

Enhancing Bibliometrics With CRIS-Systems

Lars Wenaas¹ and Jens Aasheim¹

¹Norwegian Agency for Shared Services in Education and Research (Sikt), Norway
lars.wenaas@sikt.no, jens.aasheim@sikt.no

Abstract

A CRIS-system alone can be used for a range of bibliometric analyses but also comes with natural limitations, especially the possibilities of citation-based analyses. However, a combination of a CRIS and a bibliometric database can overcome some of these limitations and utilize the best of both worlds; the accuracy of affiliation data in a CRIS and citation-data from a bibliometric source. Norway has a national CRIS-system and a national infrastructure for bibliometrics, both of which are hosted by the Norwegian Agency for Shared Services in Education and Research (Sikt). This represents a case which is particularly suited for exploration, and Sikt has developed a pilot integration between the two systems in R. The NIB Analytics pilot makes it possible to easily access and download data, integrate data and perform a limited number of standard analyses. The pilot shows potential for wider application and more accurate results when analysing data on the institutional or national level. Sikt intends to develop the pilot as an open framework together with the bibliometric community in Norway, based on the principles of open science and the Barcelona-declaration.

1 Introduction

In general, data are increasingly important for research institutions and used for administrative tasks, keeping track of production, reporting, planning, applications and research evaluation of all kinds. Bibliometrics with easily comparable (although not always representative) numerical metrics have gained a prominent, but also a debated place in this picture. Particularly analyses with emphasis on citations, these “pellets of recognition” [1], are of concern, and the Leiden-manifesto [2] is one initiative that has issued warnings on the role of bibliometrics in research evaluations. Even so, bibliometrics has acquired an important and popular evaluative role and consequently become a part of research institutions’ administrative toolbox.

CRIS-systems are rich sources of data and function as an inventory of institutions scientific production, but they rarely contain citation-data and are in any case limited in scope and coverage. While CRIS-system most often are implemented on the institutional level, Norway has a national CRIS-system, the Norwegian Research Information Repository (NVA). The system’s main function has historically been as a reporting tool for scientific production to the ministry of education and

research and it therefore keeps track of all scientific publications published by Norwegian publicly financed institutions.

On the national level, data from the Norwegian CRIS serves many other useful tasks, particularly gauging the level of open access over time on a yearly basis [3]. Each article eligible for reporting is checked against open access-sources as Directory of Open Access Journals (DOAJ), Unpaywall or publisher's reports resulting from consortia agreements, then the information is used to identify an article's current open access-status. NVA-data is also underpinning publish-and-read-agreements, where negotiations with publishers are supported and influenced by data on which journals Norwegian researchers prefer to read and publish in. COUNTER statistics enter the mix to gain insight into visiting patterns, while existing shares of open access in journals obviously are of interest in negotiations with publishers.

CRIS-systems are thus suitable for supporting a wide variety of tasks by default, but as the above examples show, many analyses also require additional data to produce insights. A field where CRIS-system falls short is in traditional bibliometric analyses involving citations. Bibliometric analyses is usually conducted with basis in a traditional bibliometric source, but these also have their own challenges. Common features with bibliometric sources are incomplete coverage and/or missing or bad affiliation data also resulting in excluded data. A CRIS-system can help address these challenges, first by identifying which articles are excluded from the analysis, and second by including articles in the analysis where an article is missing or have bad affiliation-data. This has been piloted in Norway, where Sikt maintains responsibility for both NVA and the National Infrastructure for bibliometrics (NIB).

The purpose of this paper is to outline the pilot, how data from a national CRIS and data from traditional bibliometric sources can be joined to enhance bibliometric analyses, and what benefits and obstacles that lie ahead.

2 Background: the infrastructure in Norway

The services Sikt provides in Norway are founded on the principles of open science, and there has always been a close relationship between the Norwegian national CRIS-system and open access. The newly launched NVA has recently replaced the former system Cristin (Current Information System In Norway) where one of the predecessor's functions was to accommodate green open access, by letting the users use Cristin as an 'upload-hub' for fulltexts and populate institutional repositories with metadata from the CRIS [4]. The system has recently been completely rewritten technically and was relaunched in 2025. The new system is strengthening the bonds to open science even more by merging the former functions of Cristin with repository-functionalities in a single system covering 160 research institutions and approximately 80 institutional repositories in a consolidated system covering publications from 2011 to date amounting to ~400 000 scientific records.

Open science, with its many branches, is also covering research information through the Barcelona-declaration on Open Research Information [5]. The declaration states that research evaluations and policy decisions should be made based on transparent evidence that should be accessible by those being assessed. While the declaration, also signed by euroCRIS, covers CRIS-systems, its main aim is to address the use of proprietary bibliometric sources such as for example Scopus and Web of Science (WoS). One source that aligns with the principles of the Barcelona-declaration is OpenAlex, a promising source of bibliometric data, based on the now deprecated Microsoft Academic Graph, which is maintained by OurResearch and published under an open and free license and therefore free to download and use for everyone [6].

OpenAlex has recently been included in the portfolio of NIB, which is a Norwegian infrastructure launched in 2019 to provide bibliometric data to Norwegian research institutions. Until 2025, NIB

was limited to WoS-data only, which was offered to members of NVA exclusively, even then with limitations. While OpenAlex currently only is available in NIB as a test version, it has become an important asset in a strategic move for NIB as an open science-infrastructure. NIB is currently offering both WoS and OpenAlex as a part of a multi-source-strategy, where the same analysis in principle can be performed based on two different sources, and thus give two different perspectives on the same phenomena.

3 The potential in a partnership between NIB and NVA

As a rule, every bibliometric analysis is influenced by the source of data, and consequently, no single source is able to represent a complete and accurate picture. For example, WoS is indexing a limited number of journals, and is further lacking many non-English journals and particularly research output in the social sciences and humanities [7]. While being a promising alternative to proprietary and expensive sources, OpenAlex is still struggling with data quality issues e.g. misclassified document types, which is important since letters to editor, book reviews and other material may be classified as regular research articles and thus skew analyses [8]. Both sources are also known for issues with affiliations and addresses [9], although the curated WoS in at least one study is considered stronger than OpenAlex[10]. Even with limitations, both sources support a wide range of bibliometric analyses, and access to comparative analyses using multiple sources may help improve the quality and remedy some of the shortcomings of relying only on a single source. Nevertheless, Sikt will direct efforts into making OpenAlex the main feature of NIB.

CRIS-systems, especially national ones like NVA can also function as a data-source for a wide range of bibliometric analysis, but the main benefit in bibliometrics is that NVA holds solid data on affiliations and institutional addresses. This is due to NVA's function as an institutional reporting tool for production levels to the ministry of Education and Research. This reporting, which is mandatory for all publicly financed research institutions in Norway, creates an incentive for the institutions to maintain accuracy in affiliation data for their own publications.

Some of the limitations in NIB can thus be remedied by joining data from both systems, particularly utilizing high-quality affiliations in NVA, making it possible to know the exact baseline number of records that belong to an institutional publication portfolio. A bibliometric analysis based on WoS or OpenAlex can then be validated against a verified number of records that optimally should be part of the analysis. The number of records that goes into an analysis is also increased, since relying on institutional affiliation in WoS or OpenAlex alone, usually does not catch all relevant items.

4 Bibliometrics; a challenging affair

There are a vast number of challenges that lies ahead when doing bibliometric analyses [11]. While bibliometrics is a research discipline in its own right, the average user of NIB is typically performing tasks in the administrative leg of their institution, using analyses in internal affairs on site. Staff with such responsibilities often have bibliometric tasks in addition to other tasks. There are further technical barriers as NIB currently only offers rudimentary access through SQL for fetching datasets and further analysis. This, together with complicated data models and a lack of off-the-shelf analytical tool translates to a significant hurdle for usage. Even so, the main challenge for the average NIB-user is perhaps not a technical one; bibliometrics is complicated and requires domain knowledge, methodological know-how and a basic understanding of statistics. To help lower the bar, NIB pre-

calculates several relevant indicators, including field-normalized impact and percentiles, which are recalculated monthly when data is updated. This streamlining also ensures that different users are using the same standardized and consistent measures in their analyses. Even so, the whole process remains a bumpy road for users lacking technical skills and bibliometric know-how.

Herein lies the real motivation for the pilot; if NIB shall come to use as intended, NIB needs to lower the user-threshold and facilitate easier access to both data and analyses. That means covering the whole process from harvesting data to conducting an analysis and making the output available in some usable form.

5 The analytics pilot: From data to analysis

The challenges of retrieving data from complicated data models and conducting bibliometric analyses, led Sikt to experiment with an analytics pilot in late 2025. The stated goal of the pilot was to investigate the possibilities of simplifying and streamlining the process from data to mainstream analyses. An important idea was to make use of NVA's enhanced affiliation data to improve accuracy of datasets for a single institution, which is the main concern of NIB users. The scope of the pilot thus went beyond merely combining datasets, but to make the whole process of analyses easier.

R was chosen for developing the pilot, as it is a versatile and generic open-source tool focused on statistics and analytics, which also is popular choice in bibliometric analyses. Additionally, R offers a range of downloadable free bibliometric libraries that may come to use at a later stage. Even so, R is by no means the simplest of programs for non-experts, so it was important to try to reduce the technical steps as much as possible by making the scripts executable without the need for actual programming skills or much prior knowledge of R. Besides installing necessary software, the script requires a minimum of mandatory steps (e.g. authentication and set up).

The main components of a bibliometric analytics pilot such as this one can be summarized in three steps that will briefly be explained. These are 1) match NVA-data with NIB-data, 2) simplify the data model for the two NIB-sources and 3) finale preparation of datasets for analyses.

The main key is to synchronize NVA and NIB, matching up front as many records as possible between the systems, so when the pilot fetches data from NIB, it also retrieves all keys used for merging data with NVA. The matching routine, which is performed monthly when NIB-data is updated, is currently based on DOI-matching. It includes data-enhancement and cleaning DOIs in both NVA and the NIB- databases with regex-expressions. In case a DOI is missing on a record in NVA, efforts are made to identify DOIs by querying Crossref. The overall matching score is relatively successful, with approximately 94 % match rate for recent core-publications (defined as journal articles reported to the ministry the previous three years) in OpenAlex, and similarly 84 % in WoS. A benefit that will have impact on the validity of analyses, is that a user will have information of how many publications are matched and thus how many are excluded from a specific analysis.

Both OpenAlex and WoS consist of data stored in 50-70 tables each, which all are related to one another in complicated data models. Navigating these models can be challenging even for experts, so developing the pilot also included efforts to simplify data through materialized views (MV). MVs are specific, readymade aggregations of tables in a database based on an underlying SQL-query, effectively joining and limiting data from many tables into a single table designed for a specific need. This makes it easier to both control the format of the output and retrieve data with a simple query, without the need to know or understand the underlying schema-structure.

Finally, a user can further limit the dataset to a specific Norwegian institution and/or time-period, and optionally in combination with a certain field of research. This is set as parameters by the user in

the pilot, whereafter the user can either choose to export data for further exploration in a separate program or continue to use the built-in-analysis in the pilot.

The pilot currently offers nine demo-analytics for a dataset combined from NVA and WoS, and seven for a dataset combined from NVA and OpenAlex. These range from simple inventories of production, to most published in journal or by author, or to the mean number of citations within each open access-flavor and percentiles-calculations. All tables and plots are exported as separate files, and collectively in an excel spreadsheet and a word-formatted report for further tweaking. The analyses are not yet verified and are currently in beta.

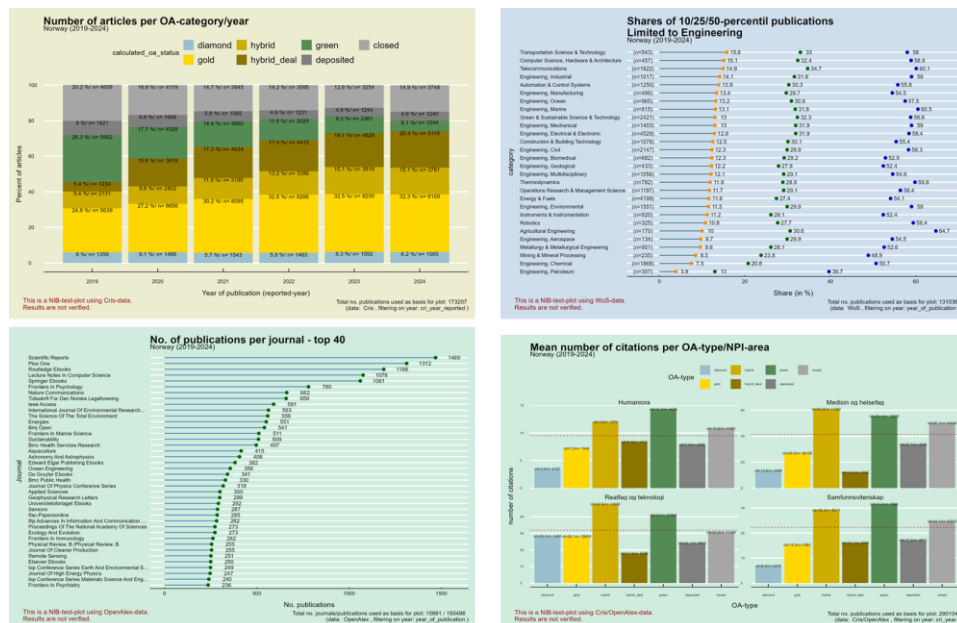


Figure 1: Sample of demo analyses.

Top left – Open access-distribution in dataset (CRIS-data only).

Top right – Percentile-distribution in dataset (WoS-data).

Bottom left – Number of publications in top 40 journals in dataset (OpenAlex-data).

Bottom right – Mean number of citations per open access-category in dataset (OpenAlex-data).

While the intent of the pilot is to make bibliometric analyses more manageable by streamlining the components in the analysis, the procedure has the additional benefit that outcomes are comparable between institutions. All indicators are calculated centrally, the selection of data is done by the same rules, the individual analysis is calculated and presented in overall the same manner instead of multiple analyses that all are using slightly different methodologies and implementations.

Future development includes upgrading the pilot to a production-ready application with verified analyses, develop the application into a framework where users are invited to contribute with new types of analyses (as separate components), and with a slightly longer horizon, implement a web-hosted service, where analyses can be run in the browser without any need for local software. Whatever direction development will take, it will be decided together with the Norwegian bibliometric community.

6 Discussion and conclusion

The pilot was made available for Norwegian institutions early in 2026 and was met with both curiosity, encouragement and interest. While it is too early to pass an evaluation on the pilot at this point, it seems clear from a purely theoretical perspective alone, that bibliometric procedures can benefit from using structured CRIS-data to gain more accuracy in analyses. The procedural approach of the pilot is also promising since all institutions perform analyses in exactly the same way and on the same (baseline) datasets, analyses will then have comparable outputs when used in reports.

The pilot does however come with a few caveats concerning responsible use and application of metrics, these challenges are presumably present with respect to any tool producing ready-to-use analytic output.

Bibliometrics is rightfully regarded as a field that requires steady hands and sober approaches. A remaining challenge with bibliometrics, even when applied methodologically correctly, is that any representation of, or judgment passed on research, rarely is expressed correctly with numbers alone. There are numerous examples where bibliometrics have been gamed to present a more flattering picture than warranted, for example by inflated self-citations [12]. Other dubious conduct often involves the use of impact factor as a surrogate for quality (see for example [13]) or comparing apples with oranges, e.g. medicine with humanities, but even if we disregard gamification or deliberate misrepresentation, an application that makes bibliometric numbers and analyses easily available, also makes it easy to churn out numbers that may misrepresent or over-simplify complex issues in a given context. This is partly what the Leiden-manifesto warns about, and there is simply no way a technical application can safeguard against irresponsible use of bibliometrics.

Further; responsible metrics also state that quantitative indicators should act in support of qualitative evaluations and not play the main role alone, but this is also out of scope for a technical application. Numbers and analyses that easily can be ready in less than a few hours, are much more available than the potentially time-consuming process of qualitative evaluations. Even with these limitations in the applicability of bibliometric analyses, they are nevertheless tempting as shortcuts.

Sikt aims to give users, and the bibliometric community, great influence on the development of NIB in the future. This includes the pilot, which hopefully will be developed into a framework where new functionality, improvements and analyses are regularly added and expanded. This should be done with an awareness against going in a direction where it becomes tempting to skip qualitative steps and instead direct analyses to be supportive of such. However, anyone involved in application development know that such frameworks are notoriously difficult to maintain and keep alive, so in the end, the fate and future of the analytics pilot is wholly dependent on the bibliometric community's own needs and decisions.

References

- [1] R. K. Merton, ‘The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property’, *Isis*, vol. 79, no. 4, pp. 606–623, Dec. 1988, doi: 10.1086/354848.
- [2] D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, and I. Ràfols, ‘Bibliometrics: The Leiden Manifesto for research metrics’, *Nature News*, vol. 520, no. 7548, p. 429, Apr. 2015, doi: 10.1038/520429a.
- [3] ‘Open Science | OA-barometeret’. Accessed: May 28, 2026. [Online]. Available: <https://www.openscience.no/oa-barometer>
- [4] L. Wenaas, N. Karlstrøm, and T. Vatnan, ‘From a national CRIS along the road to Green Open Access – and back again: Building infrastructure from CRISin to Institutional Repositories in Norway’, *euroCRIS*, Jun. 2012. Accessed: Feb. 17, 2026. [Online]. Available: <https://dspacecris.eurocris.org/handle/11366/116>
- [5] ‘Barcelona Declaration on Open Research Information’, Barcelona Declaration on Open Research Information. Accessed: Feb. 17, 2026. [Online]. Available: <https://barcelona-declaration.org/>
- [6] ‘OpenAlex: The open catalog to the global research system | OpenAlex’. Accessed: Feb. 17, 2026. [Online]. Available: <https://openalex.org/>
- [7] D. Chavarro, I. Ràfols, and P. Tang, ‘To what extent is inclusion in the Web of Science an indicator of journal “quality”?’’, *Research Evaluation*, vol. 27, no. 2, pp. 106–118, Apr. 2018, doi: 10.1093/reseval/rvy001.
- [8] N. Haupka, J. H. Culbert, A. Schniedermann, N. Jahn, and P. Mayr, ‘Analysis of the publication and document types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar’, *Quantitative Science Studies*, vol. 7, pp. 179–194, Mar. 2026, doi: 10.1162/QSS.a.406.
- [9] C. Lenke and N. Taubert, ‘Address Information in OpenAlex, Web of Science, and Scopus: First Insights’. Accessed: May 28, 2026. [Online]. Available: <https://www.open-bibliometrics.de/posts/20250923-AddressInsights/>
- [10] L. Zhang, Z. Cao, Y. Shang, G. Sivertsen, and Y. Huang, ‘Missing institutions in OpenAlex: possible reasons, implications, and solutions’, *Scientometrics*, vol. 129, no. 10, pp. 5869–5891, Oct. 2024, doi: 10.1007/s11192-023-04923-y.
- [11] A. Ferrara and S. Salini, ‘Ten challenges in modeling bibliographic data for bibliometric analysis’, *Scientometrics*, vol. 93, no. 3, pp. 765–785, Dec. 2012, doi: 10.1007/s11192-012-0810-x.
- [12] R. V. Noorden and D. S. Chawla, ‘Hundreds of extreme self-citing scientists revealed in new database’, *Nature*, Aug. 2019, doi: 10.1038/d41586-019-02479-7.
- [13] A. Casadevall and F. C. Fang, ‘Causes for the Persistence of Impact Factor Mania’, *mBio*, vol. 5, no. 2, pp. e00064-14, May 2014, doi: 10.1128/mBio.00064-14.