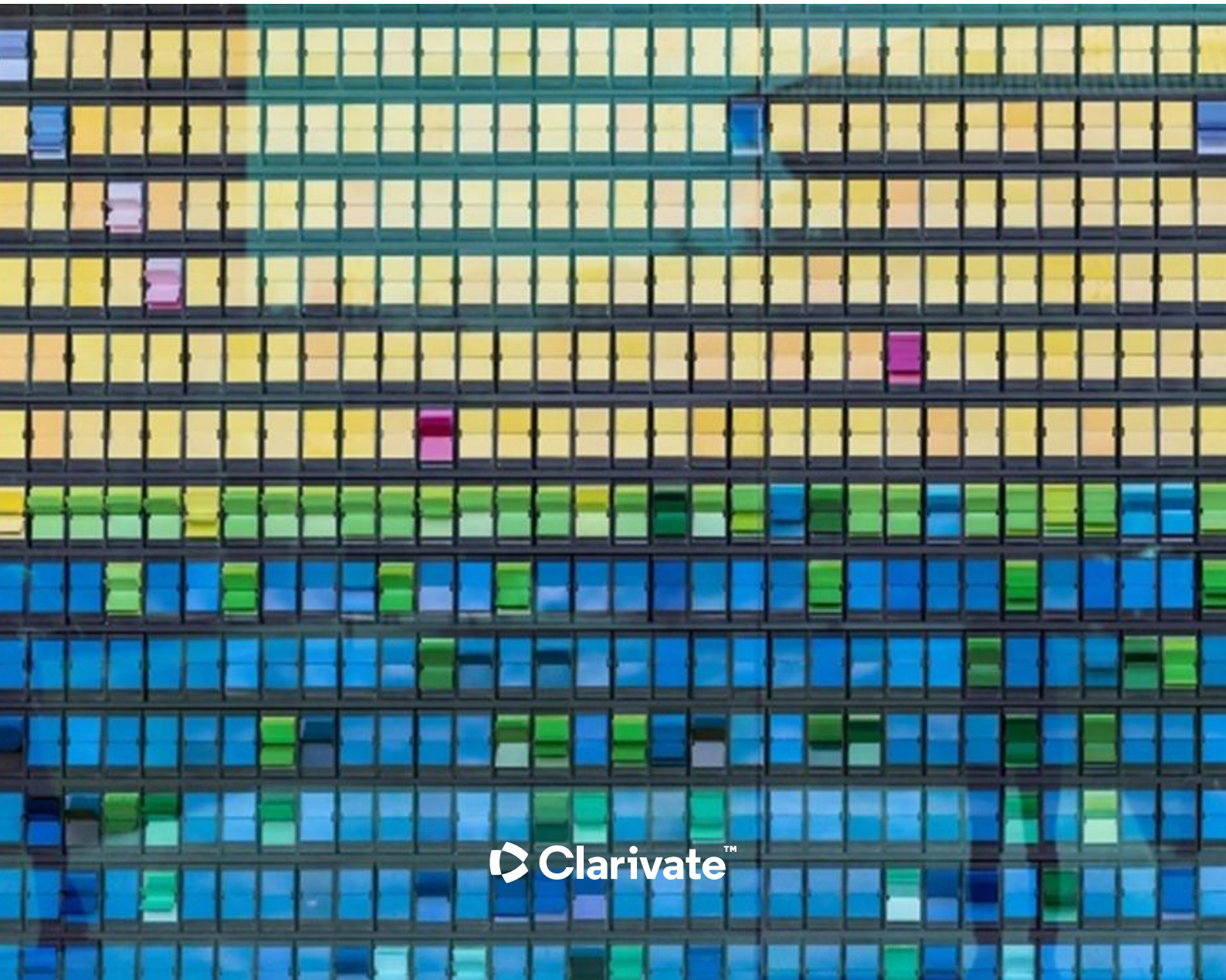
Institute for Scientific Information

Research categorization and the value of structured data

March 2026

Jonathan Adams, Dmytro Filchenko



Clarivate™

Author biographies

Dr. Jonathan Adams is Chief Scientist at the Institute for Scientific Information (ISI). He is also Visiting Professor at King's College London, Policy Institute. In 2017 he was awarded an Honorary D.Sc. by the University of Exeter, for his work in higher education and research policy. ORCiD: 0000- 0002-0325-4431. Web of Science ResearcherID: A5224-2009.

Dr. Dmytro Filchenko joined Clarivate in 2024 as Senior Director, Research & Analytics at the Institute for Scientific Information. He holds a PhD in Mathematical Modelling and Computing from Ukraine and brings more than 15 years of experience in both academia and business. His diverse leadership background spans roles including Associate Professor and Deputy Vice-Chancellor at Sumy State University, Head of Benchmarking and Technical Director. He is also a business founder. Prior to joining Clarivate he worked at QS Quacquarelli Symonds, where he directed the development of the QS World University Rankings suite and a range of other edtech and research intelligence products.

Foundational past, visionary future

About the Institute for Scientific Information (ISI)

The Institute for Scientific Information at Clarivate has pioneered the organization of the world's research information for more than half a century. Today it remains committed to promoting integrity in research while improving the retrieval, interpretation and utility of scientific information. It maintains the knowledge corpus upon which the Web of Science index and related information and analytical content and services are built. It disseminates that knowledge externally through events, conferences and publications while conducting primary research to sustain, extend and improve the knowledge base.

For more information, please visit **www.clarivate.com/isi**

About ISI reports

ISI reports offer concise and informative analyses of topical research trends, using best-in-class publication and citation data and analytics from Clarivate. This report, one of an ISI series on methods and analyses in scientometrics, explains how and why Clarivate structures research activity data to ensure that discovery, evaluation, and decision making across the global research ecosystem are based on reliable, comparable, and context aware information.

DOI: 10.14322/isi.research.categorization.and.the.value.of.structured.data

Summary

Executive summary

Good research analysis, in science, engineering or the economic and social sciences, depends on validated and structured data. A well-managed database is made even more versatile when it also has comprehensive metadata that make it interoperable with complementary data sources. What is true about research is also true for research information and analytics. This report is about the processes used by Clarivate and the Institute for Scientific Information (ISI) to ensure that the essential features for data, categorical structure, and associated metadata are central to Web of Science bibliometric database and related products.

ISI's categorization of academic journals originated in 1956 with the publication of *Current Contents*, a regular bulletin alerting researchers to new journal issues. Initial coverage of biology and medicine soon expanded to editions covering the full range of research including social sciences and arts and humanities. Each edition had sub-sections for particular discipline groups and was indexed by key metadata: author, location, keywords and so on.

As technology and publication diversity advanced, so ISI and successor divisions within Clarivate recognized and addressed a suite of challenges around publication data structure. In this report we discuss the following issues relating to the evolution of our approach:

Research culture: Culture as well as content differs between subject areas: differences in research organization, planning, management and publication. One variable is the primacy given to journals, conference proceedings, or books. To make comparisons like-for-like, such differences need to be accounted for.

> Section 2 lists the main document types in the Web of Science Core Collection and shows how their use varies between subject areas (Figure 1).
>
> Section 3 provides examples of other cultural differences between subject categories and how they affect bibliometric analysis (Figure 2, Figure 3).

Top-down categories, bottom-up topics: Classification usually starts from a broad perspective, so categories appear top-down into the detail and systems remain stable over time so information can be retrieved in familiar ways. An alternative route used to create more topical and transitory structures is to start with the finest level of detail and group upwards on the basis of common characteristics that also create a more topical structure.

> Section 4 reviews the fine and broad-grained structures used for Web of Science data and describes a citation-based topic system that ISI developed with The Centre for Science and Technology Studies (CWTS) at Leiden University.
>
> Section 5 expands on cross-content bottom-up categorization, showing how complementary databases can be linked to throw more light on key topics.

Mapping research to purpose and outcomes: Categories must be comprehensive, and in Web of Science that means not only comprehensive but also suitable for different research purposes. Research evaluation systems are usually structured for management and policy purposes. Research policy objectives are defined by a national need or social purpose, such as their impact on the economy and society. How do we link these?

> Section 6 links global research journal categories to national systems of categorization for assessment and shows how ISI mapped this for RAE 1996 (Figure 4).
>
> Section 7 discusses the methods used to map Web of Science data to UN Sustainable Development Goals.

The impact of collaboration: The need for research subject categories was well established by ISI in the 1990s, but recent ISI research has revealed that international collaboration boosts citation counts and impact.

> Section 8 shows how we categorize domestic and international collaboration in InCites Benchmarking & Analytics to identify research that is well cited compared to output of the same type (Figure 5, Table 1).

Why structured research activity data matters: Structured, verified, and interoperable research activity data are not a convenience – they are a prerequisite for credible search, discovery and evaluation.

> The Conclusion summarizes key actionable insights from the report for researchers and research analysists including research offices, funders and policy makers.

1. What is research categorization and why does it matter?

Selected, validated, and structured data provide the foundation for credible research, and that foundation is strengthened when a database is made versatile and interoperable with related, complementary databases through comprehensive metadata. What is true generally for research is equally true for research publication and performance data. Clarivate's experience over decades of methodological development – originating in the Institute for Scientific Information (ISI) – has delivered the Web of Science: a system of databases linked by shared metadata and well-defined categorical structures that support researchers, institutions, and policymakers in navigating an increasingly complex and interconnected research landscape[i].

For the research reader, there has been a long-standing need to index and to categorize academic publications. This need became more acute after 1945, as the numbers of research journals proliferated. Eugene Garfield saw that researchers needed signposts to track the rich supply of 'latest information'. That led to the foundation of the Institute for Scientific Information (ISI), now a core part of Clarivate, and the introduction of the weekly *Current Contents* bulletins. Clarivate currently indexes more than 22,000 editorially-selected journals, with each allocated to the 254 journal subject categories in the Web of Science Core Collection and processed cover-to-cover in a growing data archive that now contains 99 million metadata records and 2.6 billion citation links.

Analysis also needs structured data. A frequently cited paper may be a significant marker of influential research, but how does one citation count stack up against others? Which 'other' papers are like-for-like and useful for comparison: same subject, same year, same country/region? How wrong might your analysis be if you don't account for such variables? A classic error when first looking at publication and citation data is to calculate an overall average of citations per paper, but such an average is meaningless and misleading. The right categories make the difference.

Research categories may be subjects, disciplines, fields, types of research activity and document types. Every piece of information about research has associated characteristics, or metadata, that tell us something about the activity. What it was about, where it happened, when it was done, who did the work and with whom, and who paid for it to be done. Each of these characteristics can influence the process and outcomes of a research project. They also tell us about links to other types of data labels such as patents, research programs and policy targets. So, they must be properly recorded, displayed, and available to readers, researchers and analysts.

When these metadata are incomplete, inconsistent, or poorly structured, research discovery and analytics can be significantly distorted. Key challenges for researchers and analysts include:

- Unstructured time metadata can create misleading trends and incomplete coverage introduces analytical gaps.

- Weak or inconsistent subject categorization can undermine comparisons and lead to false conclusions.
- Inconsistent document-type metadata can invalidate comparisons.
- Unstructured coverage of non-traditional outputs can limit discovery and analytics across data sources.
- Incomplete affiliation data can distort collaboration analysis and a failure to take into account collaboration modes can give rise to misleading conclusions.

2. Research cultures and research documents

Research culture affects the ways in which people publish their research and draw on prior literature. Publication and citation patterns have become characteristic of each particular type of research activity. For example, biomedical research groups typically publish frequently, many biomedical papers are short, and they reference standard methodologies. Engineers, by contrast, often publish first in conference proceedings and later in journal articles, so they produce less frequent papers with more consolidated findings. There is consequently also a relatively larger pool of citations for biomedicine than engineering, so the citation distribution profiles naturally differ.

The principal types of research documents and their characteristics include:

- **Journal articles and reviews** are the primary output of original research across the natural sciences. Citation is a core part of research culture and the relationship between relatively high citation impact and peer esteem is well established.
- **Conference proceedings** are important as a route for rapid communication to research users in engineering and technology, and thus closer to application, but citation patterns differ from those of journals.
- **Books, chapters in books and monographs** are a key output for arts and humanities. A monograph in these fields may take significant time to build a citation profile even if it receives significant acknowledgment.
- **Grey literature** refers to many forms of reports from research groups, public bodies, think tanks, government departments and other organizations drawing on academic research. This literature is important but unstructured and has only recently started to be indexed. Links to prior research on which reports are based are inconsistent and often difficult to validate.

Any document may be of significance for search and discovery, so Clarivate categorizes all these document types in Web of Science to a common standard and structure to enable full and complete search and discovery.

However, for research evaluation, peer-reviewed journal articles and reviews are typically used as the key source of citation data. Citation analysis is usually restricted

to the citation indices of journal literature that collectively form a common currency across disciplines. It would be wrong to assess humanities research without looking at books and to assess research impact (particularly its wider societal impact) without grey literature, but simple citation counts among such publications currently remain uninformative for management and evaluation and therefore require careful and responsible use and interpretation [ii].

ISI worked with the manager of the U.K.'s cyclical Research Assessment Exercise (RAE) and Research Excellence Framework (REF) on detailed analyses of the balance of documents selectively chosen by academics for peer evaluation. Since funding is affected by that evaluation, these should reflect what the researchers think represents their most important work. It turns out that while the balance of document types varies by major field in the way we might predict, the balance within fields more surprisingly varied over time.

Figure 1. The balance of research documents selected by academics for the U.K. research evaluation cycles from 1996 to 2021. The balance across four main document types is shown for Units of Assessment (UOAs) grouped by broad faculty areas. 'Other submitted outputs' include patents, grey literature, exhibition and performance records, and other non-published material ***Error! Reference source not found.***[iii].

The balance of research output types changed across 25 years towards journal articles, from monographs in Social Sciences and from conference proceedings in Engineering. Science was always article-focused, while the Arts and Humanities have remained attached to books as a preferred publication mode.

A well-organized database such as Web of Science comprehensively covers and clearly labels output types. Understanding these differences in document types and publication practices is not only important for discovery but also has significant implications for how research performance is analyzed. The next section considers

how publishing behavior shapes citation behavior, making rigorous categorization essential for meaningful evaluation.

3. Why research categorization is essential for citation analysis

A critical document characteristic, right across research literature, is that citations accumulate over time. It has long been known that citations to journal articles typically arrive rapidly in the years immediately after publication and that the annual increase gradually falls to a citation count plateau after around ten years. Because of the pattern of citation accumulation, recent papers have fewer citations on average than those published longer ago.

If we take today's data and plot the simple average of citations per paper, tracking any given country/region forward to current from some point in the past, then a data plot shows a decrease in the average number of citations per paper as we get closer to the present day. This reflects the time that older papers had to be cited. So, to improve the like-for-like nature of data in a citation analysis, we need to categorize all our papers, globally and nationally, by year of publication and then recalculate citation impact comparing each paper to the relevant annual global average.

But time is not all that needs to be taken into account. Because subject cultures and publication practices differ, as we noted for biology and engineering, our like-for-like research data must be categorized not only by year but also by subject to produce Category Normalized Citation Impact (CNCI). In CNCI, each paper's citation count is 'normalized' as a ratio compared to the global average for the papers of the same paper type, in its year of publication and in the relevant Web of Science journal category (Figure 2).

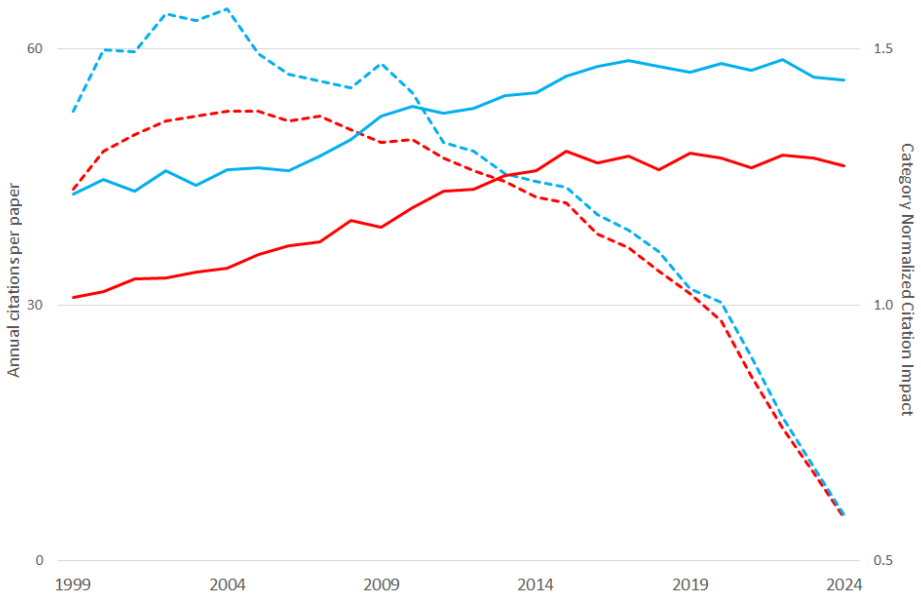


Figure 2. Calculations of annual citation impact for the U.K. (blue) and for Germany (red). The dashed lines show the result using a simple average of citations accumulated to date for the papers published in each year. The solid lines show the result when these same citation counts are normalized against the global average for both the Web of Science subject category to which the journal is assigned and the year of publication. (Data source: Web of Science Core Collection.)

The difference between simple and relative (Category Normalized) citation impact is central to our interpretation of research performance. For Germany and the U.K., in our example, we see, first, that relative impact is in fact sustained compared to the simple global average and, second, that relative impact actually rose for both countries after 2000 through to 2014.

Misinformation caused by using 'simple citation impact' at country/region level would become even greater at institutional level because of specialist portfolios. For example, an institution with a technology mission is not like-for-like in any comparison with an institution with a medical focus. Raw citation data would be distorted by the faster citation growth rate of the biomedical papers, undermining planning within institutions as well as giving a distorted external perception.

A new time variable – early access

Another variable to consider for each document is its actual publication date, which may differ from the nominal publication date on the cover of the relevant journal. The difference is between (i) the date when the publisher assigns a document to a journal issue or (ii) the date when the document actually became available online in a public database. This is why Web of Science now has an "Early Access"' document type. After the early access document has been assigned to a final issue and has a volume, issue, and page number, the "Early Access" tag is removed but both online and final publication dates are indexed and accessible for discovery and analytics.

From the perspective of the individual researcher, who is looking for information at the earliest possible point, early access is beneficial. The indexing of these papers by Web of Science enables rapid discovery. However, for the analyst, and for the research manager using bibliometric reports, early access creates a potential complication because availability ahead of assignment to a journal issue is not uniform across subject categories or countries/regions. Even within a single journal issue, some papers may have had earlier online availability – and the potential to be cited sooner – than others, so comparability may be compromised.

Subject categories – oversight and detail

Research in the 1980s showed that the rate at which citation counts rise over time varies between subject areas. Journals therefore need to be categorized using a scheme that groups culturally related subjects. The descendants of the initial ISI journal categories are now in Web of Science (254, fine grained) and Essential Science Indicators (22, broad grained).

We can see that older papers, in any discipline, have more citations on average than the latest publications. As noted, papers in biomedical journals tend to get cited soon after publication. Papers in engineering typically take longer to get cited and the rate of accumulation is lower than in biomedicine. Because of differences in citation behavior, and the pool of citations, the average citation count for biology papers plateaus at a higher count than for engineering. Within major disciplines other differences emerge: for example, molecular biology is cited more often than whole-organism sciences (Figure 3).

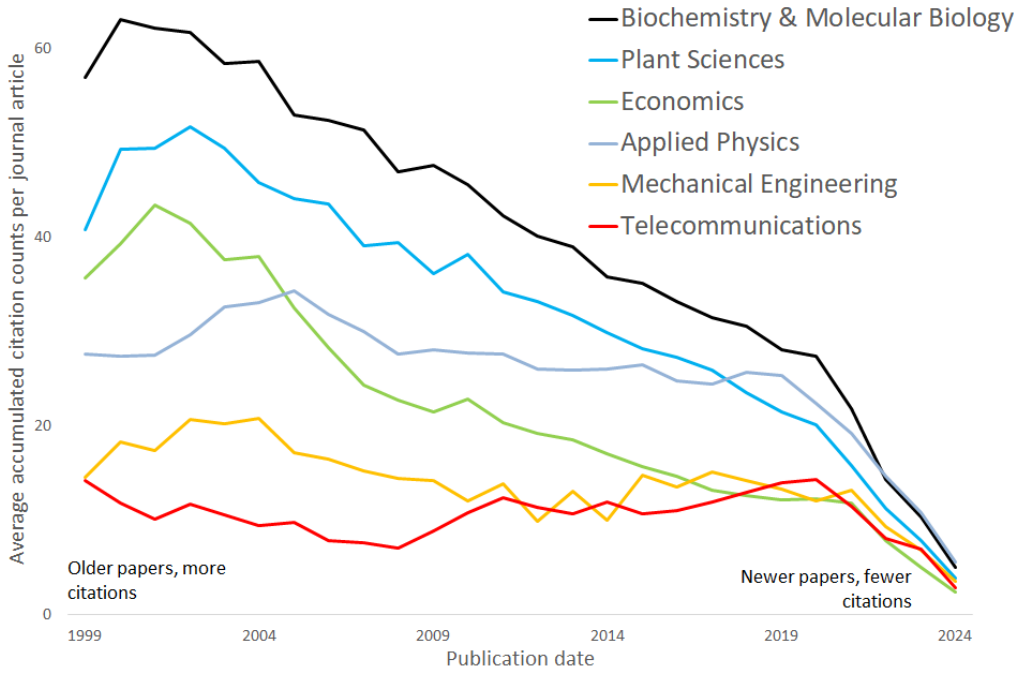


Figure 3. Average accumulated citation count per journal article rises over time and does so at different rates in different subjects; e.g., a Biochemistry & Molecular Biology article is cited 5 times on average in 2024, but the average citation count for 1999 articles has risen over 25 years to 60. This analysis shows five of 254 specific journal-based subject categories in Web of Science. These categories account for detailed cultural differences between subjects and support management information.

The data in Figure 3 do not mean that Biochemistry & Molecular Biology has greater research impact than Mechanical Engineering or Telecommunications. Telecommunications are unquestionably key innovative technologies, but their significance would not be evidenced by the observable citation rates for their basic research output. This difference may arise because biomedical researchers write many biomedical papers and each has many references, leading to a bigger pool of citing sources than in engineering.

Objectives, timescales, experimental methods and patterns of discovery differ between research domains and so do their publication and citation rates. Research cultures continue to evolve, so these differences are dynamic and ISI actively monitors such change.

Clarivate groups all research publications, including journal articles, conference proceedings and books, by subject and discipline. All are available for publication and citation analysis in the InCites research analytics platform.

- The 254 Web of Science journal categories are foundational, stable and detailed categories capturing a set of journals with close links and many cross-citations. They identify consistent areas of research across decades.
- The 22 Essential Science Indicators (ESI) journal categories give a managerial overview. These do not include the arts and humanities because of their reliance on monographs. There are 21 broad 'school' categories, such as Physics and Economics, plus one category for Multidisciplinary journals.

Together, these category systems provide an excellent starting point for the conceptual 'maps' that guide a researcher's search or a bibliometric analysis. The ESI overview is essential for scene setting and gives a perspective of the research landscape for an analyst or research manager to view the landscape. Web of Science then enables exploration once the map has been created, focusing on particular areas but knowing the key signposts from the big picture.

4. How citation networks identify subject categories

Robust and stable subject categorization is essential for meaningful citation analysis and longer-term management. The next question is whether categories can be identified and maintained in a rapidly evolving research landscape.

Web of Science and ESI categorization are guided 'top-down' by relationships between journals, their similarities in content, and cross-authorship. New categories can, and are, added from time to time. Established categories evolve as journals appear, flourish and wither. But the main landmarks are there each time the user returns to their core categories.

A more topical approach is a 'bottom-up' methodology created by citation links between individual articles, ignoring the specific journal in which they were published. There are costs and benefits to this.

- Boundary thresholds of similarity/difference must be set. That means processing the citation links between all of many millions of research articles, counting the cross-links between pairs, and deciding where a general threshold can be determined. This can be scaled to coarse- and fine-grained lists.
- The system is dynamic. New citations are continually being added as new articles appear and cite older material. A dynamic structure can be seen as unstable because of shifts and changes but, although these changes require the user to relearn the map each time, they reveal innovation.

A model developed in 2019 by the Centre for Science and Technology Studies (CWTS) research group at the University of Leiden, in collaboration with the ISI, underpins the first such dynamic map of Citation Topics[iv]. There are three levels with ten fixed topic groups at the macro level, each of which then divides into many meso-topics, and then into a proliferation of micro-topics. For example:

Citation Topic 2 is Chemistry, which includes

2.39 Polymer Science

2.76 2D-Materials

2.74 Photocatalysts, which includes

2.74.16 ZnO Nanostructures

2.74.1306 Electrochromism

It is infeasible to retain a working picture of such a detailed categorical structure. However, when reviewing recent research and considering interdisciplinary connections, an expert will benefit from first exploring Web of Science journal categories and then moving to more selective exploration using Citation Topics.

This approach to using citations as markers of topics significant to current researchers is refined in Web of Science by focusing only on the publications with exceptional citation counts in the last few years. This identifies 'Emerging Topics': innovations that have sparked rapid uptake elsewhere in recent publications.

Emerging Topics are grouped for researchers in Web of Science Research Intelligence and the Research Horizon Navigator tool within InCites. Like the categorical structures discussed earlier, this guides the user through stable and familiar layers to the changing landscape below. For example:

Literature & Language – 132 emerging topics in 14 categories

e.g., American Literature (2 topics) including

Global inequality, identity, and sociopolitical change

Agricultural Sciences – 120 emerging topics in 7 categories

e.g., Agricultural Engineering (4 topics) including

Sustainable environmental remediation materials

The detailed publication list at the topic level can then be downloaded, read in detail and evaluated by the expert researcher.

5. Cross-content subject categorization

While citation-network approaches reveal how research topics emerge from patterns of scholarly referencing, further structure is needed to organize knowledge consistently across the full range of research outputs where citation networks may be absent or poor.

Clarivate's innovative Research Topics classification scheme has been implemented in Web of Science Research Intelligence to enhance the accuracy, relevance, and interpretability of research discovery and analytics. By bridging content types and platforms, it establishes a model for knowledge organization within complex research ecosystems.

Research Topics classification is designed to address a long-standing challenge in research data organization by enabling cross-platform and cross-content aggregation. Traditional classification systems, such as journal-based subject categories, segment research by source or publication type. This works well with

those data, but it limits their effectiveness if the aim is to unify diverse data assets or to explore connections across publications, patents, and grants.

Research Topics overcome these limitations by establishing a unified taxonomy that can be consistently applied across heterogeneous research outputs. The three-level hierarchical structure (macro-, meso-, and micro-topics) is derived from the established taxonomy used for Citation Topics. However, unlike citation network-based approaches, Research Topics are assigned via a semantic similarity analysis of text content including titles, abstracts, and keywords. A content-driven methodology enables multiple topic assignments per item so it can extend classification beyond scholarly publications to include patents, grants, and other research inputs and outputs.

This structure addresses the critical need for cross-platform classification by:

- Integrating heterogeneous research outputs into a single coherent schema, enabling consistent tagging across scholarly literature, funding data, and intellectual property.
- Enabling cross-content discovery and analytics, allowing the same hierarchical topics to support search, filtering, and reporting irrespective of content source.
- Supporting interdisciplinary analysis through multi-topic assignment, reflecting the inherent complexity and convergence of modern research domains.

This classification system is dynamic, because the content-driven taxonomy evolves with the developing research landscape. Institutions can use this to analyze research portfolios in alignment with real research activity rather than disciplinary boundaries.

An innovative application of Research Topics would be in national research assessment exercises where institutions are encouraged to submit a diverse range of research outputs. For example, in the U.K.'s Research Excellence Framework (REF), universities may submit journal articles, books, datasets, software, policy reports, and other non-traditional outputs. Historically, evaluating such heterogeneous material within a consistent subject framework has been challenging because different content types are indexed in different databases and classified using incompatible taxonomies. Now, a university preparing a REF submission could generate a Research Topics profile of its candidate outputs, showing how journal articles, a monograph, and an industry-linked whitepaper all contribute to impact within a single high-momentum topic. This would provide panels with tangible evidence of a sustained, multi-modal research program rather than presenting a collection of disconnected artifacts.

This shows how cross-content classification not only improves discovery and analysis but also creates the foundation for aligning research intelligence with external societal goals. This is described further in the next section.

6. Mapping data categories to national assessment structures

As evaluation moves from analytical models to real-world assessment exercises, the challenge is to align internationally derived database categories with nationally specified evaluation schemes. The analysis described here confirms that publication data from different category schema produce related and interpretable clusters that can be linked. It means that our categories have real meaning in the research landscape and reflect how researchers report their work.

In Australia, discipline categories used for the Excellence in Research for Australia (ERA) system are defined as two-digit and four-digit codes using the Australia and New Zealand Standard Research Classification (ANZSRC) Fields of Research (FoR). For example:

Division 31 is Biological Sciences, which includes

Group 3103 Ecology

For the 2010 evaluation onwards, the Australian Research Council (ARC) defined journal lists for each FoR. Mapping the contents of the FoRs and comparing research in Australian institutions to that in any other country/region is straightforward because of these journal lists, which are mapped in InCites Benchmarking & Analytics. The FoR schema is just one of a series of national and international categorical systems made available in InCites, including the OECD subject categories. Every mapped scheme covers all Web of Science documents, so it is straightforward for research analysts in different countries/regions to make comparisons between institutions using the discipline categories with which they are most familiar.

The U.K. Research Assessment Exercise (RAE to 2014, then REF) system does not define the journals that map to a discipline. A mapping concordance is therefore required to link the national map to a global map for comparisons to inform research policy and management. In 1996, the U.K. Chief Scientific Adviser commissioned ISI, with the Centre for Policy Studies in Education at the University of Leeds, to develop a map between Web of Science journal categories and RAE Units of Assessment (UoAs) to see whether U.K. research could be internationally benchmarked.

The RAE publication database holds about 195,000 outputs of many types (see Figure 1). For RAE 1996, we found 131,091 journal articles and reviews across 6,146 journals indexed in Web of Science. Of these, 2,158 journals had just a single article while three journals had more than 1,000 submitted records. About two-thirds of all journals (3,988) were submitted to more than one UoA.

Evidently, and unsurprisingly, journal use overlaps between UoAs. Some overlaps were substantial: for example, for UoA Chemistry, 151 (of 525 total) journals accounted for 56% of the submitted articles but 57 of these frequent journals appeared at least once in Physics and 68 in Chemical Engineering. One-to-one mapping concordance is impossible, so a more ‘fuzzy’ map was developed. We can summarize the process using the example of UoA Physics:

Step 1, ISI tallied the number of RAE-submitted articles that were published in journals in three Web of Science journal categories core to UoA Physics research. These captured 59% of 5,976 RAE articles.

Step 2 looked at three journal categories cognate to UoA Physics and within the scope published by the Physics' panel. This addition captured another 23% of the RAE articles (82% total).

Step 3 considered possible additional journal categories. For example, adding the Mathematics journal category would have drawn in 180 journals but accounted for only 5 additional articles. The addition is ruled out as it dilutes the core.

Fixed journal lists in Web of Science are good for search and analysis but miss these overlaps 'on the ground'. The RAE's flexible approach to journal maps captures that mosaic but does not translate for international comparisons. Cross-mapping solves that problem. For Physics, the [three core] + [three cognate margin] Web of Science journal categories capture a high proportion of research deemed by U.K. physicists to represent their work while minimizing additional 'non-RAE' journal coverage.

Because the journal lists are non-exclusive between UoAs, we can also perform a similarity analysis of the RAE journal usage, comparing the number of submitted articles for each journal between each UoA. This analysis enables us to draw disciplines together in a tree diagram (dendrogram) starting with the most similar.

Publishing relationships between the physical sciences, engineering and mathematics are revealed by this. It shows that UoA Chemical Engineering publications cluster with Chemistry and Physics rather than other engineering UoAs, while UoA Mining Engineering does not cluster in this part of the network at all (it actually clusters with Earth Sciences). Such novel insights are valuable in terms of understanding 'like-for-like' grouping: whatever the label, Chemical Engineering research is more about chemistry than it is about engineering (Figure 4).

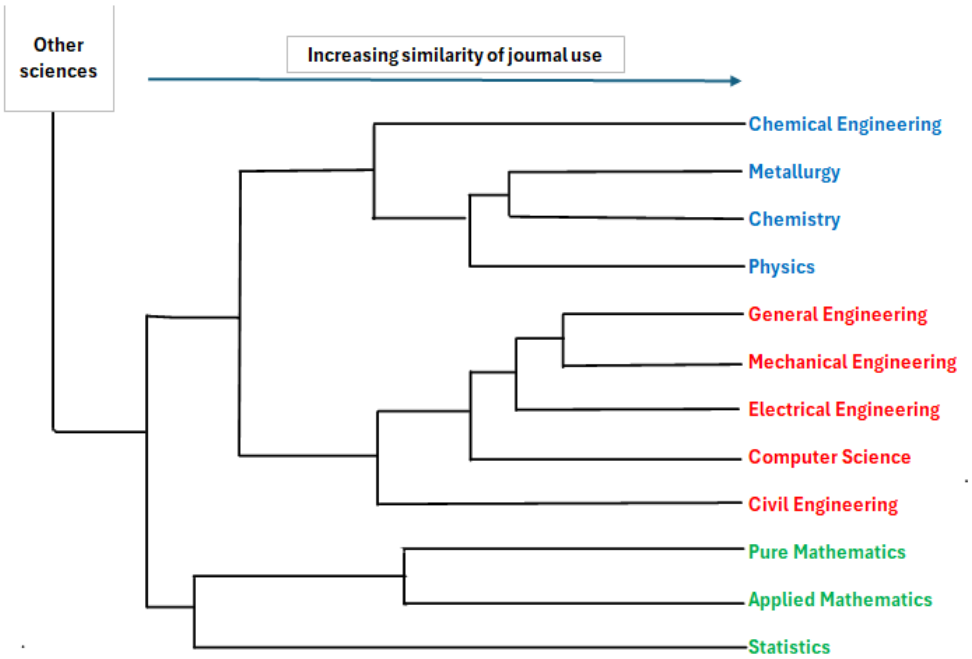


Figure 4. Progressive clustering of physical sciences, engineering and mathematics Units of Assessment (UoAs) from the U.K.'s RAE 1996 data, calculated by similarity of journal use indexed by articles submitted for assessment. This analysis shows that key publications submitted by researchers in Chemical Engineering cluster with physical sciences UoAs rather than engineering[v].

7. Mapping research metadata to objectives

The United Nations Sustainable Development Goals (SDGs) constitute a globally endorsed framework of 17 interlinked objectives aimed at addressing the most pressing social, economic, and environmental challenges by 2030. They include areas such as poverty eradication, quality education, gender equality, climate action, and sustainable cities.

In Clarivate's advanced research analytics solutions, such as InCites and Web of Science Research Intelligence, the SDGs have been introduced as a classification schema to enable systematic measurement and comparative analysis of research outputs in relation to development priorities. This embodiment of the SDGs within research data frameworks responds to the accelerating imperative for evidence-based insight into how knowledge production contributes to societal impact and global transformation.

The SDG schema is constructed through a methodical mapping process, where each of the 17 goals is aligned with Citation Topics and Research Topics (at micro-level topics) for content across the Web of Science platform and in Web of Science Research Intelligence. These mappings link specific bodies of literature to the respective goals, enabling outputs to be categorized according to an aspect of sustainable development that they address. Indicators can then be computed for each SDG entity based on the associated publication sets, supporting quantification and longitudinal tracking of research contributions across the full SDG spectrum.

This classification approach combines automated clustering with expert review. As the Citation Topic or Research Topic landscape evolves through annual re-clustering, the SDG mapping is updated to sustain analytical continuity and relevance for research leaders, policymakers, and institutional strategists to evaluate progress, identify strengths and gaps, and align investment and policy decisions with global development priorities.

This approach is inherently flexible and can be extended to support alternative categorizations of societal needs or strategic objectives at both global and national levels. For example, in Clarivate's Societal Impact Framework[ii], similar mapping methodologies can be applied to PESTLE-based categorizations of research (either through SDGs as an intermediary or independently), U.K. REF areas of impact, ASIRPA levels of impact (France) or other policy or mission-oriented classification systems.

8. Categorizing international collaboration

The analytics developed for Web of Science products have used output type, time, and thematic alignment as part of data categorization since the 1990s, but recent research has shown that these are not the only factors affecting research evaluation. The structure of international collaboration has risen substantially and has become a critical dimension requiring comparable categorization.

In western Europe, for example, international collaboration accounted for less than 10% of research output in the 1980s but now represents as much as two-thirds of articles published in journals indexed in Web of Science. As their relative frequency rose, it became evident that cross-national articles were systematically cited more frequently.

Recent research by ISI[vi] has extended the principle of like-for-like categorization of research outputs. To capture this collaborative dimension, ISI developed a collaboration adjusted CNCI (Collab-CNCI), integrating collaboration into the normalization process of CNCI.

Research outputs can be filtered by defined collaboration types or classes: domestic (with no international co-authors), bilateral (with a co-author from a second country/region), trilateral (three countries/regions), and multilateral (four or more countries/regions). As with other categorical variables, an article's citation count is compared with the world average for the same collaboration type as well as the same subject category and year. We can then compare international collaborations only with other international collaborations and not with purely domestic papers.

The consequences of ignoring collaboration can be shown by comparing average annual citation impact using (i) conventional CNCI, taking year, document type and subject category into account, and (ii) Collab-CNCI, when collaboration is also considered. This shows that highly collaborative countries/regions benefit substantially from international activity, because an indicator calculated using standard CNCI is almost always higher than the equivalent indicator calculated using Collab-CNCI. However, this is not the case for Mainland China, where only around 10-15% of journal articles are internationally co-authored (Figure 5, Table 1).

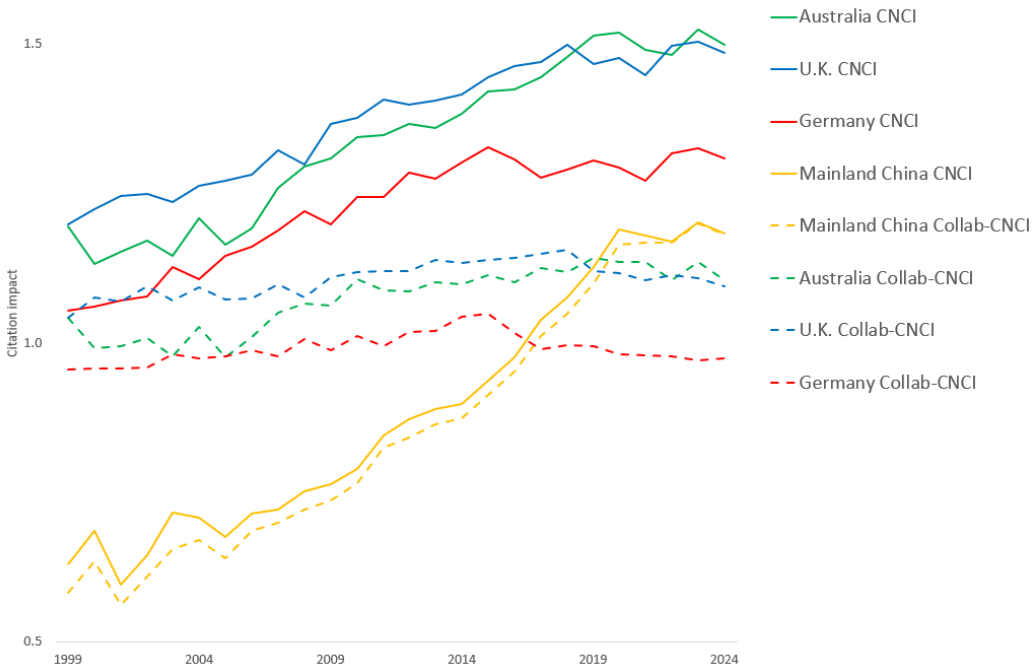


Figure 5. Trends in annual national citation impact illustrating the effect of using Collab-CNCI: categorizing papers by mode of international collaboration prior to normalizing citation counts by year and subject to calculate impact. For Australia, Germany and the U.K., the net standard citation impact (CNCI) is visibly higher than when collaboration mode is taken into account (Collab-CNCI). Mainland China has much less frequent international collaboration and its citation impact is affected very little. (Data source: Web of Science Core Collection)

This innovation does not assert that CNCI overstates citation impact. The change in the calculated indicator happens because domestic papers, with no international coauthors, typically have a lower average citation impact than internationally co-authored papers. CNCI is the overview reference indicator while Collab-CNCI expands our understanding of the detailed strengths and weaknesses in a portfolio. Deconstruction reveals that bilateral and quadrilateral collaborations may lead to citation impact indicators that are relatively higher than for domestic research, even when compared with similar collaborations among the rest of the world (Table 1).

Table 1. Collab-CNCI of articles in journals indexed in Web of Science Core Collection for the period 2020-2024. Papers are categorized and compared by collaboration types for three countries/regions.

CNCI		**Collab-CNCI**			
		All domestic	All international	Bilateral	Quadri-lateral and higher
1.50	Australia	1.06	1.06	1.11	1.34
1.18	Mainland China	1.15	1.15	1.26	1.31
1.30	Germany	0.92	0.92	0.96	1.17

The development of Collab-CNCI supports ISI's recommendations that better information emerges when we shift from metrics (CNCI alone) to profiles (a spread across collaboration types), enabling contextualized benchmarking of institutions, departments, or countries/regions. The ability to examine patterns of international collaboration alongside citation impact offers deeper insight into research activity and enables more informed policy and investment decisions. By treating international collaboration as a distinct analytical category schema, we enrich overall research intelligence of a scholarly activity that has significant influence.

9. Conclusion: Why structured research activity data matters

Our approach – selection, curation, rich and standardized metadata, and multi-layered categorization – is driven by structure as the core determinant of trustworthy research intelligence. Breadth of coverage alone does not deliver credible analytics. Like-for-like design converts raw records from trustworthy sources into reliable evidence for discovery, evaluation, and decision-making.

Why this matters to researchers

- **Find what matters faster.** Layered categorization improves recall and precision in literature reviews and horizon scans. Emerging Topics spotlight where fields are moving now.
- **Get credit that's truly comparable.** Normalized indicators protect against misleading comparisons across fields, years, and collaboration modes, ensuring like-for-like recognition of contributions.
- **Strengthen translational pathways.** Cross-content topic alignment helps researchers connect outputs to applications (e.g., patenting and funding routes) and articulate impact narratives framed to SDGs or national missions.

Why this matters to research analysts

- **Make defensible decisions.** Policy, funding, and performance assessments stand on methodologically sound, transparent foundations (CNCI, Collab-CNCI, document-type controls, early access treatment).
- **Build coherent, comparable dashboards.** Unified schemas and mapped taxonomies allow consistent KPIs across units, partners, and countries/regions, enabling scenario planning and risk-aware benchmarking.
- **Tell clearer stories with evidence.** Topic-level analytics connect capability, collaboration, and impact, supporting strategic cases for investment, hiring, and partnerships – without overstating or understating performance.

Structured, verified, and interoperable research activity data are not a convenience – they are a prerequisite for credible discovery and evaluation. By embedding categorization and normalization into every stage – from indexing to analytics – Clarivate turns a vast, heterogeneous corpus into stable, interpretable, and actionable intelligence for researchers and research analysts alike.

References

[i] 2020. Adams J., Pendlebury D., Szomszor M. The value of bibliometric databases: Data-intensive studies beyond search and discovery. Clarivate, https://clarivate.com/academia-government/lp/the-value-of-bibliometric-databases-data-intensive-studies-beyond-search-and-discovery/

[ii] 2024. Filchenko D., Pendlebury D., Quaderi N. and Adams J. A responsible framework for evaluating the societal impact of research. Clarivate, https://clarivate.com/academia-government/lp/a-responsible-framework-for-evaluating-the-societal-impact-of-research/

[iii] 2020. Adams J., Gurney K. A., Loach T. and Szomszor M. Evolving document patterns in UK research assessment cycles. Frontiers in Research Metrics and Analytics, 5, 2 (23 April 2020) https://doi.org/10.3389/frma.2020.00002

[iv] 2021. Szomszor M., Adams J., Pendlebury D., Rogers G. Data categorization: understanding choices and outcomes. Clarivate, https://clarivate.com/academia-government/lp/data-categorization-understanding-choices-and-outcomes-2/

[v] 1998. Adams J., Bailey T., Jackson L., Scott P., Pendlebury D. and Small H. Benchmarking of the international standing of research in England: report of a consultancy study on bibliometric analysis. Centre for Policy Studies in Education, University of Leeds. 108 pp. ISBN 1 901981 04 5

[vi] 2025. Adams J., Potter R., Filchenko D. Unlocking the efficiency of international research collaboration: A new intelligence solution for informed decision-making and policy planning. Clarivate, https://clarivate.com/academia-government/lp/unlocking-the-efficiency-of-international-research-collaboration/

About Clarivate

Clarivate is a leading global provider of transformative intelligence. We offer enriched data, insights & analytics, workflow solutions and expert services in the areas of Academia & Government, Intellectual Property and Life Sciences & Healthcare. For more information, please visit **clarivate.com**.

Web of Science is the world's largest publisher-neutral citation index and research intelligence platform. It organizes the world's research information to enable academia, corporations, publishers and governments to accelerate the pace of research.

Need to evaluate research at your organization?
Contact us to find out how Clarivate can help:

clarivate.com/contact-us