science Advances, 7(17), eabb9004.

Peroni, S., Shotton, D., Vitali, F. (2016) A document-inspired way for tracking changes of RDF data. Proceedings of the 1stWorkshop on Detection, Representation and Management of Concept Drift in Linked Open Data. CEUR Workshop Proceedings-

Petr, M., Engels, T.C.E., Kulczycki, E., Dušková, M., Guns, R., Sieberová, M., Sivertsen, G. (2021) Journal article publishing in the social sciences and humanities: a comparison ofWeb of Science coverage for five European countries. PLoS ONE 16(4), 0249879.

Possemato, T. (2022) Entity modelling. JLIS.it, XIII, 3, 12-28.

Rancière, J. (2009) Aesthetics and its discontents. Cambridge, Polity Press.

Sīle, L., Pölönen, J., Sivertsen, G… Teitelbaump, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: findings from a European survey. Research Evaluation, 27(4), 310–322. DOI: https://doi.org/10.1093/reseval/rvy016

Small, H. (2013) The value of the humanities. Oxford, Oxford University Press.

Stone, S. (1982) Humanities scholars: information needs and uses. Journal of Documentation, 38(4), 292–313.

Sula, C.A., Miller, M. (2014) Citations, contexts, and humanistic discourse: toward automatic extraction and classification. Literary and Linguistic Computing, 29(3), 452–464.

Tsakonas, G. (2024) Big cultures and small languages: A new paradoxography in a shifting research system. Paper presented to the Fiesole Retreat Conference (available at https://youtu.be/_15zEKBKSM4).

Tshitoyan, M.L., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O. et al. (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 571, 95-98.

Visser, M., van Eck, N.J., Waltman, L. (2021) Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies. 2(1), 20–41

Waltman, L. (2016) A review of the literature on citation impact indicators. Journal of Informetrics, 10(2), 365–391.

Wang, D., Barabàsi, A.L. (2021) The science of science. Cambridge, Cambridge University Press.

Watson-Boone, R. (1994) The information needs and habits of humanities scholars. RQ 34(2), 203–215.

Wiberley, S.E., Jr. (2009) Humanities literatures and their users. Encyclopedia of Library and Information Sciences, 3rd edition, 2197–2204.

Wu, L., Wang, D., Evans, J.A. (2019) Large teams develop and small teams disrupt science and technology. Nature, 566 (7744), 378-382.