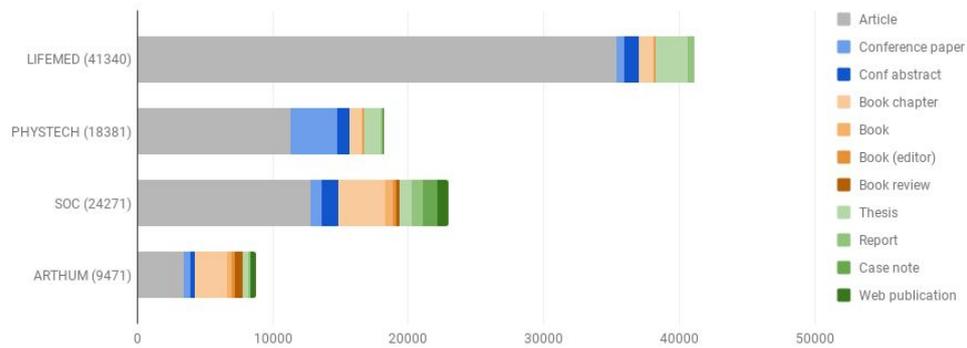
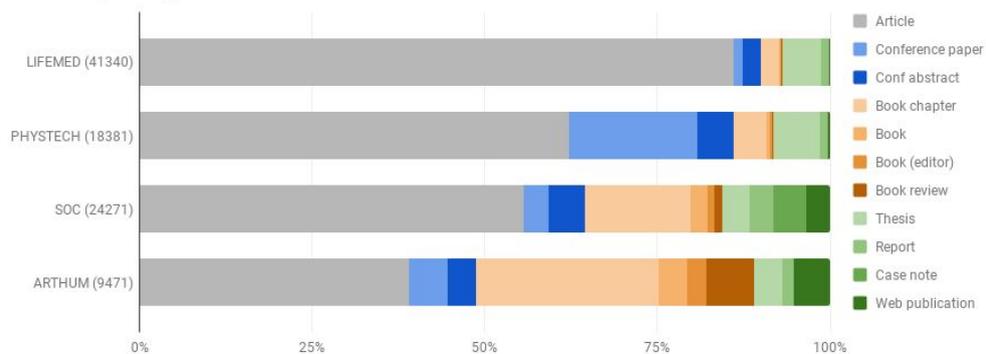


Publication cultures and Dutch research output: a quantitative assessment

KUOZ data - absolute numbers



KUOZ data - percentages



Jeroen Bosman and Bianca Kramer
Utrecht University Library

for The Association of Universities in the Netherlands (VSNU)

Publication cultures and Dutch research output: a quantitative assessment

A note on data availability.

Data underlying this report has been shared as far as possible and allowed.
At the time of publication, this excludes data from the Utrecht University / UMC Utrecht CRIS.

Utrecht, april 2019

Jeroen Bosman  [0000-0001-5796-2727](https://orcid.org/0000-0001-5796-2727)

Bianca Kramer  [0000-0002-5965-6560](https://orcid.org/0000-0002-5965-6560)

Utrecht University Library
for The Association of Universities in the Netherlands (VSNU)

report available at: <https://doi.org/10.5281/zenodo.2643360>

data available at: <https://doi.org/10.5281/zenodo.2643367>



Table of contents

Summary	3
Elements in this study	5
1. Introduction	6
Open access policies and non-journal output	6
Publication type choices as a reflection of publication cultures	6
The present study - starting with KUOZ data	8
2. Quantitative assessment of publication types by main field	13
Using multidisciplinary citation databases	13
Distribution of publication types in the different databases	14
Distribution of publication types in main academic fields	16
Overlap between databases (publications with DOI)	22
Case study: Utrecht University CRIS output (2017)	24
Using additional databases	30
Books, a special case	31
3. Open Access levels	33
Open access in national reporting	33
Open access levels of content covered in Web of Science, Scopus and Dimensions	33
Open access levels of content covered in additional databases	35
License types of OA output, by field	36
Open access books	39
4. Additional databases for filling gaps - an assessment	41
Coverage and filtering	41
Completeness (share of output captured)	42
Reusability of data, data license	43
Ease of mapping of fields	43
Solving issues of additional databases	43
5. Conclusions and recommendations	45
References	48

Summary

Lack of insight in publication cultures: a problem for open access policies?

Open access availability of research outcomes in the Netherlands has increased significantly over the last few years. That rising level of open access is evident for peer reviewed journal articles, supported by arrangements with publishers, direct financial support of authors, and promotion of open access publishing in journals, together with sharing versions of those articles in repositories. This focus on journals and on articles raises the question whether we have the full picture to inform open access policies. We lack insight in the importance of non-article output, and that may even be more marked in specific disciplines.

Goals: describe publication culture variety and its effects, make recommendations

The goals of this study are to describe the disciplinary variety in publication cultures, to analyse its effects on comprehensiveness and bias in reporting open access levels and to present recommendations for various stakeholders in the research ecosystem. More concretely we look at these questions:

What are the different publication types of recent Dutch university output, what are levels of open access availability and what are the pros and cons of the various methods for tracking those?

For the Association of Universities in the Netherlands (VSNU) the answers to these questions should be valuable for their policies in monitoring and advancing open access and making sure these account for issues and opportunities around open access in all fields of research.

Variety in current research output

Dutch universities generate an immense variety of research output. The composition of output, as measured by counting publications, varies significantly between academic fields and thus also between institutions. Much but not all of the university output is registered in research information systems and publication numbers are aggregated at national level in KUOZ data by the VSNU. For all four main fields reported on here, output beyond journal articles is significant. For Social sciences and Arts/Humanities in particular (with over 40% and over 60% of output respectively not being regular journal articles) looking at journal articles only ignores a significant share of their contribution to research and society. It is also in these two main fields especially that the registered output includes substantial shares of popular and professional publications.

Inclusion of the variety of publications in bibliometric tools and analyses

Bibliometric analyses are dependent on databases with publication information. There is severe bias in many of these databases, with more comprehensive coverage for some publication types than for other and, partly by that, also better coverage for some fields than for others. The three main citation databases (Web of Science, Scopus, Dimensions) that allow analyses by affiliation/institution do each cover some variety of publication types (including conference material and book chapters), but this is still relatively little compared to

the variety of what is being published. Some additional databases (especially NARCIS and BASE, which both aggregate information from institutional repositories) add value because they are more inclusive. However, they have limitations in disaggregation by field or institution, and they also have incomplete or less consistent data.

In depth analysis of disaggregated data from the Utrecht University CRIS shows that looking just at the three main citation databases leaves out very substantial shares of research output. The share of Utrecht University output that is not detected by Web of Science, Scopus or Dimensions is predominantly non-article output, though in Social sciences and Arts/Humanities just using those databases also leaves out very substantial shares of journal articles.

Open access of different publication types

Determining open access levels and developments is complex, due to the many forms of open access and available methods to detect open access availability of a publication. The most efficient way to detect open access uses digital object identifiers (DOIs) and uses Unpaywall, a database with information on indications of open access availability based on DOI input. Detecting open access of non-article output is difficult, because the publication types often lack DOIs, or are not covered in databases that allow export of these DOIs. Some additional databases, especially those that harvest repositories (like NARCIS, BASE and OpenAIRE) include open access information provided in the metadata of the repository content they harvest. The value of this information is thus in large part dependent on the quality of repository metadata.

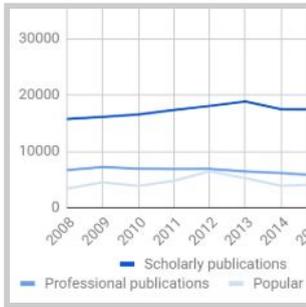
Licenses chosen for open access sharing of publications are also part of publication culture and show clear differences between fields. That is most clear in the relatively high usage of NC (non-commercial) and ND (no derivatives) clauses in Creative Commons licenses used in full gold journals in Arts/Humanities.

Recommendations for various stakeholders

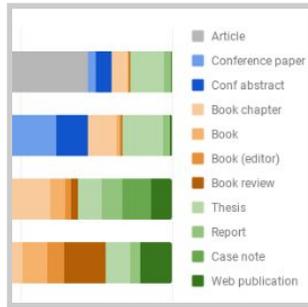
This study confirms that because of large differences in publication culture between disciplines, current approaches to tracking open access levels are too limited. It also shows that for more inclusive monitoring, improvements in coverage and detection are necessary. We offer various recommendations for authors, publishers, institutions, database providers, aggregators and open access detection tools to increase the possibilities to get a comprehensive overview of publication types and open access levels thereof in all disciplines. The overall recommendation is to increase the use and open availability of metadata and step up usage of permanent identifiers (at least for authors, institutions, funders and publications). More inclusive reporting is possible when, apart from these, institutions are as comprehensive as possible in registering output in their CRIS and making sure that all metadata are in their repository, with an open license on those metadata. The latter also holds for aggregators of repositories (e.g. NARCIS). Database providers could make steps in offering richer export/downloading options with as few restrictions as possible on the usage of metadata. Finally it would be welcomed if open access detection tools took steps to facilitate detection of OA for non-article output.

With commitment and actions from various stakeholders, it would be possible to create a more fair, accurate and nuanced insight of open access developments across fields.

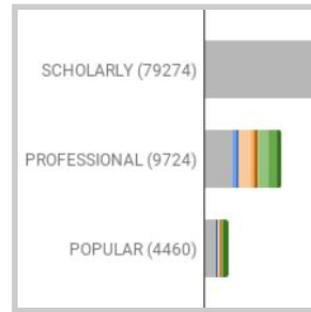
Elements in this study



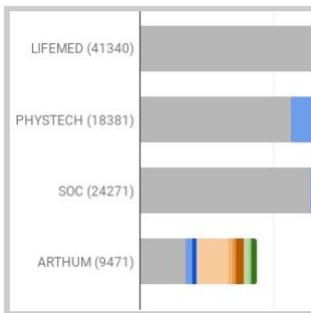
Publication culture ▶▶



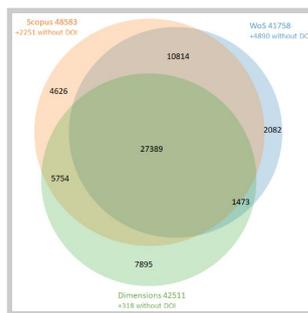
Publication types ▶▶▶



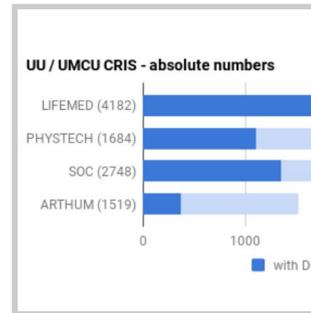
KUOZ data ▶▶



Academic fields ▶▶



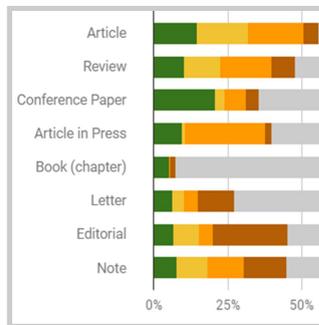
Citation databases ▶▶▶



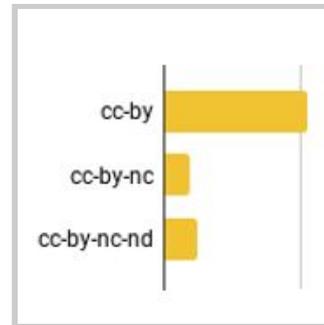
Case study: CRIS ▶▶



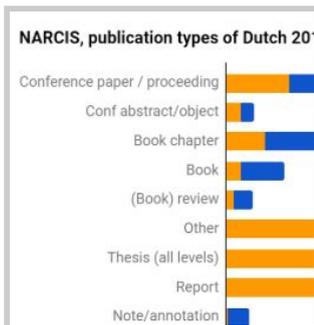
Case study: Books ▶▶



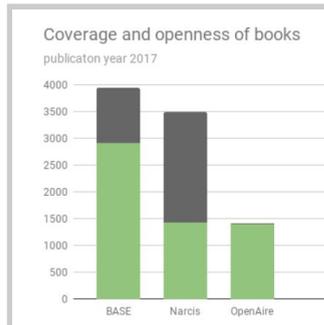
Types of OA ▶▶



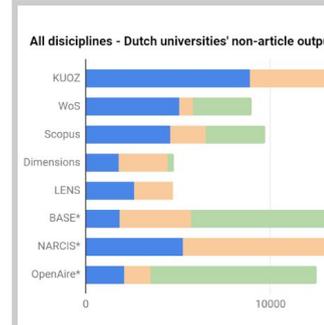
Licenses ▶▶



Other databases ▶▶▶



Case study: Books (OA) ▶▶



Filling the gaps ? ▶▶

1. Introduction

Open access policies and non-journal output

For a number of years, VSNU has been pursuing open access policies in order to promote availability and usage of scholarly output of its member institutions. Focus has been on gold (publisher provided) and to a lesser extent green (self archiving) routes to open access. Attention has been mainly on open access for journal articles, because that constitutes the largest amount of publications and because facilitation by publishers of open access for journal articles - in terms of business models, policies and infrastructure - is further advanced than for other types of output.

There remain substantial challenges in further advancing open access for journal articles - new requirements from funders (Plan S) not being the least of them - but VSNU feels there is a need to look at open access for other types of output as well, for two reasons. First, there is a lot of potential in non-journal article output that is important to have openly available for scholarly and non-scholarly usage. Secondly, with uneven distribution of types of output in the various disciplinary fields, some fields may indirectly have been disadvantaged and lacked full support for opening up their specific types of output.

To support choices in moving towards supporting the transition of other types of output towards open access, the VSNU want to first gain insight in who is publishing what: what is the usage of the various publication types in the main disciplinary fields of academia? Additionally there is a need to find out what are current ways to assess what proportion of these publication types is open access. To provide those insights, VSNU committed a study into the quantitative aspects of publication cultures at Dutch universities. Utrecht University Library has conducted this study.

Publication type choices as a reflection of publication cultures

The types of publications used in different fields are a reflection of what many call a field's publication culture, or broader even: cultures of scholarly communication. These can be seen as sets of needs, norms and practices in publication and communication of research that develop through time in the various fields. They are a complex interplay between types of research, types of research objects and their representation in data/text/imagery, types of authorship and collaborations, types of funding, degrees of formalization of norms, language, citation and writing styles, and types of discourse and evidence. They determine which specific channels/media/document types authors choose and how they use them.

Publication cultures are perhaps best considered as a subset of research cultures. The latter have been detailed for the Netherlands in the context of research data management (KNAW 2013, p.19).

The discussion around variety and change in publication cultures is not only relevant for open access policies. It plays a role in discussions on societal versus scholarly impact and in interdisciplinary research projects (De Jonge Akademie 2012). It is perhaps even more debated in the context of researcher status and evaluation of grant proposals. To keep criteria for assessment of publication lists in grant decisions up to date NWO (2013, 2016,

translated in 2018) commissioned studies on publication cultures in humanities and Social sciences in which the main aspects of publication cultures on which the (sub)fields vary were publication types, authorship and, for humanities, language.

Number of co-authors
Author order on co-authored publications
Choice of publication types and their perceived value
Languages used in publications
Audiences one tries to reach with various types of publications (scholarly, professional, popular)
Licenses chosen for open access publications
Citation styles
Shelf life of publications
Importance of journal hierarchy or publisher brand
Peer review: single, double, triple blind, or open identities
Acceptance of publishing before peer review in e.g. preprints or working papers

Table 1. Some examples of aspects of disciplinary publication cultures

Of course even if the relative popularity of document types in different fields is the same, that does not mean that the way the publication is used and valued is the also the same. If a cardiologist says “I wrote a book on heart failure”, that might mean a complete different thing than if an historian says “I wrote a book on interpretations of the Srebrenica genocide”. Even though the format is the same - a book - the way the format is used and its standing and role in the respective publication cultures can vary strongly. Also, publication cultures are in constant - albeit relatively slow - development. The mere fact that style manuals (e.g. the Chicago Style manual or the APA publication manual) have updated editions every few years indicates that publication norms and practices are constantly changing.

Even if in a certain field the mix of publication types does not change, that does not mean that the publication culture dynamic does not change *how* authors use the various document types. For instance, the share of books in the total number of publications in a field can remain constant while the language used therein, the number of co-authors, the citation styles and types of audiences targeted with those books change.

This study is restricted to publication types, even though these are but one aspect of publication culture and that in turn is but one aspect of research communication. Next to practical limitations of time and data availability, the main reason is that the publication type aspect of publication cultures is arguably the most relevant for open access policies. Still, the restriction clearly is a limitation that needs to be kept in mind. Further study of, for instance, the relevance of language orientation and author numbers in disciplinary publication cultures may at some point be warranted. This study uses the term publication types both for what are considered publication types (book, journal) and what are actually document types, that can be published in certain publication types (e.g. in journals: letter, article, review; e.g. in

books: thesis, monograph). The data underlying figures and calculations in this study are available on Zenodo (Kramer and Bosman, 2019).

The present study - starting with KUOZ data

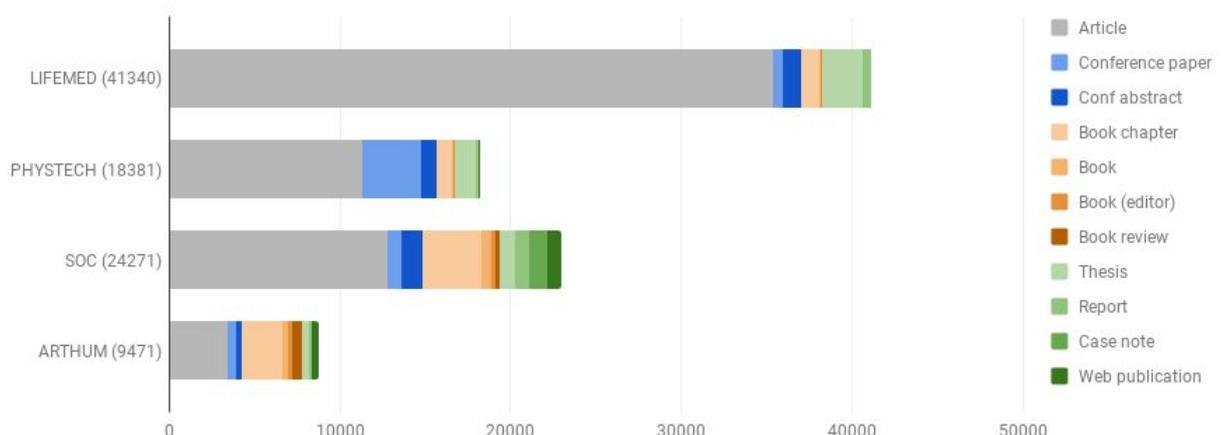
This brief report delivers the results of this study. It is accompanied by the data that have been collected. The main research question is:

What are the different publication types of recent Dutch university output, what are levels of open access availability and what are the pros and cons of the various methods for tracking those?

The study has been limited on purpose because of the limited amount of time available. For that reason, historical developments and international developments are out of scope and there has been only a quite limited study of the literature. The study is descriptive in nature and only hints at possible explanations of patterns found. It includes recommendations for VSNU on how to best track usage and open access levels of various types of publications.

Starting point for the present study is the so-called KUOZ data that has been collected for VSNU by the institutions for a number of years. These provide insight into publication volumes by year, type, institution and field, with the latter operationalized as 8 so-called *HOOP (Higher Education and Research Plan)*-areas. KUOZ data are generated from the universities' current research information systems (CRIS). They are based on VSNU-agreed standards for labeling publications in the so-called 'Definitieafspraken wetenschappelijk onderzoek' (VSNU 2018, p.17). Figures 1 and 2 summarize the main KUOZ data.

KUOZ data - absolute numbers



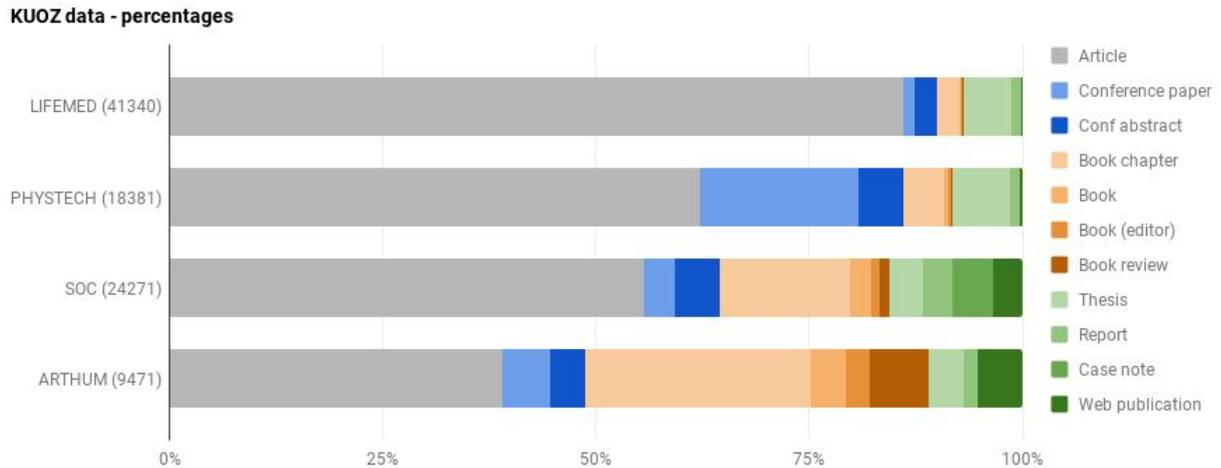


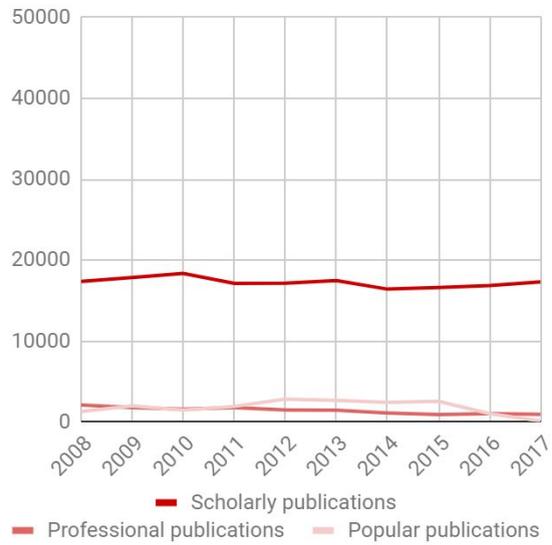
Figure 1. Publication types in the four main academic fields, 2017. Source: VSNU, KUOZ data 2017.

It is clear that journal articles are extremely important but not in any way the only relevant output type. This confirms that any study ignoring non-article output is limited. Though we might say that scholarly communication in life and medical sciences is dominated by the journal article, other fields are characterised by a substantial and even majority role for non-journal output. Naturally in all fields, doctoral theses are important, though admittedly in Life sciences/Medicine and Physical sciences/Technology these are often comprised of journal articles. The specific nature of some subfields causes specific publication types to show up here. In the Physical sciences/Technology field we see conference proceedings (from e.g. computer science and engineering) and in the Social sciences field we see case notes (from the law subfield) and quite some report publications (likely important in e.g. sociology, economics and again law). The Social sciences and Arts/Humanities report substantial numbers of web publications. These can vary from regular blog posts, opinion pieces in online publications, to toolkits and dossiers. But of course the most striking non-article output is anything appearing as book, part of a book or as review of a book. It is part of the life blood of humanities, but also still very important in the Social sciences (perhaps less so in the behavioural part of Social sciences, though that is not looked into here). It is interesting to see that the relative shares of full books versus book chapters does not differ much between Social sciences and humanities.

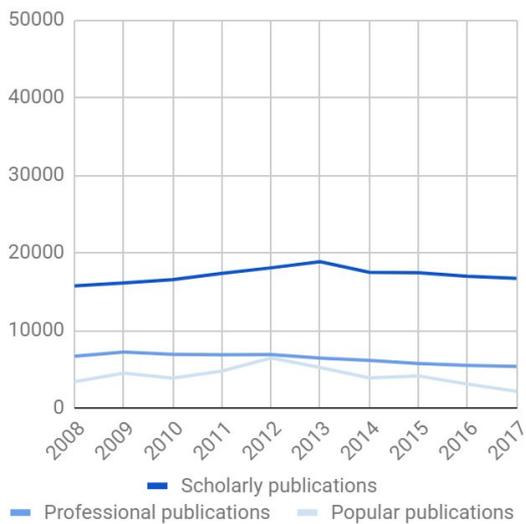
Life sciences & Biomedicine



Physical sciences & Technology



Social sciences



Arts & Humanities

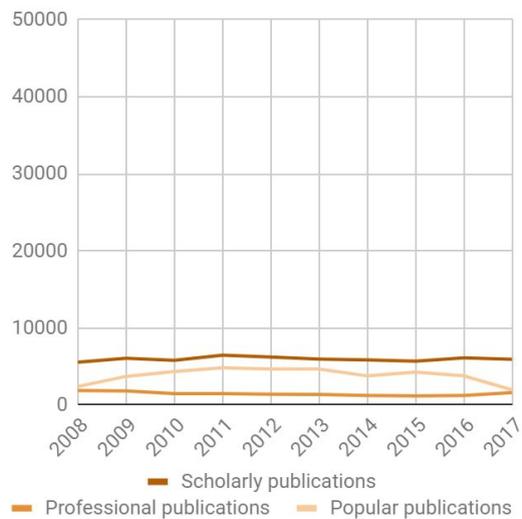


Figure 2. Scholarly, professional and popular publications. Source: VSNU, KUOZ data - longitudinal.

Apart from their composition by publication type, the products of academics in the various fields also differ in the composition by audience type, discerning between pure scholarly, professional and popular work. Whereas in Life sciences/Medicine and Physical sciences/Technology, reported professional and popular output is limited, it is substantial in Social sciences and especially humanities, in a relative as well as absolute sense. It must be noted that the value attached to these various types of output vary. In a certain sense a successful high quality history book oriented at the broader popular market can be just as important and valued as highly as a pure scholarly monograph. In the same way an influential report advising the government on e.g. educational policies can be valued highly in the Social sciences. There are incentives for researchers and institutions to publish in pure scholarly/scientific channels and at the same time there are incentives to invest in public engagement, for instance through popular publications. It is remarkable that popular publications are on the decline in the last 4 or 5 years, in all fields. Is this a reflection of

stronger incentives for pure scholarly publications? And does this mean public engagement loses out as a consequence, or is public engagement increasingly taking forms not captured by output reported in the current information systems (CRIS) of our institutions? Professional publications also show decline, but that decline is slower and more long term than that of popular publications.

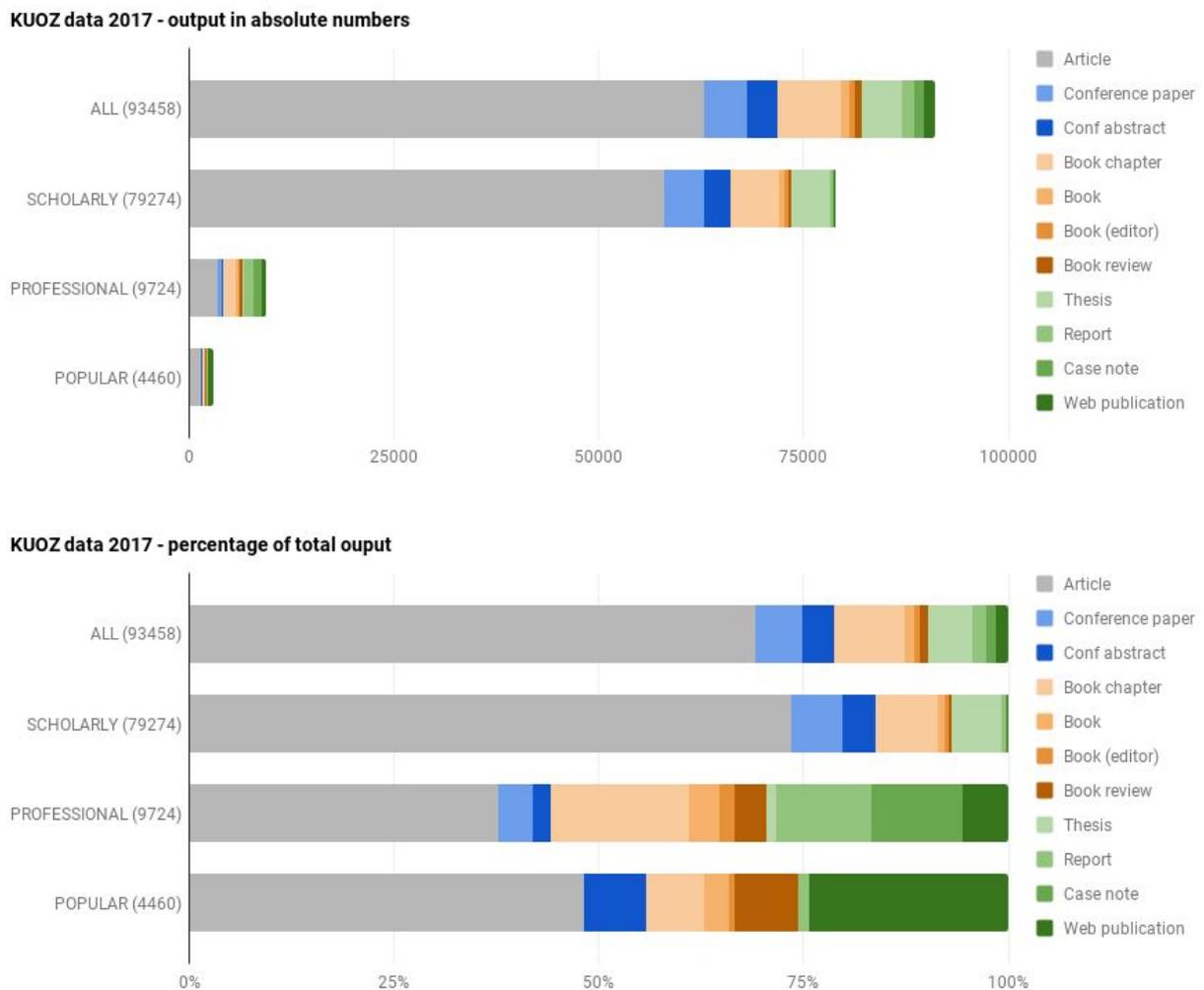


Figure 3. Scholarly, professional and popular publications, by publication type, 2017. Source: VSNU, KUOZ data 2017.

The use of various types of publication is not independent from the split by target audiences (fig. 3). For professional and popular audiences the journal article is substantially less dominant than for scholarly peers, with greater roles for books, reports and case notes (for professional audiences) and web publications, especially for popular audiences. Whereas the latter may be expected to be mostly openly accessible by their nature, that may not always be true in the same way for reports, case notes and notably, books. For web publications in particular there is also the issue of sustainability and archiving of the output. Overall we can say that for *all* fields and for all types of audiences focussing on just journals and on just scholarly output delivers a reduced view on what is going on and what is valuable to peers and society at large.

Though the KUOZ data from VSNU itself is valuable for reference and as a first source of information to create an overview, it remains aggregate data that cannot be broken down further. Thus, it can not be used to assess open access levels, because it lacks information on individual publications. The CRIS data of the separate institutions do provide more detail but are not openly available and based on partly unknown and possibly diverging registration practices.

To gain more insight this report first explores data from large multidisciplinary citation databases in chapter 2, to provide detailed breakdowns by field and publication type and compare these to the KUOZ data. That is followed in chapter 3 by an analysis determining open access levels. Using additional databases to get reliable and comparable data for non-journal output is a challenge in itself that is reported on in chapter 4. As a case study and to corroborate findings based on public and proprietary databases this study also includes more detailed CRIS data from Utrecht University. In chapter 5, general conclusions and recommendations are given.

2. Quantitative assessment of publication types by main field

We assessed the differences in publication types in the four main fields by looking at the extent to which different publication types are represented in a number of multidisciplinary databases (as well as in a few other sources). It is important to realize that the coverage of publication types in any database represents not only the production of those publication types, but also their inclusion in those database(s). Therefore we also looked at the extent to which the proportional representation of publication types in multidisciplinary databases matches the self-documented research output by universities (by looking at the KUOZ data and, at a more fine-grained level, the data from one university's CRIS).

Using multidisciplinary citation databases

We chose to use three large multidisciplinary citation databases: Web of Science, Scopus and Dimensions, as these are regularly used for comparative bibliometric evaluations and most Dutch universities have access to at least one of them. All three databases are multidisciplinary, cover multiple publication types and offer fine-grained search and retrieval functionalities such as harmonized field search for organizations, that allows for unambiguous identification. They are also all closed-access, requiring a paid license to access their full content and functionality, and limiting open sharing of the data derived. Through our university's licenses, we accessed Web of Science (without the Book Citation Index, that has to be acquired separately) and Scopus. We also had access to the full version of Dimensions during the period of this study.

Publication types

Web of Science, Scopus and Dimensions each cover a variety of publications, and allow detection of a number of publication types through search filter options or retrievable metadata. A complete overview is provided in the dataset accompanying this report (Kramer and Bosman, 2019). Web of Science and Scopus both allow for detection of *articles*, *books/book chapters*, *conference papers/proceedings*, *letters*, and *reviews*. Each also have a number of publication types unique to the database. In general, Web of Science distinguishes more document types (42) than Scopus (16). At the time of analysis, detection in Dimensions was limited to 5 document types: *articles*, *chapters*, *monographs*, *proceedings* and *preprints* (since then, the category *books* has been added). It is important to realize that similar content can be labeled differently in different databases. For example, document type categories in Dimensions are broader than in Web of Science and Scopus: for articles, non-article journal material (editorials, letters etc) is included as well. In addition, even though document types are in principle covered and detectable, the degree of coverage may be very limited (as with books/book chapters in Web of Science, without the Book Citation Index) or biased towards certain fields (as with preprints in Dimensions, that at the time of sampling only harvested [biorXiv](#)).

For this study, we looked at the full output of Dutch universities from 2017, broken down by publication type as assigned by the respective databases.

Affiliations

To assess the representation of research output from individual organizations (such as Dutch universities), a database should allow for output detection at that level. Web of Science, Scopus and Dimensions all allow for searching/limiting results for specific organizations, and all use a form of harmonization to account for multiple name and spelling variants. Web of Science and Scopus use their own, proprietary, classifications, and Dimensions uses the openly available [GRID classification](#). The presence of a harmonized field tag does not guarantee complete retrieval of an institution's output in the database though, as this also depends on whether the tag is assigned to all output or not. Another point to consider is that since neither database has the possibility to limit to corresponding authors unambiguously (with Web of Science listing one or more corresponding authors, Scopus only one and Dimensions none at all), all output of an organization is included, irrespective of whether the corresponding author is from that organization or not. This is, however, in line with choices of what research output is reported in universities' own CRIS and in the KUOZ data.

For this study, we identified the output of 14 Dutch universities (including university medical centers) using their harmonized field names in each database. An overview of field names used is provided in the dataset accompanying this report (Kramer and Bosman, 2019).

Fields

Web of Science and Scopus each have their own classification for research fields, both at journal level. In both databases, a journal can belong to more than one field category. Dimensions uses the [ERA/ANZSRC Fields of Research](#) classification and applies machine learning to classify papers at article level (Herzog and Kierkegaard Lunn, 2018).

For this study, the 151 research areas in Web of Science, 27 major subject areas in Scopus and 22 2-digit fields of research codes in Dimensions were mapped to the main fields in Web of Science (Physical sciences, Technology, Life sciences/Medicine, Social sciences and Arts/Humanities), with Physical sciences and Technology combined, resulting in 4 main fields. These were used throughout the study. The full field mapping is included in the dataset accompanying this report (Kramer and Bosman, 2019).

In contrast to the journal- or article-based classification of fields in the three citation databases, the KUOZ data have been assigned fields based on Dutch universities' organizational structure. Publications are assigned to a specific HOOP-area based on the faculty/department(s) where the work was carried out. For this study, HOOP-areas were mapped to the main fields derived from Web of Science, with exception of the HOOP-area 'Nature', which encompasses both Physical sciences/Technology and Life sciences/Medicine, and was split equally among these two categories. This may not reflect the actual proportion of output from both fields in this HOOP-area, but because of the aggregated nature of the KUOZ data, it was not possible to make a more precise distinction.

Distribution of publication types in the different databases

Output from Dutch universities from 2017 was retrieved from Web of Science, Scopus and Dimensions in June/July 2018. For each database, the number of publications per document

type were analyzed for the 4 main fields used in this study and compared with the KUOZ data (fig. 4). In each database, the most frequently occurring publication types were included in the analysis, ensuring a concentration ratio of at least 90% in each field (meaning that the publication types included cover at least 90% of the output).

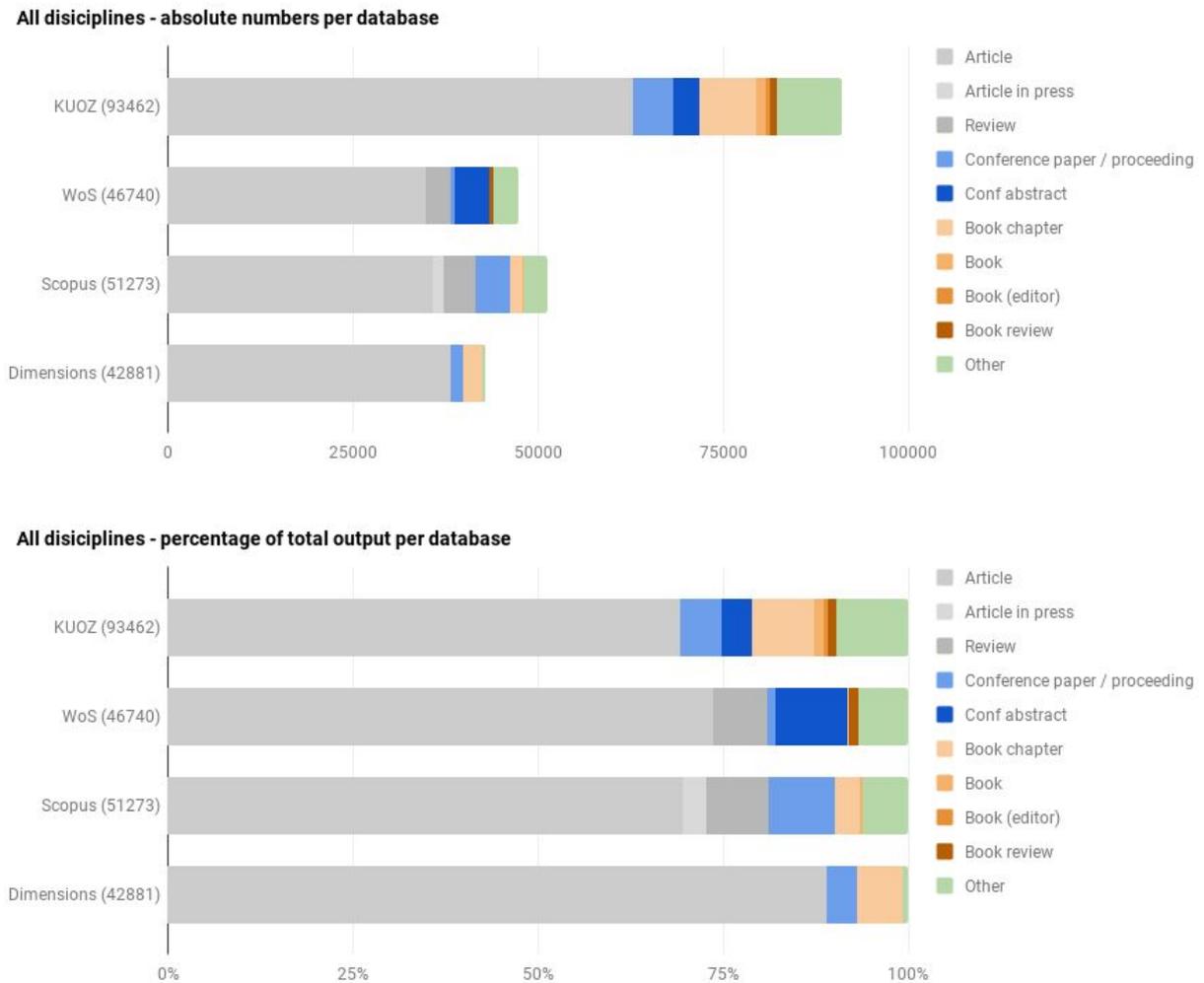


Figure 4. Total number of publications and shares of publication types, for each of three multidisciplinary databases, compared to KUOZ data. Sources: VSNU, KUOZ data 2017; Web of Science, Scopus, Dimensions.

The total number of publications in KUOZ data is almost twice as high as the number retrieved from each of the citation databases. The main reason for this is that the KUOZ data are a summation of the reported output by each university, without duplication. A rough estimation, comparing Web of Science data for articles & reviews from individual Dutch universities in 2017 (Kramer and Bosman, 2018) with data collected in this study indicates that about 20% of aggregated output are duplicates (46,895 articles & reviews in aggregated output vs. 38,264 in deduplicated output).

Another factor in the larger output as reported in the KUOZ data is the inclusion in KUOZ of non-scholarly publications (15% of the total output in KUOZ), which to a large extent are not included in the citation databases used in this study.

In comparing the proportional share of different publication types in KUOZ with that in the other databases, the factors mentioned above are important to keep in mind. If certain publication types (in general and/or in specific fields) are more often the result of collaboration between multiple institutions, these will be overrepresented in the KUOZ data. We have not tried to correct for this. As for the effect of non-scholarly publications in KUOZ, non-scholarly output has a larger proportion of non-article publications than scholarly output (see fig. 3). However, due to the relatively low numbers of non-scholarly output, the overall distribution of publication types does not differ greatly from that of scholarly output only. Two notable exceptions to this are case notes and web publications, which are almost exclusively considered professional and popular research output, respectively.

When comparing absolute numbers and proportional shares of publication types in KUOZ with what was retrieved from Web of Science, Scopus and Dimensions (fig. 4), it is clear that in all three databases, articles are by far the most common publication type. Beyond that, however, different databases have different 'strengths': while Web of Science is strong in conference abstracts (actually containing more than are included in the KUOZ data) and book reviews, Scopus has more conference proceedings, and a larger proportion of book chapters as well. Of the three citation databases, Dimensions has the most book chapters (both relatively and in absolute numbers), though still considerably less than are reported in the KUOZ data. Only Scopus contains a limited number of books, likely edited volumes of which chapters are included individually as well. Looking at 'other' publication types (non-article, non-book and non-conference output), editorials, letters and erratums are present in Web of Science and Scopus, while the former two are included in 'articles' in Dimensions. Theses, case notes, and web publications are present in the KUOZ data (the latter two mainly as non-scholarly publications), but not in either of the three other databases.

Distribution of publication types in main academic fields

What does the difference in coverage and retrieval of publication types mean for how the main fields are covered in the three citation databases studied? Or stated differently: how is the diversity in publication types (also known as bibliodiversity) in different fields reflected in these citation databases? For this, we looked at the different publication types retrieved from Web of Science, Scopus and Dimensions¹ for each field, and compared those to the output as reported in the KUOZ data.

In the sections below, we discuss the types and shares of non-article output per academic field in more detail. In all fields though, articles (including review articles) are the most numerous publication type in the KUOZ data as well as in the citation databases. For all fields except Physical sciences/Technology, the number of articles retrieved from the three citation databases is considerably less than the number reported in the KUOZ data, with the largest differences observed for Social sciences and Arts/Humanities. Interestingly, for Physical sciences/Technology, the number of articles retrieved in both Web of Science and Scopus is equal to or even exceeding the number reported in the KUOZ data. This could be partly be due to the way field assignment was done in the KUOZ data (with the HOOP area 'Nature' split evenly between Life sciences/Medicine and Physical sciences/Technology,

¹ Methodological note: in Dimensions, only 35599 publications (83%) had field classification(s) assigned - the percentages in the figures discussed here were calculated based on this number.

which might have resulted in a misclassification of articles and other publication types belonging to Physical sciences/Technology). Another possible explanation, esp. for Web of Science, is that e.g. conference proceedings in KUOZ might have been labeled as articles in the citation database(s), thereby increasing the share of articles. We have not checked these assumptions, so they remain hypotheses.

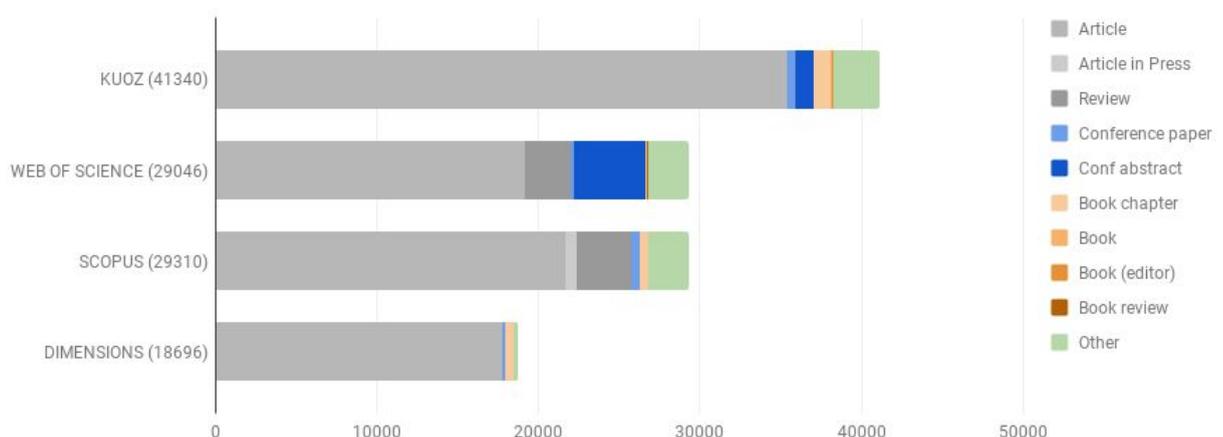
Life sciences/Medicine

In Life sciences/Medicine (fig. 5), the largest non-article output shares are from **theses**, **conference abstracts** and **editorials/letters**. Conference proceedings and book chapters make relatively small contributions. Theses are reported in KUOZ, but not covered in any of the three databases studied. Editorials/letters can be identified in Web of Science and Scopus (and are included in ‘articles’ in Dimensions). Book chapters are included to some extent in Scopus and Dimensions, with Scopus also including a small proportion of conference papers.

Conference abstracts are included in Web of Science only, where they are the most common non-article publication type for Life sciences/Medicine, with absolute numbers exceeding those reported in KUOZ. This may reflect an underreporting in KUOZ data, or differences in classification of what are considered conference abstracts. In addition, the field classification of the KUOZ data (with HOOP area Nature split evenly between Physical sciences/Technology and Life sciences/Medicine) may mean that some conference abstracts classified as Physical sciences/Technology may, in fact, belong to Life sciences/Medicine in the KUOZ data.

Dimensions is the only database that includes preprints, and since at the time of sampling, [biorXiv](https://www.biorxiv.org/) was the only preprint server harvested by Dimensions, it is no surprise that these contribute to the output in Life sciences/Medicine in this database.

LIFE SCIENCES & MEDICINE - absolute numbers



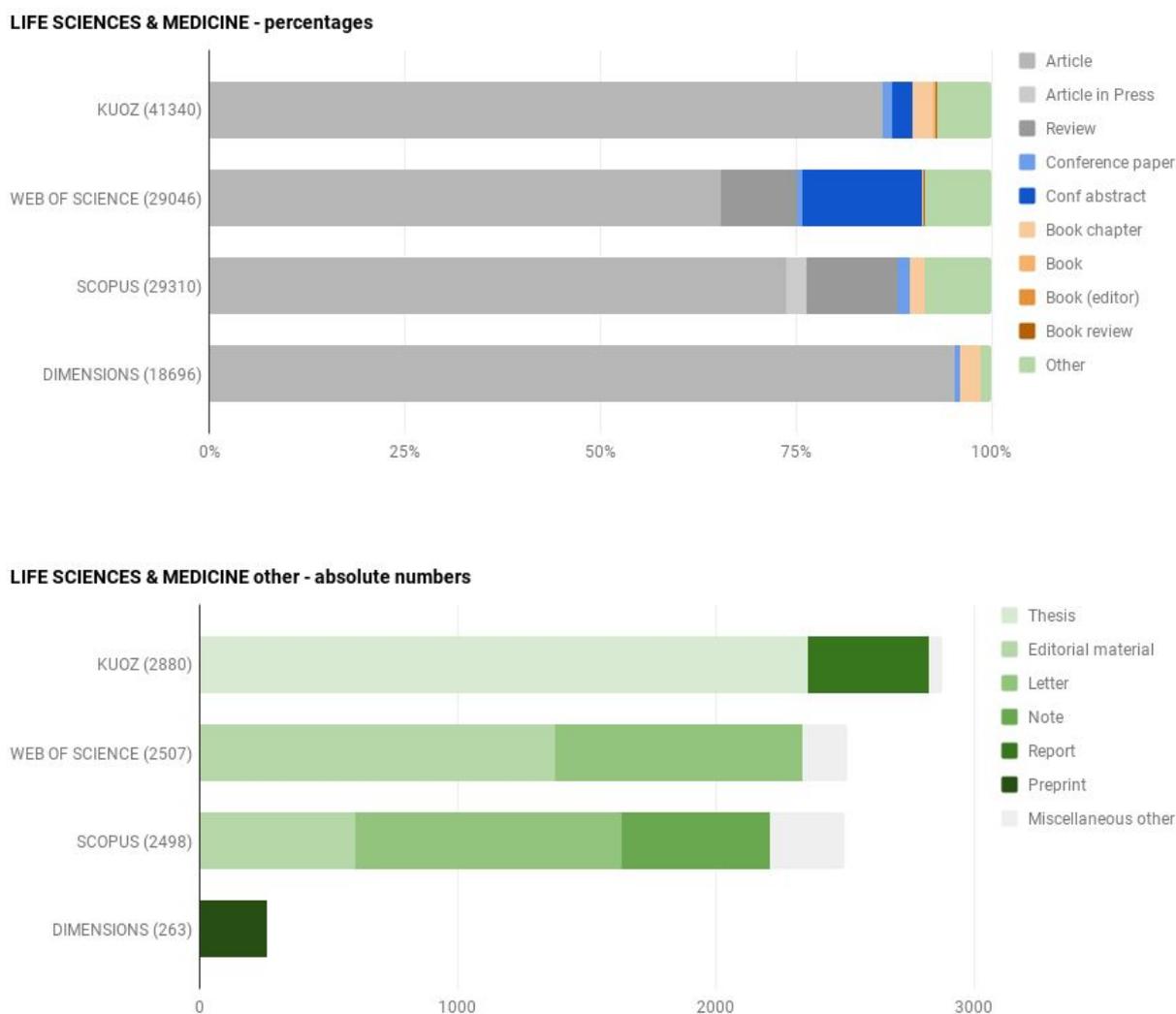
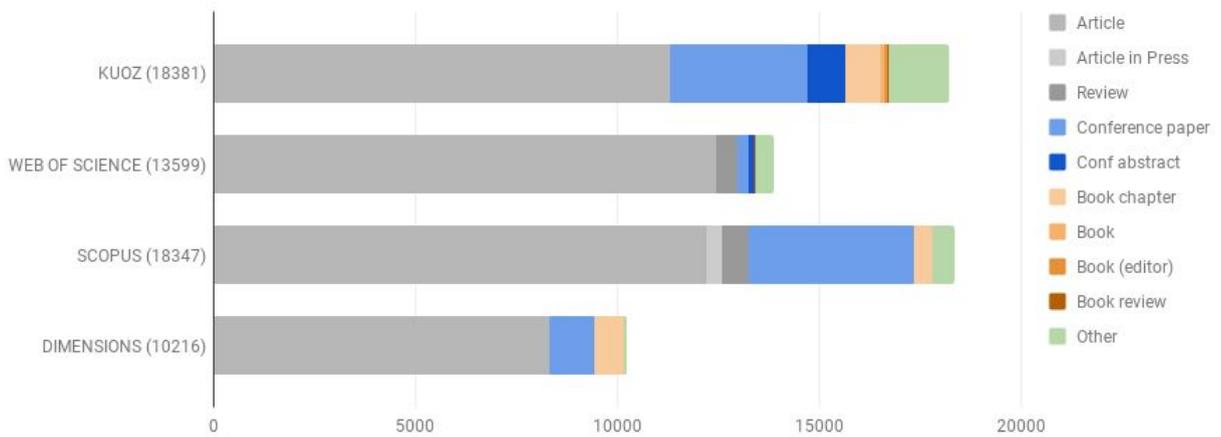


Figure 5. Total number of publications and share of publications in Life sciences/Medicine, for each of three multidisciplinary databases, compared to KUOZ data. Sources: VSNU, KUOZ data 2017; Web of Science, Scopus, Dimensions.

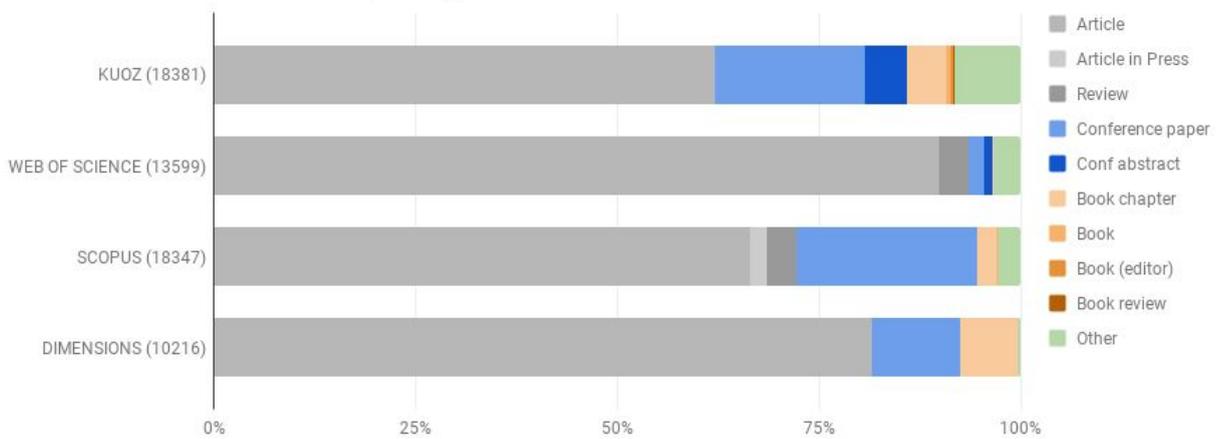
Physical sciences/Technology

In Physical sciences/Technology (fig. 6), the largest non-article output shares are from **conference proceedings, book chapters** and **theses**. Conference abstracts, editorials and reports make up smaller percentages of output. Conference proceedings in this field seem well represented in Scopus and to a lesser extent in Dimensions. Both of these databases also include a certain amount of book chapters in these fields. Theses and reports are reported in the KUOZ data, but not included in either of the three citation databases, while editorials can be identified in Web of Science and Scopus only. Regarding conference abstracts, these are present in Web of Science, but mostly attributed to Life sciences/Medicine rather than to Physical sciences/Technology (see fig. 5). As noted above, this might be an artifact due to the way fields are assigned in the KUOZ data (with the HOOP area 'Nature' divided equally over Life sciences/Medicine and Physical sciences/Technology).

PHYSICAL SCIENCES & TECHNOLOGY - absolute numbers



PHYSICAL SCIENCES & TECHNOLOGY - percentages



Physical Sciences & Technology other - absolute numbers

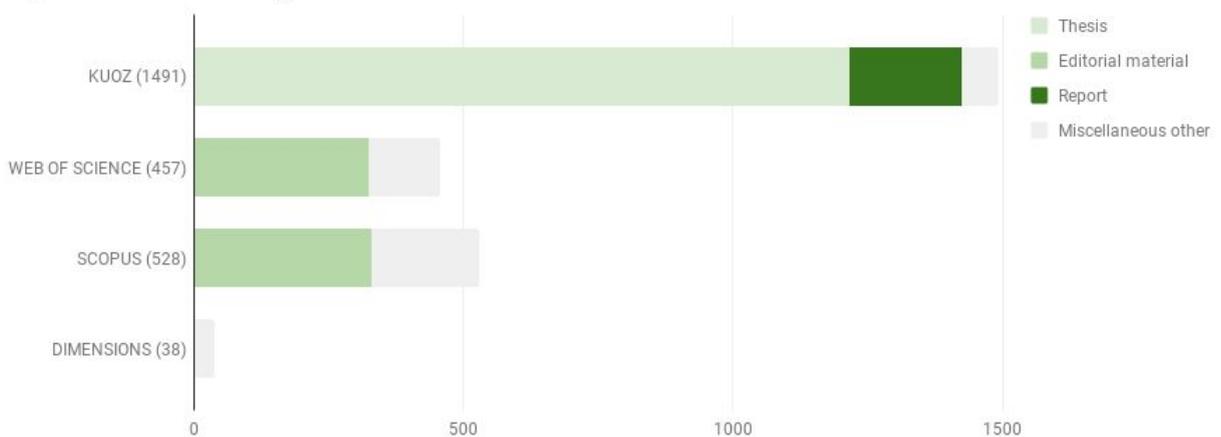


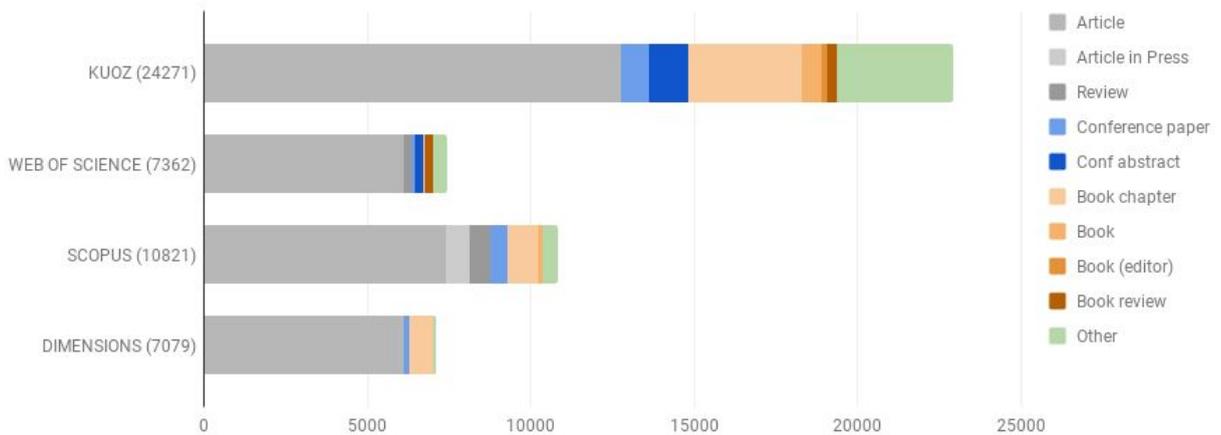
Figure 6. Total number of publications and share of publications in Physical sciences/Technology, for each of three multidisciplinary databases, compared to KUOZ data. Sources: VSNU, KUOZ data 2017; Web of Science, Scopus, Dimensions.

Social sciences

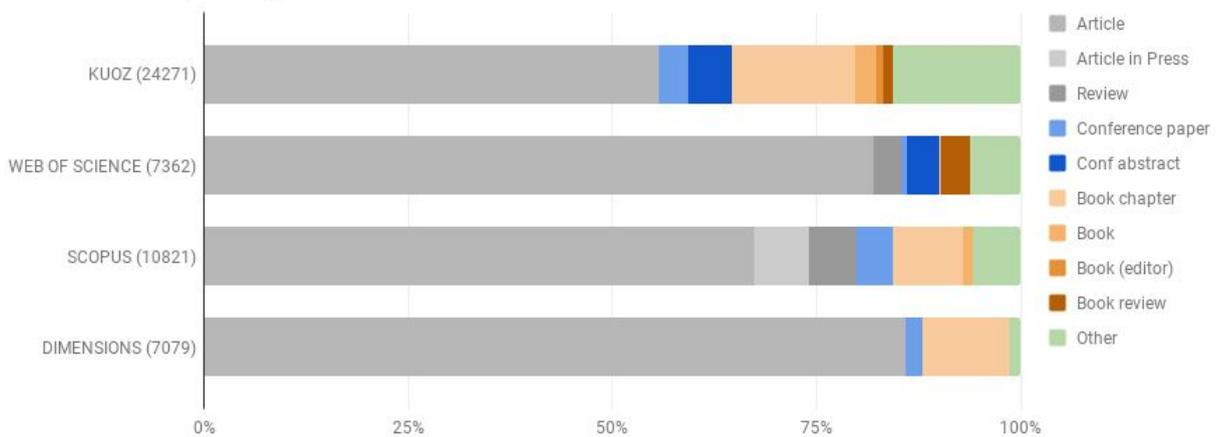
In Social sciences (fig. 7), non-article output makes up almost 50% of research output as reported in the KUOZ data. The largest non-article output shares are from **book chapters**,

conference abstracts and **case notes**, with conference abstracts, theses, reports and web publications also contributing sizeable (and more or less equal) shares. The representation of this bibliodiversity in Social sciences in citation databases is limited: book chapters and conference proceedings and conference abstracts are covered to some extent in Scopus and Dimensions, and conference abstracts in Web of Science. However, case notes (a publication type specific to the field of Law), theses, reports and web publications are present in the KUOZ data, but not retrieved from the citation databases studied. Regarding article output, the number of articles retrieved from each of the three citation databases is considerably less than the article output reported in the KUOZ data.

SOCIAL SCIENCES - absolute numbers



SOCIAL SCIENCES - percentages



Social Sciences other - absolute numbers

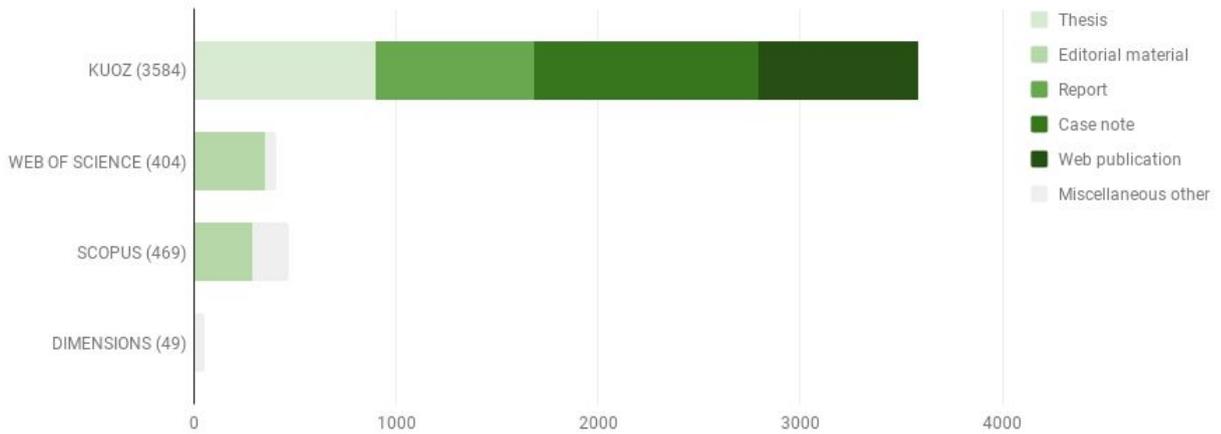
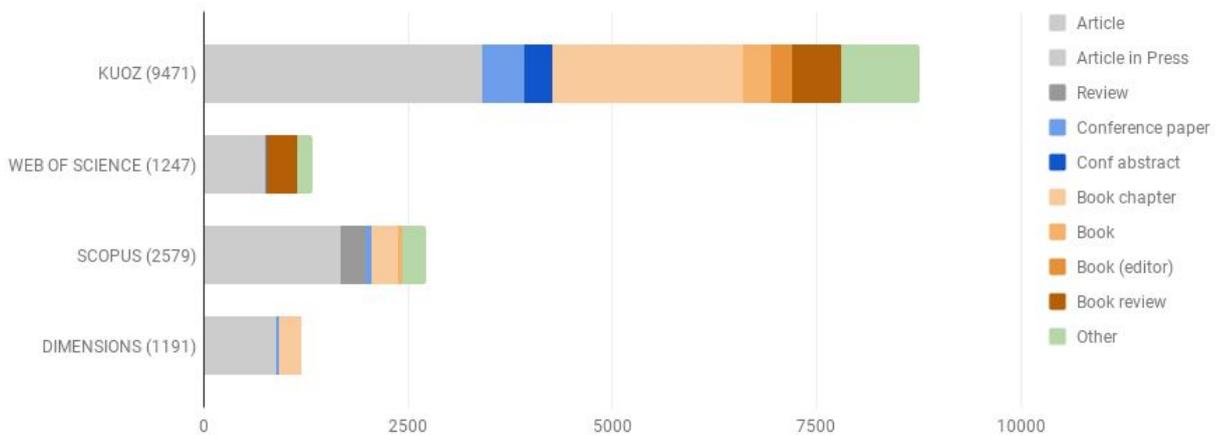


Figure 7. Total number of publications and share of publications in Social sciences, for each of three multidisciplinary databases, compared to KUZ data. Sources: VSNU, KUZ data 2017; Web of Science, Scopus, Dimensions.

Arts/Humanities

In Arts/Humanities (fig. 8), over half of all research output in the KUZ data is non-article output. In this field, **book chapters** are the most frequent non-article output type, with **book reviews**, **conference papers**, **web publications**, **conference abstracts**, **books** and **theses** contributing sizeable (and more or less equal) shares. Some book chapters are included in Scopus and Dimensions, and book reviews can be found in Web of Science. But for Arts/Humanities most other publication types are mostly absent from the three multidisciplinary citation databases included in this study.

ARTS & HUMANITIES - absolute numbers



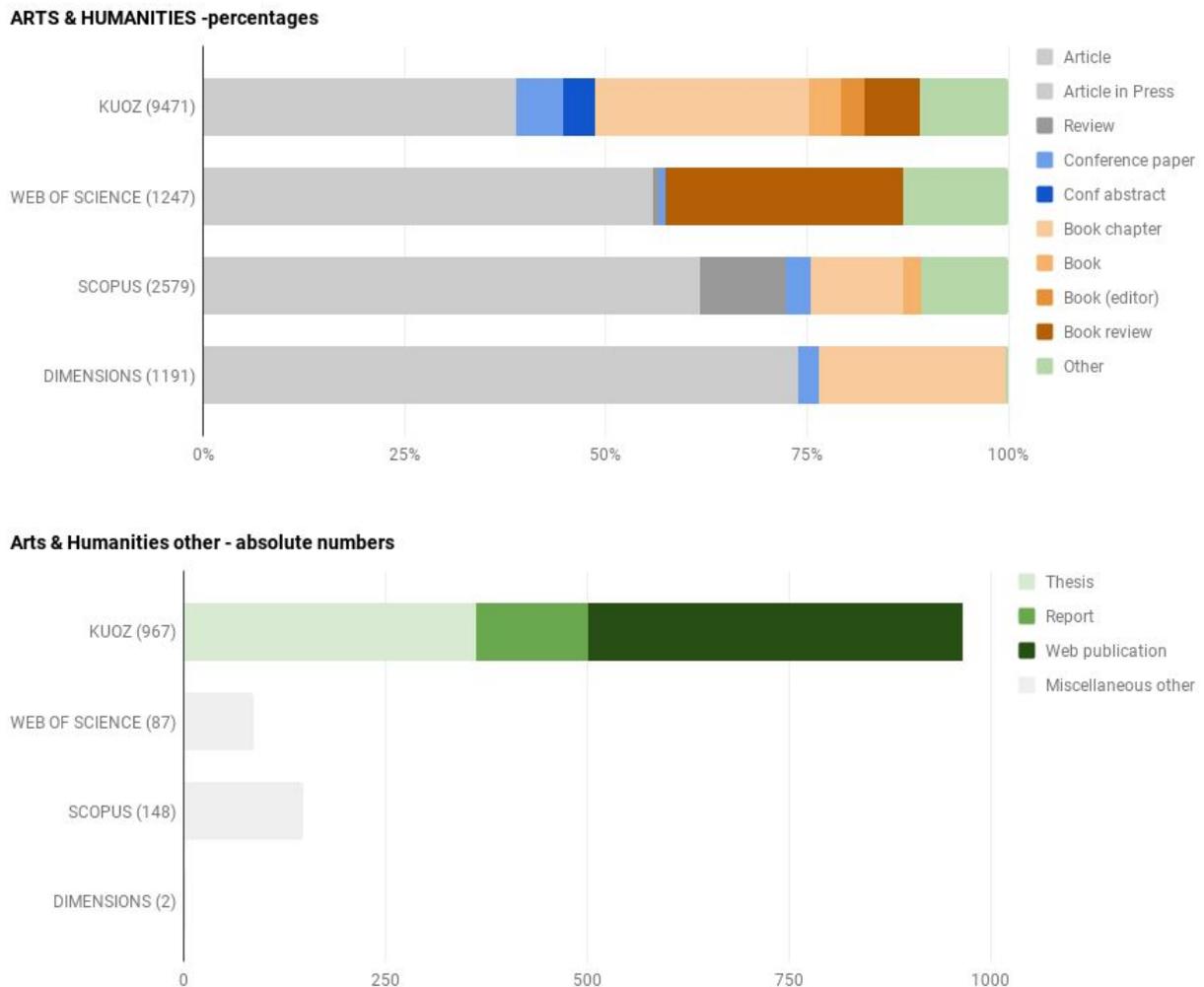


Figure 8. Total number of publications and share of publications in Arts/Humanities, for each of three multidisciplinary databases, compared to KUOZ data. Sources: VSNU, KUOZ data 2017; Web of Science, Scopus, Dimensions.

Overlap between databases (publications with DOI)

The data above illustrate the respective coverage of publication types in each of the three citation databases as compared to the KUOZ data (overall and for each field). However, the fact that a certain publication type has corresponding relative shares (or absolute numbers) in different databases, is not, in itself, an indication that the databases contain the same publications. Assessing whether the same publications are present in the different databases studies requires identification of these publications on an individual level. While theoretically this could be done by a combination of title, author and source matching, we confined ourselves to the subset of publications with a DOI, and used this DOI as reliable identifier of individual publications. In all three citation databases, at least 90% of publications have a DOI (WoS 90%, Scopus 96%, Dimensions 99%) with high percentages for articles (98%, 97% and 99%, respectively) and book chapters (92%, 94% and 100%) alike. Conference papers in Scopus less often have a DOI (82%) than in WoS and Dimensions (98% and 100%). The publication type in our sample that lacks DOIs in the majority of cases is conference abstracts (only 20% have a DOI). Conference abstracts are only included in Web of Science, and as seen above, mostly in Life sciences/Medicine. In general, Web of Science can be expected to have more material without DOI, because it ingests full journals

(including publication types without DOI, like conference abstracts as shown above). It also includes print journals without DOIs, and other publication types without DOI.

Analyzing the overlap between research output retrieved from Web of Science, Scopus and Dimensions revealed that only 43% of publications with DOI are retrieved from all three databases (fig. 9). The other publications are either not included in all three databases (e.g. because the journal or the specific publication type is not ingested), or not retrieved (e.g. because affiliation detection is less complete, or the publication year recorded differs). Overlap between Web of Science and Scopus is largest, with Web of Science having the least unique publications in this sample. Dimensions has the most unique titles, possibly because it ingests all metadata from [Crossref](#) (the organization that assigns DOIs to scholarly content and keeps a registry of metadata for that content). This includes titles and content types not covered in Web of Science in Scopus.

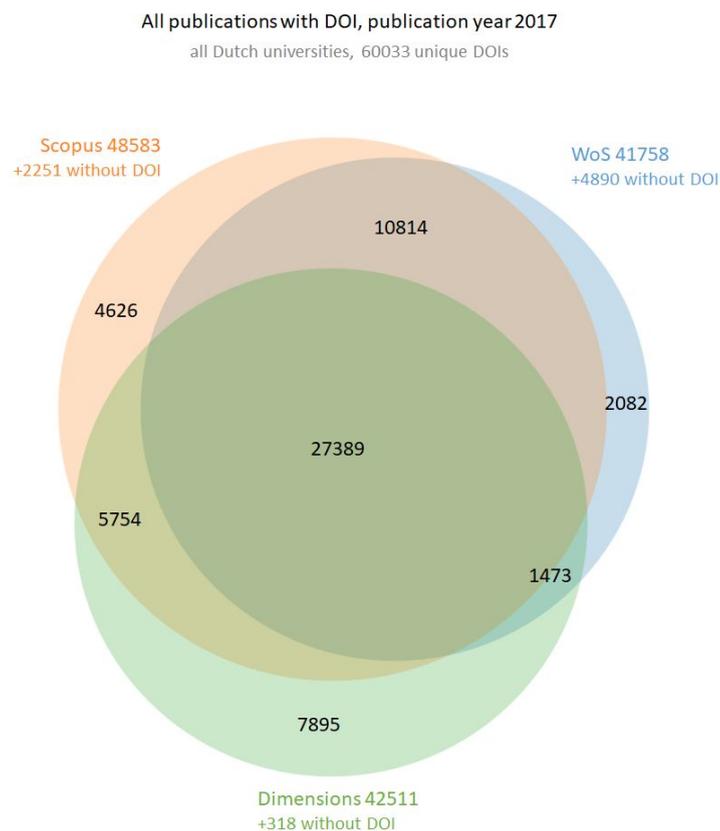


Figure 9. Overlap in research output from Dutch universities (publication year 2017) between three large citation databases. Overlap was determined for publications with DOI only. Venn diagram not entirely proportional. Sources: Web of Science, Scopus, Dimensions.

Taken together, Dutch research output from 2017 in the three citation databases amounts to 60,033 unique publications (with DOIs). This is still considerably less than the number of publications reported in the KUOZ data (93,462). This will be partly due to duplicates in the KUOZ data (as discussed previously), but the KUOZ data also contain publication types not included in the large citation databases (particularly theses, reports and web publications, as well as books). Unfortunately, because the KUOZ data only contain aggregate publication

counts, it cannot be asserted what share of publications in the KUOZ data have a DOI, nor what the overlap is between the output reported in the KUOZ data with the output retrieved from the three citation databases for the share of publications with a DOI.

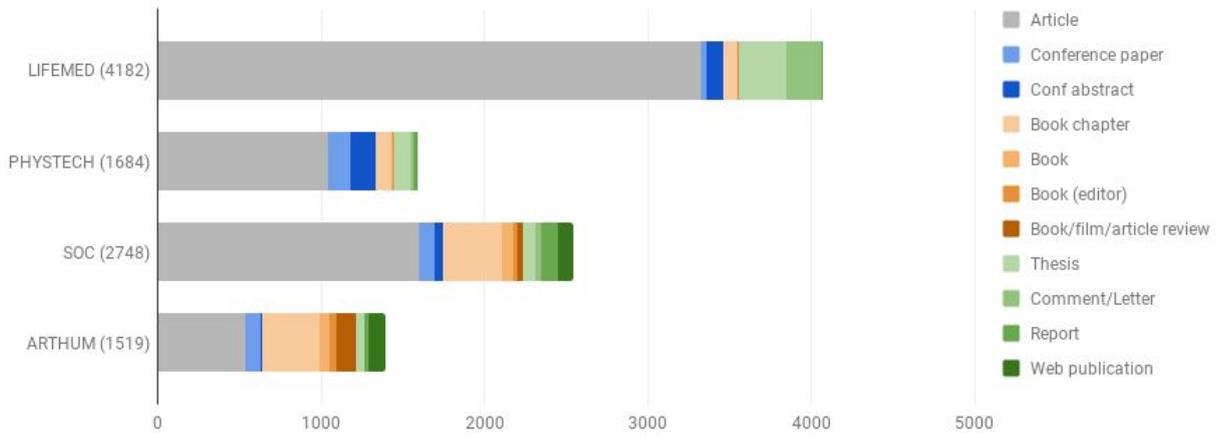
To get more insight into the proportion of Dutch research output that is covered by the main citation databases, we took the CRIS output of our own university (Utrecht University) as a case study. This allowed us to complement the aggregated KUOZ data with a subset of data for which information is available at the level of individual publications.

Case study: Utrecht University CRIS output (2017)

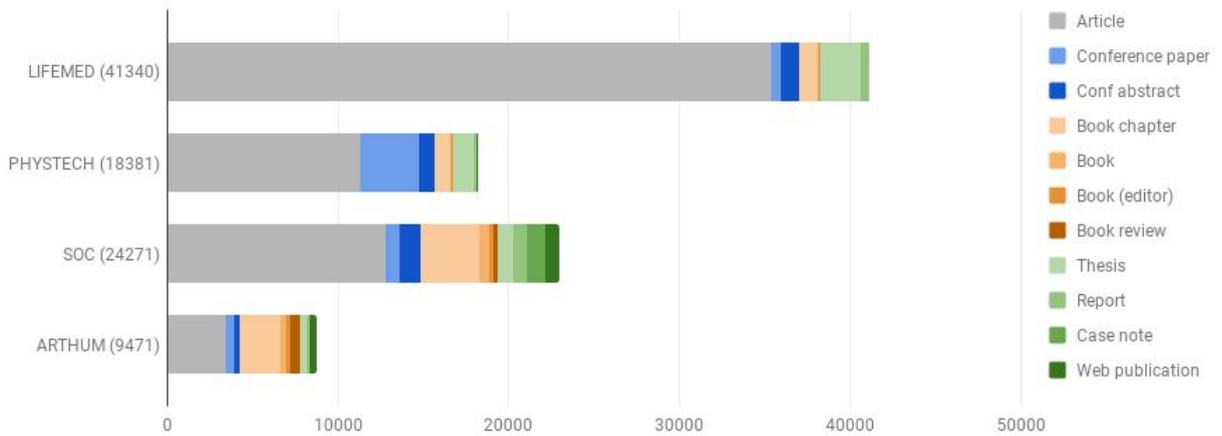
The full 2017 research output from Utrecht University and UMC Utrecht (insofar as reported in the CRIS) was downloaded from the CRIS (Utrecht University Pure and UMC Utrecht Pure) on August 14, 2018. Publications were deduplicated and assigned to one or more of the four main fields used in the rest of this study, based on the faculties and/or departments of the author(s) of each publication. Publication types used in the CRIS were mapped to the publication types in KUOZ as closely as possible.

The total number of unique publications retrieved from the UU/UMCU CRIS was 9903, compared to 7887 in Web of Science, 7776 in Scopus and 6727 in Dimensions. Utrecht University is a broad, multidisciplinary university, and the distribution of publication types for each main academic field largely matches the distribution of the KUOZ data (which contains data from the CRIS of all Dutch universities) (fig. 10). Some differences include a relatively large share of comments/letters in Life sciences/Medicine (a publication type not included in the KUOZ data), and a difference in the shares of conference papers (less in UU/UMCU) and conference abstracts (more in UU/UMCU) in Physical sciences/Technology. In the KUOZ data, Physical sciences includes output from the three technical universities in the Netherlands (Delft, Eindhoven and Twente), whose publication culture could well include more conference papers relative to conference abstracts. In Utrecht, where Physical sciences/Technology represents output from the departments of physics, and chemistry, in addition to that of mathematics and computer science, conference abstracts can be expected to play a larger role in research output. Case notes (a publication type specific to the field of law) are part of the research output of Utrecht University, but their share is much lower than in the KUOZ data and they are thus not included in the charts.

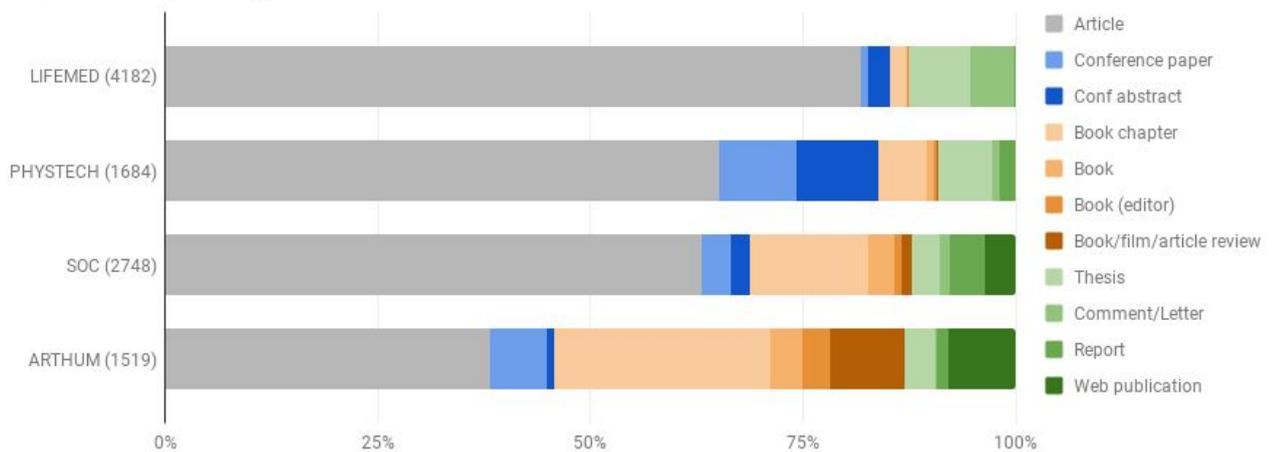
UU / UMCU CRIS - absolute numbers



KUOZ data - absolute numbers



UU / UMCU CRIS - percentages



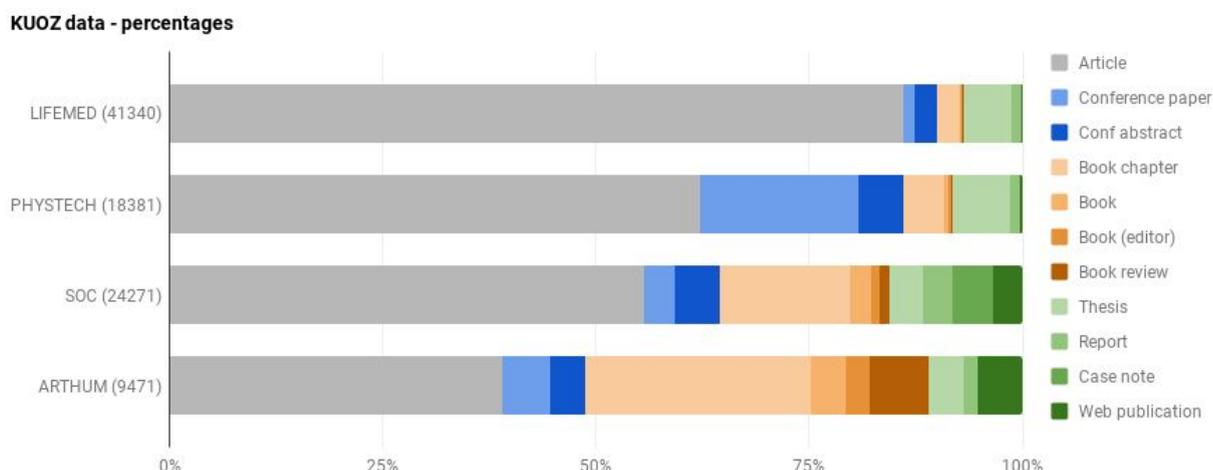


Figure 10. Comparison of publication types in the four main academic fields between UU/UMCU CRIS (A,C) and KUOZ data (B,D). Sources: UU / UMCU CRIS, VSNU, KUOZ data 2017.

Unlike the KUOZ data, which are only available at aggregate level, the Utrecht CRIS data allowed us to check the share of publications that have a DOI available, and to assess the overlap between a university’s reported research output and the coverage and retrieval of that output in the three large citation databases studied.

Figure 11 shows the proportion of each publication type that has a DOI included in UU/UMCU CRIS. Overall, only 67% of publications have a DOI, which is in stark contrast to the sample of total Dutch research output retrieved from Web of Science, Scopus and Dimensions, where this percentage is 90% or more, as shown earlier. While for articles, the percentage of DOIs is only slightly lower for the UU/UMC CRIS sample (86%) than for Dutch output in the citation databases (97-99%), there is a much stronger difference for other publication types included in the CRIS as well as in the citation databases, such as book chapters (20% vs. 92-100%), conference papers/proceedings (29% vs. 82-100%). In both UU/UMCU CRIS and Web of Science, only a minority of conference/meeting abstracts have a DOI (13% vs. 20%, respectively).

A brief methodological note: in the UU/UMCU CRIS, two types of conference abstracts and two types of conference papers are distinguished. For each publication type, one is labeled as conference contribution (conference abstract / conference paper, respectively), and one as journal- or book/report contribution (meeting abstract / proceeding, respectively). In the analyses so far, we have combined these categories, but here we show them separately. There is a clear difference in the share of publications that have a DOI, with more contributions to journals and books/reports having DOIs than contributions to conferences.

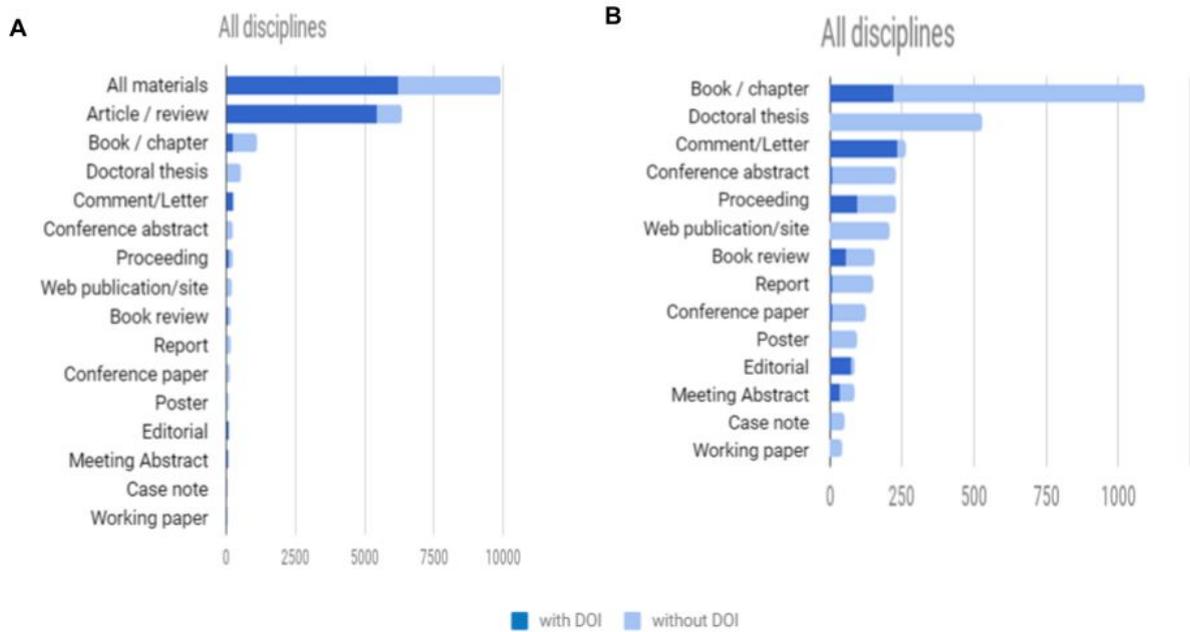


Figure 11. Research output (2017) in UU/UMCU CRIS with and without DOIs. A: most frequent publication types; B: most frequent publication types other than articles. Source: UU / UMCU CRIS.

On one hand, the observed differences may point to selection bias in the coverage of the large citation databases in favour of sources that also happen to have a DOI. An extreme example is Dimensions, which ingests all metadata from [Crossref](#) (the organization that assigns DOIs to scholarly content) - logically, all this content has DOIs. In addition, the CRIS also contains non-scholarly content (professional and popular publications) that is less likely to have a DOI. Finally, inclusion of DOIs in the CRIS might be incomplete, leading to false negatives in the CRIS data.

The differences in use of publication types in the main academic fields, combined with the proportion of these publications types that have a DOI, means that any analysis limited to publications with a DOI (e.g. checking open access status with Unpaywall) disproportionately affects Social sciences and especially the humanities. In these fields, only 49% and 24% of publications in CRIS have a DOI, respectively. For Physical sciences/Technology, this is 66% and for Life sciences/Medicine, 85% (see fig. 12).

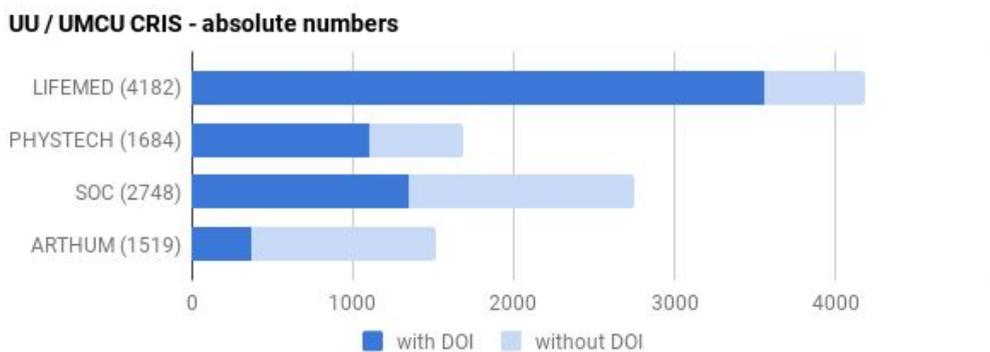


Figure 12. Proportion of research output in UU / UMCU CRIS with and without DOIs, per main academic field. Source: UU / UMCU CRIS.

For the share of research output in UU / UMCU CRIS that does include a DOI, overlap between the research output retrieved from Web of Science, Scopus and/or Dimensions can be assessed. Only 33% of publications are retrieved from all four databases, and all databases have content not included in or retrieved from any of the other databases (fig. 13). 15% of DOI-content in Pure (870 of 6177 publications) is not retrieved from any of the other databases studied, either because the publication venue or the specific publication type is not ingested in one of the citation databases, because affiliation detection is less complete or the publication year recorded differs. The fact that the three citation databases together contain a lot of research output not included in the CRIS could point to the fact that the CRIS was still incomplete for the preceding publication year at the moment of download (August 2018), or is unable to get a lot of output registered.

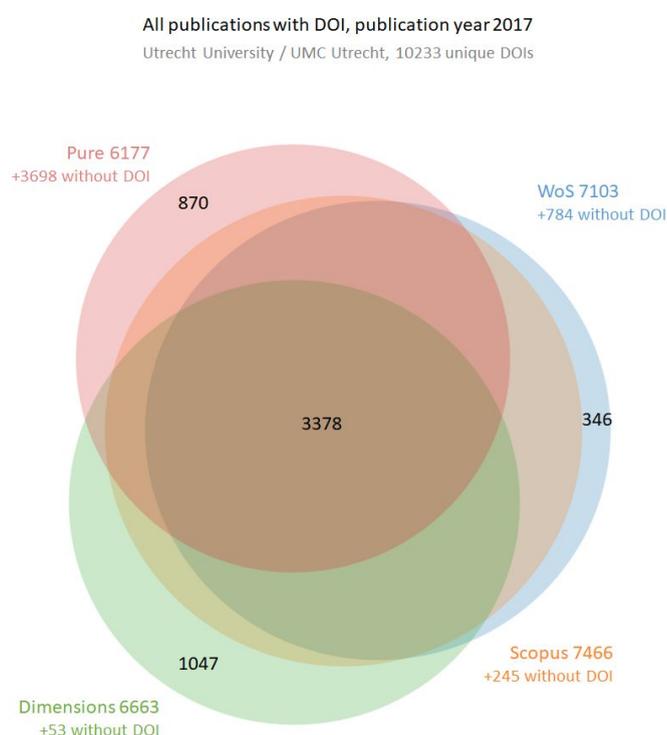
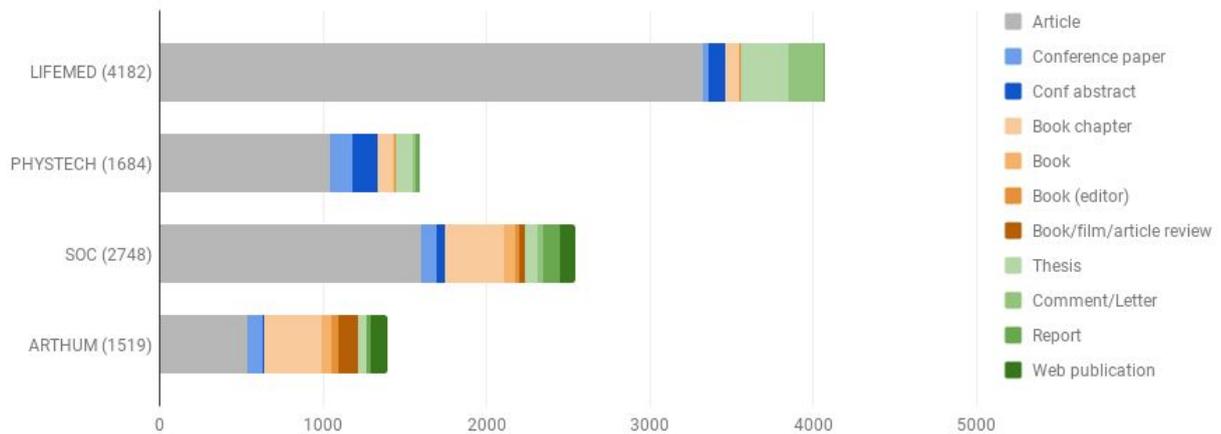


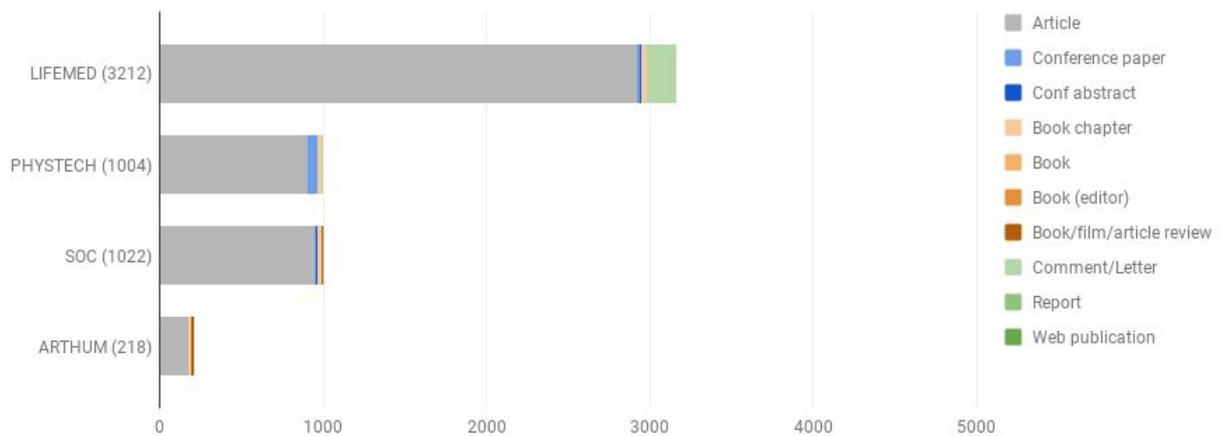
Figure 13. Overlap in research output from UU / UMCU (publication year 2017) between Pure (UU/UMCU CRIS) and three main citation databases. Overlap was determined for publications with DOI only. Venn diagram not entirely proportional. Sources: UU / UMCU CRIS, Web of Science, Scopus, Dimensions.

Given that one-third of publications in the UU / UMCU CRIS does not have a DOI, and that of publications with a DOI, some 15% is not retrieved from any of the large citation databases, what are the implications of only using DOI-content from these large citation databases in studies on the research output of an institution or group of institutions? Figure 14 shows the proportion of research output from the UU/UMCU CRIS per main academic field that is retrieved from either of the three citation databases, as well as the proportion of output that is left out when only these databases are used and only DOI-containing output is considered.

UU / UMCU CRIS - absolute numbers



UU / UMCU CRIS - detected in WoS, Scopus or Dimensions (with DOI) - absolute numbers



UU / UMCU CRIS - not detected in WoS, Scopus or Dimensions - absolute numbers

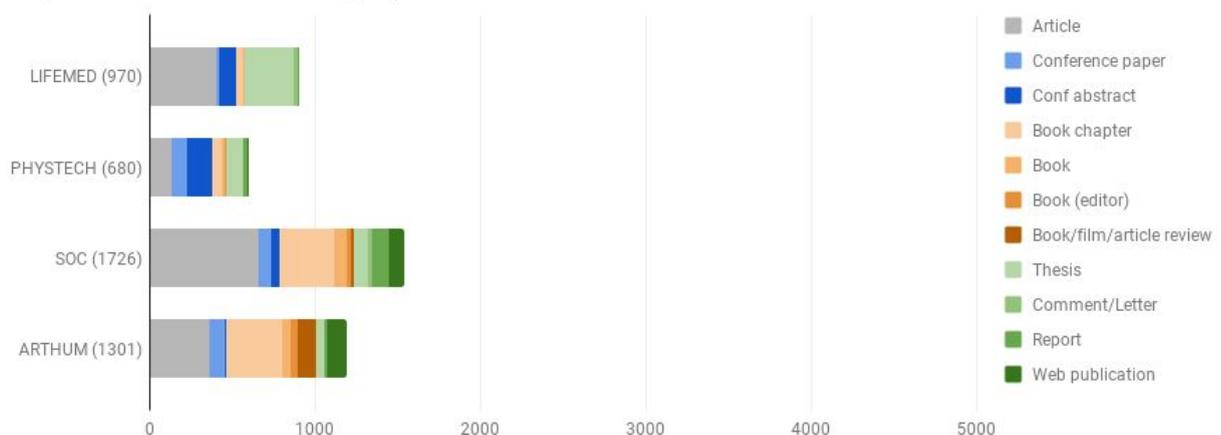


Figure 14. Proportion of research output, by publication type, from UU/UMCU CRIS that is retrieved from one of the three large citation databases when matched on DOI, as well as the proportion of research output that is not covered or retrieved in this way. Sources: UU / UMCU CRIS, Web of Science, Scopus, Dimensions.

As is clear from these figures, the proportion of CRIS output retrieved from the three large citation databases in this way is extremely homogeneous, almost exclusively consisting of articles. Thus, relying on either of the main citation databases (which are largely restricted to

content that contains DOIs) for analysis of research output means losing the bibliodiversity present in the CRIS (for all academic fields), as has been shown before (Van Leeuwen et al. 2016). Specific effort is needed to ensure all research output from an institution or group of institutions is considered and included - for instance, by also using other databases for publication types not included in the three main citation databases. At the same time, information from other databases (including large citation databases) can also complement information on research output as present in universities' CRIS systems and, consequently, the KUOZ data.

Using additional databases

Web of Science, Scopus and Dimensions are not the only multidisciplinary databases that can be used to assess the composition of Dutch university output by publication type. Four main ones that are openly available and actively maintained are NARCIS, BASE, OpenAIRE and LENS. The first three are mainly aggregating metadata from institutional and subject repositories, while LENS uses data from Microsoft Academic, Crossref and PubMed to fill its database. All four are non-commercial. NARCIS is maintained by DANS and restricted to Dutch institutions. BASE is maintained by Bielefeld University Library and has a global scope. OpenAIRE is an European Union initiative and is hence restricted to European countries. LENS is from Australia and a joint initiative of Cambia and Queensland University of Technology. All four do allow, with some restrictions on size/frequency and (commercial) usage, downloading of data in sets or via an API. The biggest problem with the three repository harvesting databases is that they either lack affiliation information or do not allow to select multiple institutions. For that reason, data presented here are for the Netherlands as a whole (or at least, based on all Dutch repositories, including those of universities of applied science). LENS does allow restricting to sets of institutional affiliations and data here are hence for Dutch universities and medical centres only.

The three repository harvesting databases show a publication type composition that is relatively rich in non-article output (fig. 15). This is strongest for NARCIS, especially considering that the other category is entirely made up of non-article output. OpenAIRE does not do a good job capturing Dutch repository content compared to BASE and NARCIS. LENS also captures less, though that is partly caused by the fact that the selection from LENS made here is restricted to universities. LENS is comparable with Dimensions in amount of Dutch university output covered. But just as with Dimensions, non-article output is weakly represented. It is regrettable that BASE and NARCIS, that both do provide a more varied image of the output compared to the main citation databases, have substantial shares of records with unknown publication type.

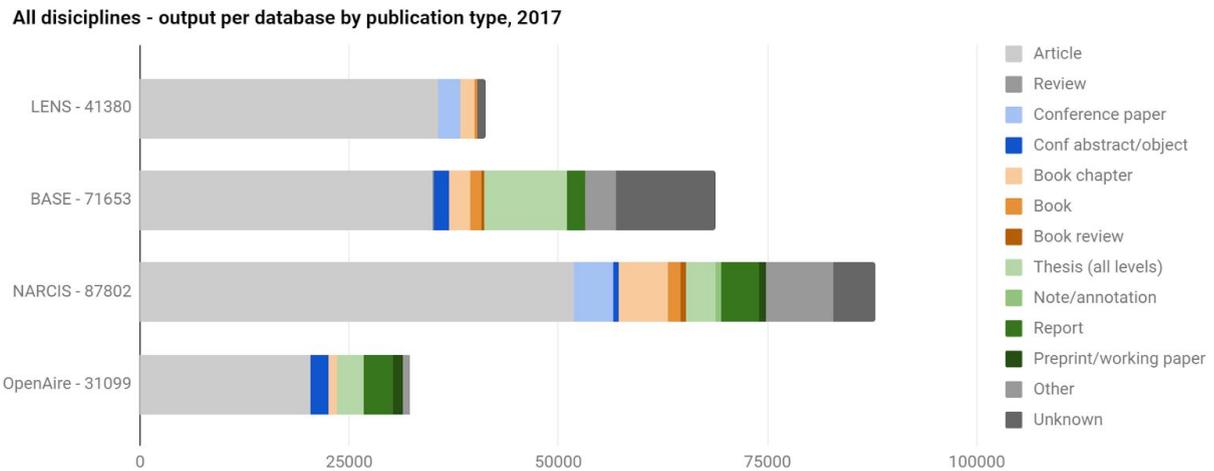


Figure 15. Additional multidisciplinary databases, Dutch output by publication type, 2017. Sources: LENS, BASE, NARCIS, OpenAIRE.

Finally, in looking for database alternatives, theoretically there is also an option to use an extensive range of disciplinary databases to assess output. This option has not been explored here. With the dozens of (sub)fields and hundreds of databases that approach would very likely be too labour intensive, complex and unsustainable. The usability of these databases would also depend on the inclusion of affiliation information.

Books, a special case

Books deserve special attention, because of the quantitative and well as qualitative importance of books for most of Arts/Humanities subfields as well as some Social sciences subfields. Using Utrecht University CRIS data it is possible to provide more detail, for instance on publishers of book output. That detail may be useful information for open access policy. The data indeed shows that Physical sciences/Technology and Life sciences/Medicine have very modest book output, more so considering their sheer size (fig. 16). What is less well known is the very long tail of small publishers in Social sciences and especially Arts/Humanities, where the largest 10 publishers publish not much more than half of the book output.

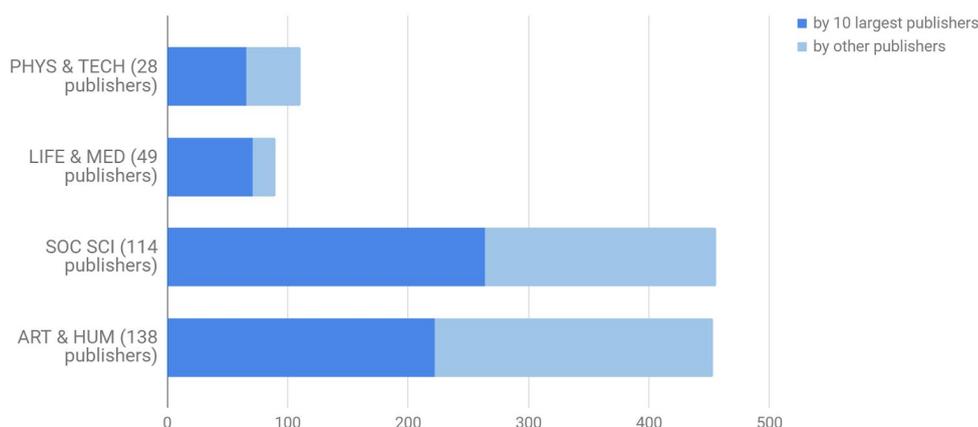


Figure 16. Books and book chapters in Utrecht University / UMC Utrecht output, per discipline, 2017, with share of largest 10 publishers. Source: UU / UMCU CRIS.

Looking at which publishers exactly are chosen most frequently (fig. 17) a picture arises of a small number of broad publishers that cater for all fields (Springer, Taylor & Francis), some that are strong in just a few fields (Oxford UP, Wiley) and some that are really specialised (Brill, IEEE, Ars Aequi, BSL, Epsilon)

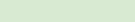
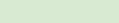
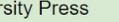
LIFE & MED	90		PHYS & TECH	111	
SpringerNature	35		SpringerNature	29	
Elsevier	10		Epsilon	12	
Nederlandse Hartstichting	7		Geological Society of London	4	
Bohn Stafleu van Loghum	5		Cambridge University Press	3	
Taylor & Francis	5		IEEE	3	
Wiley	3		SciTePress	3	
Koninklijke Van Gorcum	2		Taylor & Francis	3	
Wageningen Academic Publishers	2		Utrecht University	3	
Balans	1		VELON	3	
Boom	1		Verloren	3	
SOC SCI	456		ART & HUM	453	
Taylor & Francis	47		Taylor & Francis	46	
SpringerNature	44		SpringerNature	33	
Boom	35		Amsterdam University Press	32	
Edward Elgar Publishing	32		Brill	31	
Wolters Kluwer	28		Verloren	19	
Wiley	21		Oxford University Press	14	
Ars Aequi	18		Boom	13	
Oxford University Press	14		De Gruyter	12	
Eleven international publishing	13		Ubiquity	12	
SDU	12		John Benjamins	10	

Figure 17. Ten most common publishers of books and chapters, per field, for Utrecht University / UMC Utrecht output, 2017. Source: UU / UMCU CRIS.

3. Open Access levels

Open access in national reporting

Currently, assessing open access (OA) levels is usually confined to peer-reviewed articles. This is a direct result of current OA-policies and mandates on reporting, including that of the VSNU, that focus on article output. Dutch universities report yearly to the VSNU on the OA-level and types of OA of their article output (VSNU 2018), using the Definition framework monitoring Open Access (VSNU 2017). Universities generally use the information in their CRIS systems for reporting on OA-levels - partly manually, partly by automation and use of external services as [Unpaywall](#), which allows the detection and classification of OA-status of scholarly articles that have a DOI.

Open access levels of content covered in Web of Science, Scopus and Dimensions

We used Unpaywall in June/July 2018 to assess OA-levels and types of OA for research output from 2017 from Dutch universities as identified in Web of Science, Scopus and Dimensions, limiting detection to publications containing a DOI². Unpaywall also provides information on licenses used, and its data is openly available for non-commercial use. In all, 60,033 unique publications with DOIs were retrieved from these databases. As discussed earlier and also shown in Figure 18A-B, selecting on DOIs disproportionately favours articles over other publication types. Trying to take into account more publication types may thus imply having to deal with higher shares of lacking DOIs and thus more effort needed to analyse open access availability or licenses.

The issue of lacking DOIs is strongest in Web of Science (fig.18 A), a reflection of that database having a better coverage of publication types that often lack DOIs, and also of journals that still do not assign DOIs to their articles. Dimensions, sourced from DOI-registration organisation Crossref, thus almost by definition lacks this problem. Lacking DOIs and thus OA-detection challenges are most frequent with some non-article output like meeting abstracts and book reviews (in Web of Science) and conference papers (in Scopus). To what extent this problem will naturally solve itself through time by higher awareness among publishers of the importance of persistent identifiers remains to be seen.

² Methodological note: Web of Science, Scopus and Dimensions all use Unpaywall to give information on OA-status directly in the database. However, each database can and does make its own decisions on how to interpret and display OA information. For this reason, we used Unpaywall separately in this study.



Figure 18. OA-levels and classification of 2017 research output of Dutch Universities, per document type. Publications retrieved from the three large citation databases; OA-levels retrieved from Unpaywall. A: most frequent document types; B: most frequent document types other than articles, C: OA classification. Sources: Web of Science, Scopus, Dimensions, Unpaywall.

Open access levels found this way are 45-55% for articles (fig. 18A). This might paint a slightly too positive picture of OA availability of scholarly articles overall, because these three databases tend to have a bias towards fields that are more likely to have higher open access levels (see e.g. Piwowar et al. 2018 and Bosman and Kramer 2018 for a discussion on OA-levels of various (sub)fields). Zooming in to non-article output (fig. 18B), OA-levels drop significantly, with low levels for conference material and especially book chapters.

A contributing factor for the lower detection of non-article material, even that with DOIs, could be the focus of Unpaywall on detecting open access versions of scholarly articles. If publication venues for non-article formats (e.g. full OA book publishers, or general websites for reports) are not harvested by Unpaywall, publications hosted there will not be detected as OA, even when they are openly available. It would be good to check the extent to which this is indeed a limitation of using Unpaywall to detect OA levels of non-article publication types.

Unpaywall allows the detection of the following types of OA:

- full gold (journal in DOAJ)
- hybrid (journal not in DOAJ, article with open license)
- bronze (journal not in DOAJ, article without open license)
- green only (article in repository, not also OA in a journal)

For publications that are not part of journals (such as books/book chapters, conference abstracts) as well as for conference proceedings not published in journals (but with DOIs), this method of classification means that these publications will never be classified as full gold, but either as hybrid or bronze, depending on the license detected. If books/book chapters or conference proceedings are only detected as OA in a repository, they will be classified as green, similar to journal articles only available through a repository.

Analysis on types of open access (fig. 18C) reveals publication type-specific shares of these various OA types. Where hybrid and gold are important for articles and reviews, we see a much stronger role of bronze OA in editorial material and letters and a relatively strong role for green in proceedings and conference papers. If book chapters are open access at all, that is even predominantly achieved via the green route.

The data here show great variability in OA availability and forms of OA across publication types. This highlights the importance of information on publication type details. In Dimensions, this information is more limited than in the other two multidisciplinary databases, with all journal content being included under ‘articles’. This limits analysis of opportunities and problems in OA availability and detection thereof. On the other hand Dimensions does - uniquely among the three multidisciplinary databases - provide data for preprints (bioRxiv at the time of sampling, with SSRN added since), which is potentially very useful in terms of OA policy setting. Thus, sourcing multiple databases, each with their strengths and weaknesses, is currently the best approach for input in policy development, provided one has access.

Open access levels of content covered in additional databases

It is interesting to see to what extent additional databases beyond WoS, Scopus and Dimensions have added value in tracking open access of Dutch universities’ output, especially for non-article publication types. In a general sense their added value may lie in greater coverage of non-article output (esp. NARCIS and BASE) and open availability (all four) (fig. 19-20).

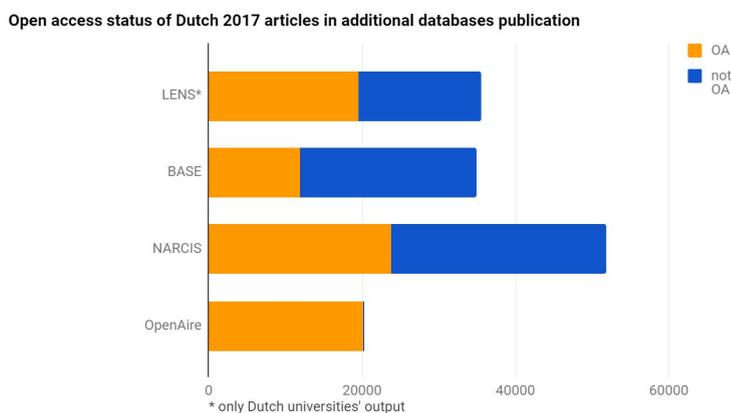


Figure 19. Open access status of Dutch 2017 articles in additional databases. Sources: LENS, BASE, NARCIS, OpenAIRE.

Open access detection is based on Unpaywall data integrated in the database (LENS) and on repository info (other three). At the time of sampling, OpenAIRE harvested open access publications only, which made it less useful for determining open access levels. The (additional) coverage of non-article content by LENS is limited (because of LENS' sources). In terms of diversity of coverage NARCIS and BASE are the most interesting, covering and differentiating content that is often not indexed by large citation databases (e.g working papers). NARCIS often finds larger open access shares. This could be influenced by the fact that for some universities, only open access content seems to be harvested by NARCIS. It would be important to check the comprehensiveness and stability of coverage and open access detection over time. In general, aggregators such as NARCIS and BASE get their open access information from the metadata of the repository content they harvest. The value of this information is thus in large part dependent on the quality of repository metadata. Repository based databases have many dependencies (input in CRIS systems, inclusion in repository, harvesting from thousands of different repositories using different publication type labels). If one compares the open access numbers found here with those found by WoS, Scopus and Dimensions, clearly NARCIS, BASE and OpenAIRE find more evidence of open access books and book chapters, next to theses.

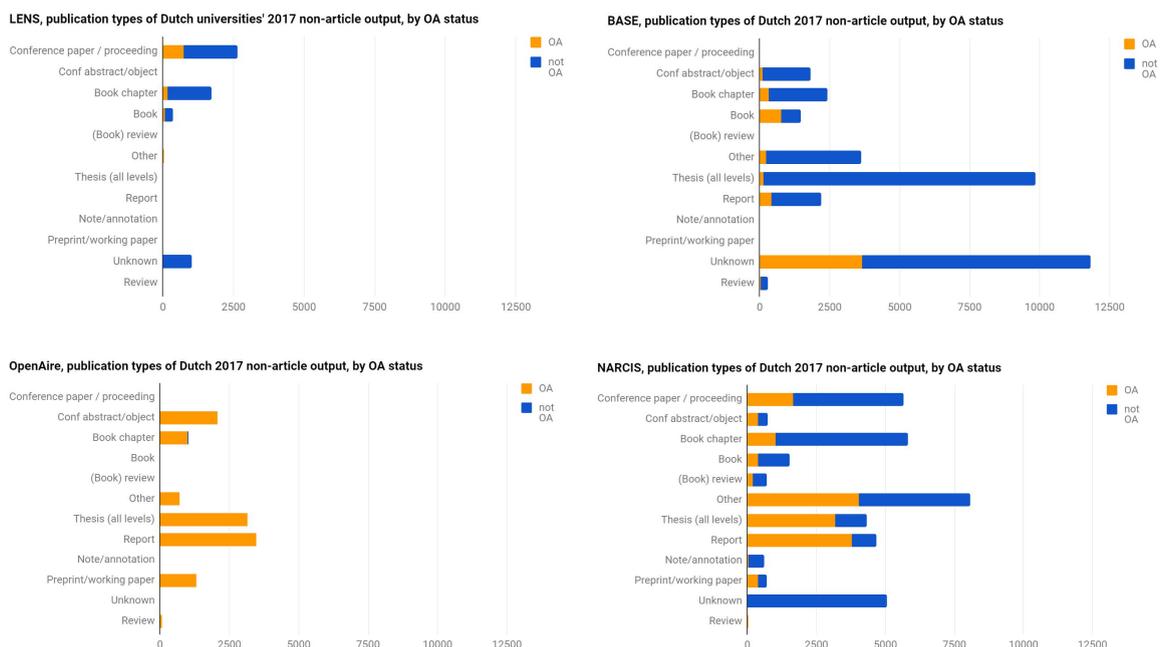


Figure 20. Additional databases coverage of Dutch universities' non-article output, per publication type, by OA status, 2017. Sources: LENS, BASE, NARCIS, OpenAIRE. Check date: 201904.

License types of OA output, by field

In open access and especially open access policy, it is desirable to indicate what rights users have beyond access and reading. For publication, the most widely accepted way to do this is to attach a Creative Commons (CC) license that either waives all rights (CC0), just requires attribution (CC-BY), or add additional restrictions on commercial use (with the NC-clause), derivative creation (ND-clause) or licensing of downstream work (SA-clause), or

combinations thereof. Next to the CC-licenses there are also custom licenses from publishers. Lacking licenses mean there is less certainty on what one can do with the publication or file (aggregating, republishing, changing/adding/remixing/forking, mining etc.) and may indicate less certainty on openness status and availability over time. This is complex matter in terms of exact rights, acceptance and use in the various publication cultures. CC-license information is however machine readable and can be analysed using Unpaywall, given that DOIs of the publications are known. The overall picture (fig. 21, using the 60,033 unique DOIs of Dutch university output from Web of Science, Scopus and Dimensions, analyzed with Unpaywall to get information on licenses) is quite clear: green open access in repositories mostly lacks a license and open access in full gold open access journals predominantly has a CC-BY license. Licenses of hybrid open access material are mixed: about half CC-BY and the other half spread over CC-BY-NC-ND, CC-BY-NC and “implied OA”, an Unpaywall term indicating there is some evidence of an open license but it could not be verified exactly.

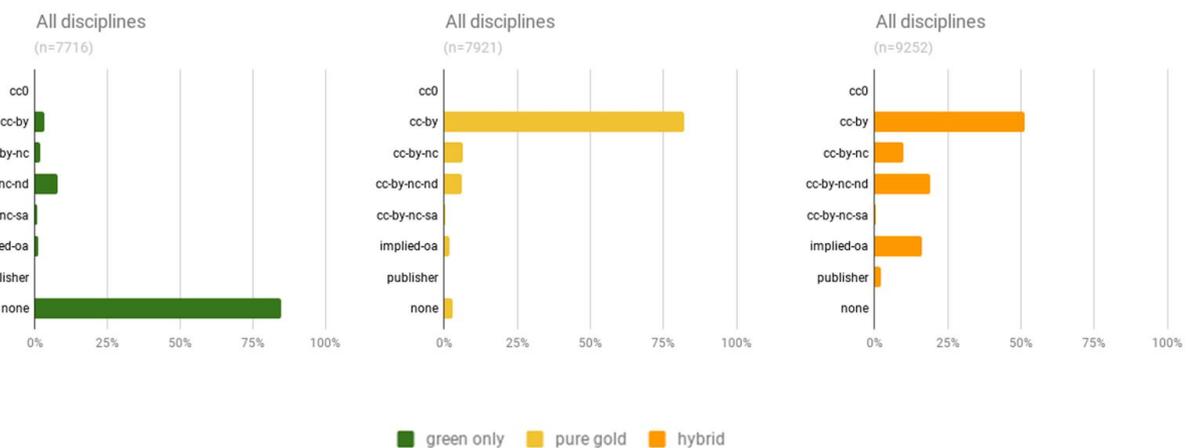


Figure 21. Licenses of open access output of Dutch universities, 2017. Sources: Web of Science, Scopus, Dimensions, Unpaywall.

Looking at the licenses attached to publications in the four main fields (fig. 22) brings nuance to the story. As this is still only for material with a DOI it must be taken into account that the picture is influenced by DOIs lacking for specific publication types. Licenses of material in repositories are still mostly lacking, though in Arts/Humanities and Social sciences there is a substantial amount with a relatively restrictive CC-license, especially NC-ND, which matches the license type required for green OA sharing by some publishers. For publication in full gold journals the predominance of the CC-BY license is strongest in Physical sciences/Technology, a bit less in Life sciences/Medicine and Social sciences and less still in Arts/Humanities, where NC and ND elements in the license are quite common (though still a minority). For hybrid OA, the proportion of CC-BY is more or less equal across fields, with around 50% of open access articles in hybrid journals carrying this license. For the other half, we mostly see more restrictive licenses (with NC- and ND elements) used in Physical sciences/Technology and Life sciences/Medicine, while Social sciences and Arts/Humanities have the highest frequency of the “implied-OA” “license”. Especially in these fields, then, there might be a reluctance of traditional publishers, while allowing hybrid OA, to fully embrace CC-licenses.

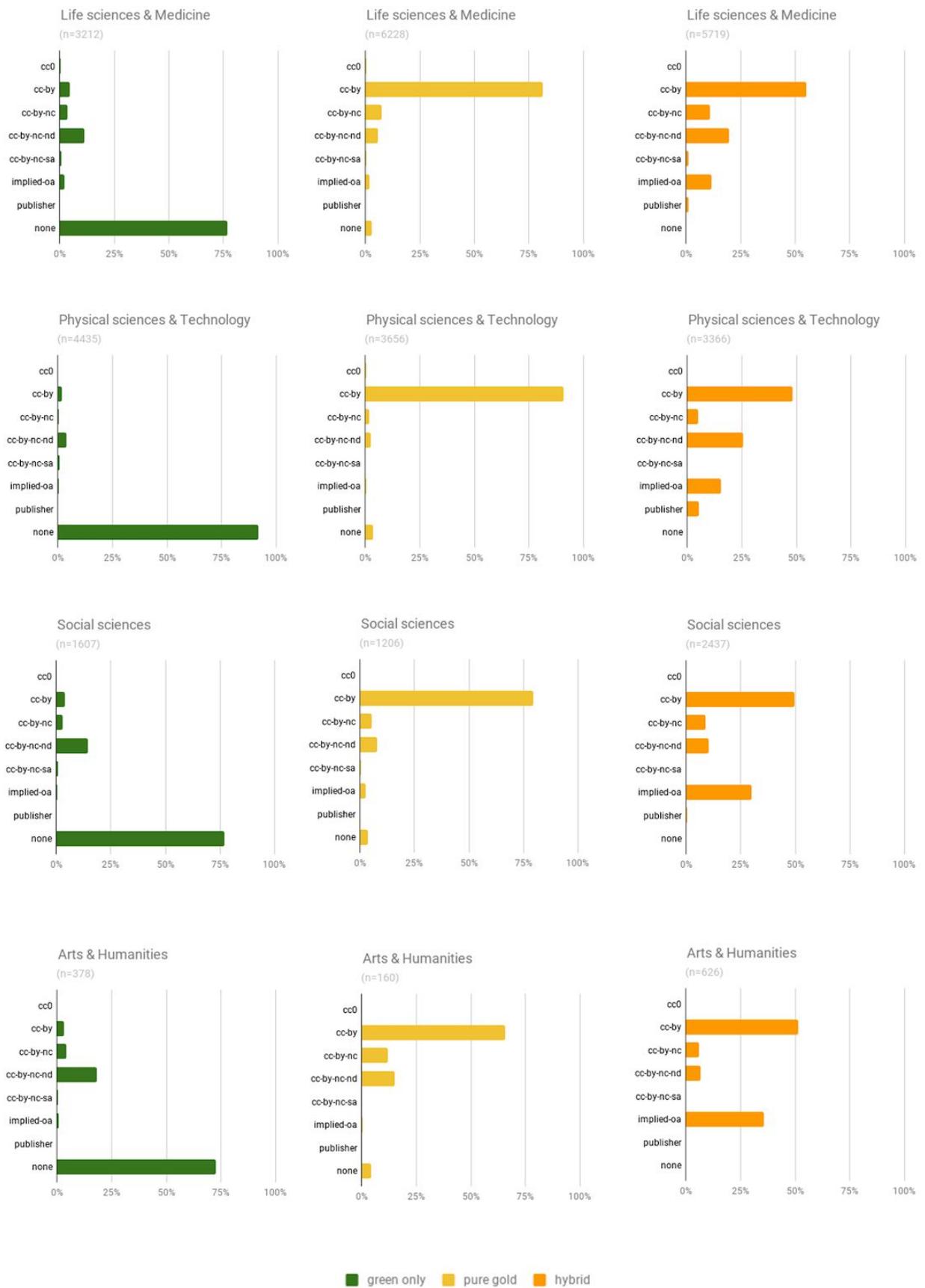


Figure 22. Licenses of open access output of Dutch universities, 2017, per academic field. Sources: Web of Science, Sopus, Dimensions, Unpaywall.

Open access books

Books are arguably the most important non-article publication type and as demonstrated play an important role in all fields, as full books or book chapters. That justifies a further look into information on the open access status of books in the main citation databases but certainly also in the additional databases, as citation databases have known limitations regarding book content coverage (see fig. 4). As figure 23 shows, three of the additional databases (BASE, NARCIS and OpenAIRE) indeed cover way more Dutch open access book material, with LENS having numbers comparable to those of Dimensions. Whereas Dimensions, LENS and Scopus only contain a few hundred records of open access Dutch book content, OpenAIRE, NARCIS and especially BASE cover up to several thousands of these. Theoretically there are two explanations. The same content can be detected as open access in the three repository aggregators but not in the three citation databases (e.g. because of lacking DOIs). Another explanation is that the citation databases lack in coverage of book content that is open access, e.g. chapters from humanities and social science book publishers that are shared as green open access in repositories but are not at all in the citation databases, as often with books from smaller publishers. More detailed research in this area is warranted.

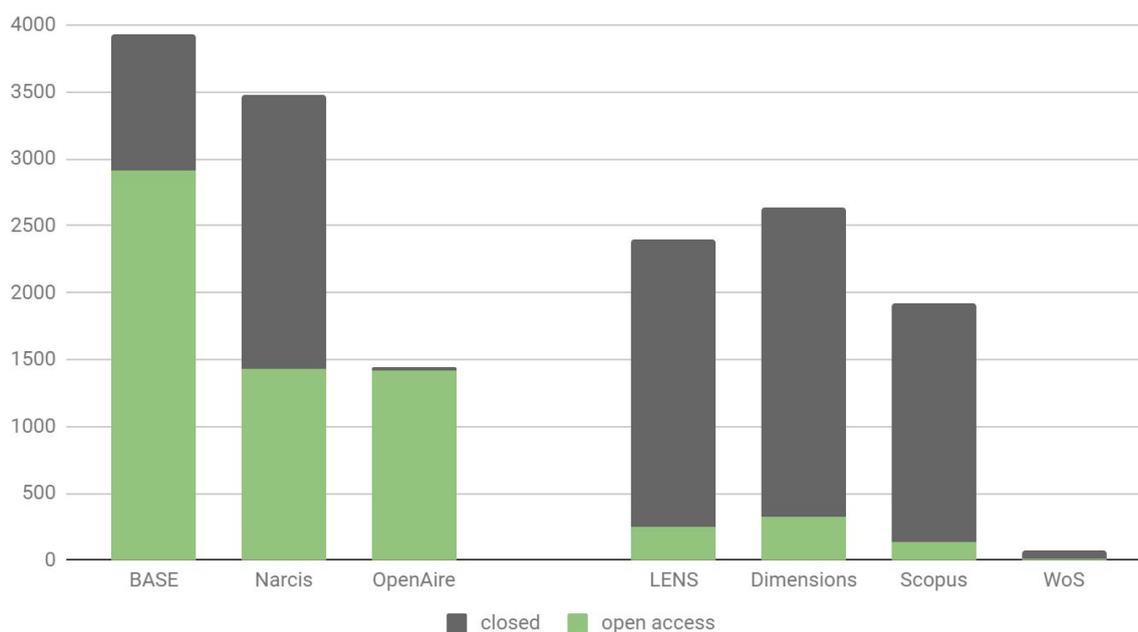


Figure 23. Coverage and openness of books & chapters from Dutch universities, 2017, in 7 databases. Sources: BASE, NARCIS, OpenAIRE, LENS, Dimensions, Scopus, Web of Science. NB our WoS version has no book citation index and OpenAire did not seem to aggregate non-OA documents. Check date 201811, except OpenAire and LENS: 201904.

One of the explanations for the lack of open access is that because of lacking DOIs open access even, if it is available, cannot be detected by the Unpaywall-based detection mechanism, which is also deployed by the citation databases themselves. However, as figure 24, with results from an analysis of book output of Utrecht University, shows, the book content in those databases overwhelmingly has DOIs assigned to it. So, the lack of open access book content in the citation databases is not caused by absent DOIs but by that content simply not being open access or not being detected as OA, depending on coverage

of Unpaywall. Increased coverage of Utrecht University book material would not immediately also lead to increased OA-levels for that in the citation databases as the content not yet covered is overwhelmingly (>80%) without DOIs. As long as that is not resolved, determining OA-levels for book output will be difficult with citation databases, regardless of coverage increase. That means that determining OA-levels for books can currently best be done in the CRIS of universities or in repository aggregators, if OA versions of books and chapters are indeed deposited in the repositories.

Finally, databases that are more dedicated to book content like collated library catalogues (e.g. Worldcat) or book search engines (Google Books) have limited value for these analyses, despite their impressive coverage, as they lack affiliation information.

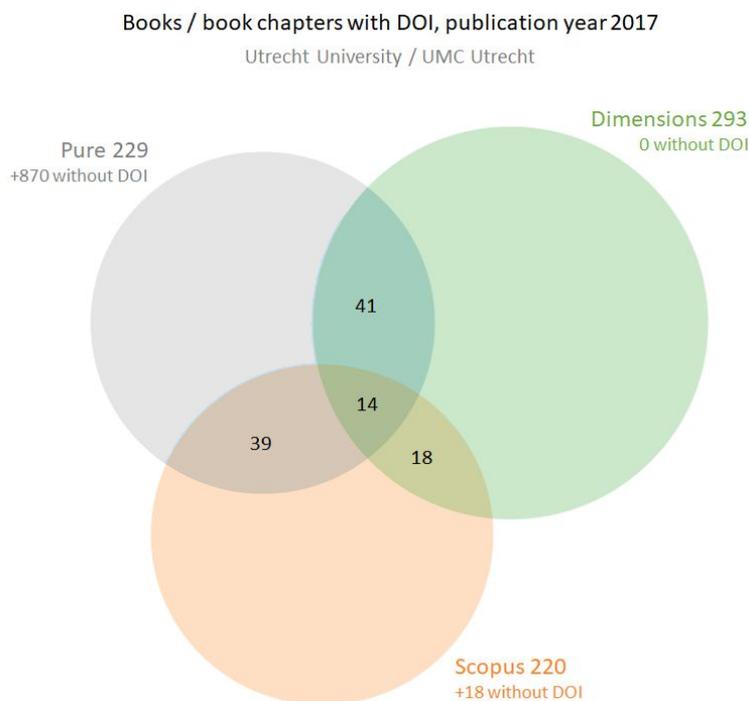


Figure 24. Overlap of database coverage of books and book chapters with and without DOIs in Utrecht University / UMC Utrecht output, 2017. Sources: UU / UMCU CRIS, Web of Science, Scopus, Dimensions.

4. Additional databases for filling gaps - an assessment

This report has frequently shown that citation databases have some limitations. Additional databases should be considered and indeed have provided some valuable data for the purpose of this study. That justifies looking more systematically at their potential for filling gaps but also at pinpointing any barriers in their application. Here we look at coverage and filtering, completeness, reusability and ease of mapping fields.

Coverage and filtering

Using databases with affiliation information beyond WoS, Scopus and Dimensions offers some additional coverage (table 2), for e.g. datasets, fully non-English material, and preprints. Alas, databases such as Google Scholar, Google Books, CORE and Worldcat, though useful for discovery, are not helpful for our purposes because they lack affiliation information. Though a valuable database, we left out ScienceOpen, because it has relatively limited publication type detail (it only discerns articles / posters / proceedings / peer review).

type	coverage (possibility)							
	WoS	Scopus	Dimensions	LENS	Base	Narcis	OpenAIRE	CRIS (Pure)
article	✓	✓	✓	✓	✓	✓	✓	✓
audio	x	x	x	x	✓	x	x	x
biographical item	✓	x	x	x	x	x	x	x
book	x	✓	✓	✓	✓	✓	✓	✓
book part / chapter	✓	✓	✓	✓	✓	✓	✓	✓
case note (law)	x	x	x	x	x	x	x	✓
conference paper / proceedings	✓	✓	✓	✓	✓	✓	✓	✓
data/software paper	✓	✓	✓	✓	✓	✓	✓	✓
dataset	x	x	x	✓	✓	✓	✓	x
letter	✓	✓	✓	✓	✓	✓	✓	✓
material fully in non-EN lang.	x	x	✓	✓	✓	✓	✓	✓
patent	x	✓	x	✓	✓	✓	✓	x
poster	x	x	x	x	x	x	x	✓
preprint	x	x	✓	✓	✓	✓	✓	✓
presentation slides	x	x	x	x	x	x	x	x
report	x	✓	x	✓	✓	✓	✓	✓
review	✓	✓	✓	✓	✓	✓	✓	✓
review of book/film/art	✓	x	✓	✓	✓	✓	✓	✓
video	✓	✓	x	x	✓	x	x	x

Table 2. Coverage of publication types in affiliation-enhanced databases beyond WoS, Scopus and Dimensions.

In some cases the publication type is covered but there is no filter to restrict the search to it, making analyses difficult. This is a problem with a number of databases for preprints, reviews of books/films etc., letters and data papers.

Because the main citation databases (WoS, Scopus, Dimensions) have known coverage limitations, it is interesting to see whether those limitations can potentially be overcome by using additional databases. Four databases that do at least allow filtering by affiliation country, may indeed have something to offer (fig. 25). LENS, BASE, NARCIS and OpenAIRE have obvious additional value for the coverage of (Dutch) output, and they are also openly available, which is a crucial advantage in allowing anyone to make (comparable) analyses.

Completeness (share of output captured)

In sheer size and coverage of the various non-article publication types NARCIS is much closer to completeness than either WoS, Scopus or Dimensions (fig. 25). It is even much more comprehensive than BASE and OpenAIRE, especially in capturing book output (note that reports are labelled as books). The smaller size of BASE and OpenAIRE compared to NARCIS for Dutch repository output is surprising because either directly or indirectly all source the same set of institutional repositories. Apparently OpenAIRE and BASE make some additional choices in their harvesting. For the completeness of information in repository harvesters, next to these harvesting choices of aggregators (and harvesting push decisions of repositories), it is important to have all output metadata in the repository, whether there is a full text available or not, whether that is open or not.

Two of the three databases that are mainly based on harvesting repositories (BASE, NARCIS and OpenAIRE) cover substantially more book material than the main citation databases and all three cover substantially more 'other non-article output' (such as doctoral dissertations and reports). LENS, a relatively new citation and patent database based on Crossref, Microsoft Academic and PubMed, does not cover larger numbers of non-article output compared to the main citation databases, though of course the exact records may be different and thus constitute additional coverage. In depth record level comparison is necessary to determine the exact amount of additional coverage.

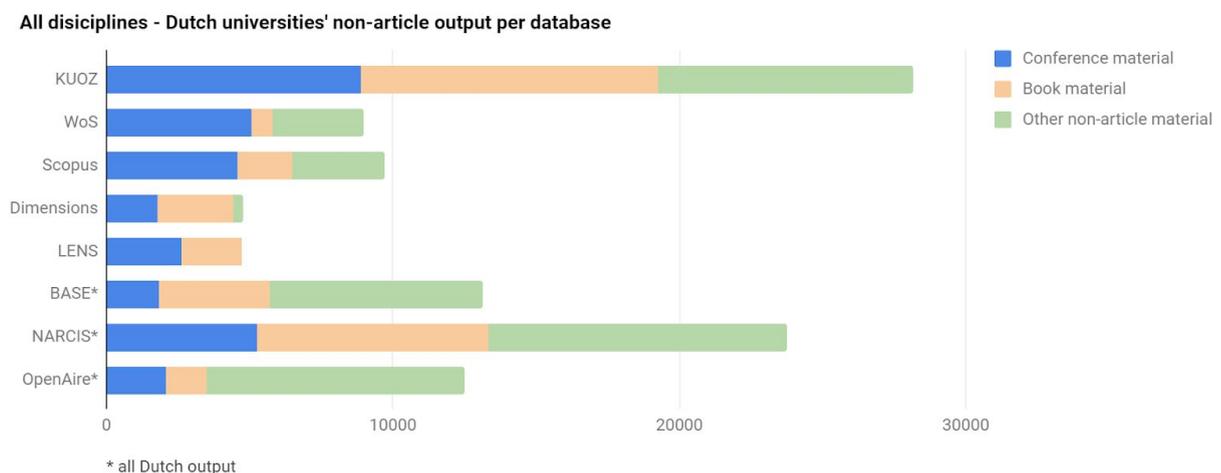


Figure 25. Dutch universities' main groups of non-article output of additional databases compared with KUOZ and main citation databases, 2017. Sources: VSNU, KUOZ Data 2017, Web of Science, Scopus, Dimensions, LENS, BASE, NARCIS, OpenAIRE.

Reusability of data, data license

Reusing data requires the technical possibility to download or import and the legal right to reuse them and ideally openly share them, for the sake of verification and reproducibility. Table 3 shows the current situation. It is clear that there are still barriers, despite these database being freely available. Especially more clarity on licenses and perhaps more open licenses (CC0) would be good for doing analyses and sharing results of that.

	LENS	BASE	NARCIS	OpenAIRE
Downloading options	<ul style="list-style-type: none"> • CSV / RIS / BibTEX / JSON export 	<ul style="list-style-type: none"> • API • OAI-PMH for partners 	<ul style="list-style-type: none"> • Harvesting with OAI-PMH 	<ul style="list-style-type: none"> • API • CSV export
Downloading restrictions	<ul style="list-style-type: none"> • Max. 50,000 records in export 	<ul style="list-style-type: none"> • Register IP-address or range • Specify use case 	<ul style="list-style-type: none"> • Need to register • Limited use of 'organizations'-field and "persons"-field 	<ul style="list-style-type: none"> • Max. 10,000 records per day via API • Max. 2000 records CSV export
License for downloaded data	<ul style="list-style-type: none"> • Data has ODC-BY license 	<ul style="list-style-type: none"> • No license specified 	<ul style="list-style-type: none"> • Downloading and reusing allowed 	<ul style="list-style-type: none"> • CC-BY

Table 3. Data usage allowances of additional databases.

Ease of mapping of fields

The more databases one needs to paint a full and nuanced picture of open access in all fields, the more mapping of publication/document types will have to be done. In combining information from LENS, BASE, NARCIS and OpenAIRE, mapping publication type categories is not easy and often introduced compromises. For example one database may have a category thesis, another doctoral thesis, still other separate ones for bachelor and master theses. Without diving into the actual records it is difficult to assess the effects of lumping these together, especially when numbers in those databases that do have fine-grained categories are very unbalanced and different with for instance one having thousands of Dutch doctoral theses but only a handful of bachelor and master theses and the other the opposite of that. A final remark on this: even if the same categories and labels for those categories are used, the definitions used behind the screens for assigning those categories could well be different, and the resulting cumulative categories used in studies combining information from various databases might still unknowingly be amalgams.

Solving issues of additional databases

It is difficult to assess the reliability of the data from repository harvesting databases. Without full agreement on standards for registration, repository aggregators are bound to have some unbalance caused by differing and changing practices at institutions and in countries. Harvesting of the repositories is automatic and metadata checking, enhancement and harmonization is labour intensive. This can be challenging when the organisations providing these additional databases are small, with limited resources. It is important though they receive and act on feedback on data quality. Some issues remain in practical use of these additional databases for analytical purposes. These issues are:

- Affiliation control and affiliation harmonization (all)
- Lacking functionality to filter on multiple institutions (BASE, NARCIS, OpenAIRE)

- Limitations on data export for off-site analysis (all)
- Limitations of data sources (e.g. lacking DOIs for BASE, NARCIS, OpenAIRE)
- Low bibliodiversity (LENS)

Some of these limitations could easily be improved (filter functionality, export allowances) or may improve over time (DOIs in repositories, more publication types in the sources that LENS uses). Affiliation harmonization requires substantial investment though (as long as institutional IDs are not added to affiliation information in publications). Here, the new [ROR-initiative](#) (Research Organization Registry) might act as a catalyst. ROR is a community-led project (in which Crossref also participates) to develop an open identifier for research organizations.

5. Conclusions and recommendations

This descriptive study confirms the expectation that focussing on just articles does not do justice to the great variety in academic output. This holds for all fields, but is especially pronounced in Social sciences and Arts/Humanities. Some examples: for Social sciences, important types of content that are lacking from many databases, including sometimes CRIS, are working papers (economics), case notes (law), and to a lesser extent reports (that are in repositories, but less in scholarly citation databases, also often lacking DOIs). Inconsistent labeling further means that reports are sometimes labeled as book material, making meaningful analysis even harder. For both social sciences and humanities, a sizeable amount of output has no DOI, hindering analysis with tools like Unpaywall for OA-detection and coverage in DOI-based databases (e.g. Dimensions). There is also the additional problem of databases assigning content to subject categories at the journal level, which potentially skews observed patterns.

On a broader scale, our first conclusion is on the awareness of (biblio)diversity. Differences in publication culture are easily overlooked. Academics themselves are often not fully aware of different values and practices in other fields and the effects of publication cultures easily escape superficial perception. Focussing on the most frequent (publication types) and the biggest (publishers) risks missing opportunities in research communication, for instance in promoting open access. Awareness of what is behind aggregated data, like KUOZ data, helps devising more targeted policies. The data presented here shows that it is possible to look further than the journal article. It also shows that providing more fine-grained descriptions leads to better understanding of differences in open access across fields, a topic of current discussion.

A second conclusion concerns database dependence. Almost invariably, it matters significantly what database you source information from when studying publication patterns. Coverage varies considerably, definitions and labels of publication types diverge, metadata differ. This means that one at least needs to be aware of the characteristics and limitations of the database used and of the effects and opportunities of using other or additional databases. Generally speaking it is advisable to always combine insights based on databases that are built on publisher- or Crossref data with insights based on databases that are harvesting information from repositories. The first may often provide more control, the latter often provide broader, more inclusive coverage.

Thirdly and lastly there is an important methodological conclusion. In describing the use of various publication forms in academic fields we found a number of issues that hinder a full and fair analysis of the work of academics and with that also stand in the way of effective design and evaluation of open access policies. The issues revolve around availability and completeness of metadata. If that is lacking any analysis will to some extent be plagued by problems of coverage, findability, traceability, comparability and reproducibility.

To improve the situation the various stakeholders could consider the following recommendations. These combine issues derived directly from this study with some relevant other developments.

- Authors/researchers
 - Whenever the publisher facilitates/allows it, use ORCID when publishing/sharing.
 - Choose platforms for sharing also based on availability of PIDs (DOIs) and ORCID.
- Publishers / publishing organisations
 - Make sure all books and reports as well as the chapters contained in them (if they have separate authors) have a DOI or comparable PID. This also holds for other content from publishers, including professional publications.
 - Use ORCID and facilitate using author contributorship roles, e.g. via the CRediT taxonomy.
 - Make all metadata, including citations, open.
- Universities/libraries
 - Align definitions of output types used in practice when creating records in a CRIS.
 - Make sure that anything that has or gets a DOI or other PID has that recorded in the CRIS and repository.
 - Make sure that registration of research outputs in CRIS matches changes in research communication culture (e.g. by also registering preprints).
 - Have all output metadata in the repository, whether there is a full text available or not, and whether that is open or not.
- Database providers
 - Offer an API, with the least amount of download restrictions.
 - Harmonize affiliation information.
 - Include and use standard IDs where available (for institutions (GRID or future ROR), authors (ORCID), funders (FundRef), journals (ISSN), publications (DOI) and conferences (ID in development).
 - Work towards an open citation database.
- Aggregators (including NARCIS)
 - Allow and create easy ways for exporting data (e.g. by offering an API-service).
 - License all data CC0.
- Open access detection tools (= Unpaywall)
 - Consider harvesting full OA journals outside DOAJ.
 - Consider harvesting sites offering full OA (and often CC-licensed) content outside DOAJ, e.g. open access book publishers.

The problem of lacking DOIs is a very clear one that cannot be solved overnight. Some recommendations for assessing open access of non-DOI output:

- Using repository and repository aggregator data, e.g. NARCIS.
- Using multidisciplinary databases that more comprehensively cover Arts/Humanities journal output (e.g. Google Scholar, JSTOR), though they lack affiliation info.
- Amassing data from field specific databases. Though most will have document/publication type filters, very few will have affiliation search that is need for this type of analysis (e.g. InspireHEP for high energy physics) and many lack mass export options.

- Using Google / Google Scholar scraping to find freely downloadable versions (see e.g. Martín-Martín et al, 2018). This is a very generic approach, but there are problems in metadata quality and it is very difficult to differentiate between types of open access.
- Using databases that include non-DOI output and have a comprehensive OA filter, e.g. 1Findr.

With some of these recommendations followed up and improvements in place it would be possible to create a more fair, accurate and nuanced insight of open access developments. It would allow for any institution or groups of institutions to create an overview - be it ad hoc or in the form of a monitor - of open access shares per publication type and per year and where relevant also per publisher. This 'holy grail' is within reach, but does require agreeing on standards and commitment and actions from various stakeholders.

References

- Bosman, J.M. and Kramer, B.M.R. (2018) Open access levels: a quantitative exploration using Web of Science and oaDOI data. [Preprint] PeerJ Preprints 6:e3520v1 <https://doi.org/10.7287/peerj.preprints.3520v1>
- Herzog, C. and Kierkegaard Lunn, B. (2018) Response to the letter 'Field classification of publications in Dimensions: a first case study testing its reliability and validity'. *Scientometrics* 117 (1): 641-645. <https://doi.org/10.1007/s11192-018-2854-z>
- KNAW (2013) Responsible research data management and the prevention of scientific misconduct. Amsterdam: KNAW. https://www.know.nl/nl/actueel/publicaties/responsible-research-data-management-and-the-prevention-of-scientific-misconduct/@@download/pdf_file/20131009.pdf
- Kramer, B.M.R. and Bosman, J.M. (2018) Open Access levels of Dutch universities' output 2016-2017 (articles & reviews): green, gold, hybrid and bronze - May 2018 [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.1251569>
- Kramer, B.M.R. and Bosman, J.M. (2019) Publication cultures and Dutch research output: a quantitative assessment (data). [Dataset]. Zenodo. <https://doi.org/10.5281/zenodo.2643367>
- Leeuwen, T.N. van, E. van Wijk, E. & P.F. Wouters (2016). Bibliometric analysis of output and impact based on CRIS data: a case study on the registered output of a Dutch university. *Scientometrics* 106: 1. <https://doi.org/10.1007/s11192-015-1788-y>
- Martín-Martín, A., Costas, R., van Leeuwen, T., & López-Cózar, E. D. (2018). Evidence of Open Access of scientific publications in Google Scholar: a large-scale analysis. *Journal of Informetrics*, 12(3), 819-841. <https://doi.org/10.1007/s11192-015-1788-y>
- NWO (2013) Onderzoek publicatieculturen NWO-MaGW: resultaten. The Hague: NWO. <https://doi.org/10.17026/dans-zng-a7c6>
- Nederlandse organisatie voor wetenschappelijk onderzoek (2016) Onderzoek publicatieculturen Sociale en Geesteswetenschappen. Den Haag: NWO. <https://www.nwo.nl/binaries/content/documents/nwo/algemeen/documentation/application/sgw/onderzoek-publicatieculturen-sociale-en-geesteswetenschappen/Onderzoek+publicatieculturen+Sociale+en+Geesteswetenschappen.pdf>. Published in English in 2018 as: Study into Publication Cultures Social sciences and Humanities. The Hague: Netherlands Organisation for Scientific Research. https://www.nwo.nl/binaries/content/documents/nwo-en/common/documentation/application/sgw/study-into-publication-cultures-social-sciences-and-humanities/Study+into+Publication+Cultures_ okt2018_UK.pdf.
- Piowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, Farley A, West J, Haustein S. 2018. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* 6:e4375. <https://doi.org/10.7717/peerj.4375>
- Salman, J., M. Kleinjans & D. Weijers, eds. (2012) Kennis over publiceren: publicatietradities in de wetenschap. Amsterdam: De Jonge Akademie. http://www.dejongeakademie.nl/shared/resources/documents/Kennis_over_publiceren_121218.pdf
- VSNU (2017) Definition framework monitoring Open Access. VSNU. http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/Open%20access/Definitief%20Definition%20framework%20OA_VSNU-20160217.pdf [accessed April 9 2018]
- VSNU (2018) Definitieafspraken wetenschappelijk onderzoek - toelichting bij KUOZ. Den Haag: VSNU. https://www.vsnu.nl/files/documenten/Feiten_en_Cijfers/VSNU_Definitieafspraken_onderzoeksinzet_en_output_KUOZ.PDF
- VSNU (2018) Percentages of open access publications in 2016 and 2017. VSNU. https://www.vsnu.nl/en_GB/percentages-open-access-publications-2016- [accessed April 9 2018]