Inventory of open textbooks

Working group project report: Reusing open educational resources

Acceleration plan
Educational innovation with ICT

digital educational resources

Inventory of open textbooks

Working group project report:
Reusing open educational resources

Acceleration Plan for Educational Innovation with IT
Zone Digital (open) educational resources

Acceleration plan
Educational innovation
with ICT

Ruben Jacobse
Michiel de Jong
Sylvia Moes

Editor
Nicole Bakker

Contents

Introduction

The transition from commercial to digital and open educational resources provides a host of opportunities for the professional development of lecturers. It also reduces dependence on commercial publishers for the publication of educational literature. Despite this, the transition is still in its infancy in the higher education sector in the Netherlands.

The Reusing open educational resources working group of the Towards digital (open) educational resources zone of the Acceleration Plan Educational Innovation with IT recognises that opportunities exist to stimulate this transition. The working group expects that, if it is much easier to reuse open textbooks, they are more likely to be used and reused.

Lecturers use teaching materials that they have control over, and open textbooks can be used flexibly in lessons and degree programmes. Furthermore, open textbooks give students much greater access to educational literature, since they no longer have to depend on relatively expensive textbooks with restricted access.

The working group is part of the Towards digital (open) educational resources acceleration zone. In this zone, representatives of six universities and two universities of applied sciences in the Netherlands are working on optimising the use of digital (open) educational resources in higher education. The zone aims to make it easier for lecturers and students to compile and use an optimal mix of educational resources, selected from an increasingly diverse range of materials and aligned with the educational vision of the lecturers and institution, as well as the needs of the students.

However, if we are to facilitate the use of open textbooks, we first need to know what is available. As no such overview has been made for the Netherlands, the working group has made a start on producing one in this project. The inventory of the available literature focuses on open textbooks that are available to the general public and that meet the selection criterion ‘suitable for university education’. The inventory represents the first part – phase I – of the project.

Phase I of the project involved the following activities:

1. Inventories of open textbook repositories made by Delft University of Technology and VU Amsterdam were combined and extended, resulting in a fairly comprehensive overview of sources of open textbooks.
2. The open textbooks in these repositories were combined in a single overview and categorised by subject area and other metadata.

Based on the inventory of open textbooks, an analysis was then made of their adaptability. An important aspect of open textbooks is the opportunity that they give lecturers to adapt the material for use in their own teaching. We found that many open textbooks are published in a digital format that makes it difficult or impossible to modify the material. The working group therefore conducted a technical review of software for converting difficult-to-adapt files into easily modifiable formats.

Phase II of the project involved the following activities:

1. An inventory was made of the most common formats of open textbooks in the repositories.
2. A technical review was conducted of methods for importing open textbooks into an environment in which they can be adapted. An analysis was also made of how to reduce an open textbook to a series of unformatted elements, to make it easier for lecturers to adapt the content to meet their needs.

This report describes the implementation and results of both phases of the project, plus comments and recommendations for future study. Note that it is an abridged version; the full report including technical details can be requested by sending an email to: leermaterialen@versnellingsplan.nl.

1 Phase I: Inventory of open textbooks

1.1 Introduction

Open textbooks are available in various repositories and from various publishers. There are also platforms that index and provide access to the textbooks offered by these repositories and publishers, through a process known as 'harvesting'. Several of these platforms give users the possibility to export the metadata of their collections, the four main platforms being Directory of Open Access Books (DOAB), British Columbia Open Textbooks (BCOpen), Open Textbook Library (OTL) and Unglue.it.

The content of the collections provided by these four platforms is analysed using metadata, which is encoded in the Machine-Readable Cataloging (MARC) standard for bibliographic information. This is provided in binary MARC21 format with the extension .mrc, or in MARCXML format with the extension .xml. The bibliographic information is however not complete for every textbook, and the metadata describing the subject area of the textbook is often missing.

This lack of descriptive metadata can however be resolved by also using the Library of Congress Classification (LCC) system. For every book that has been assigned an LCC code, mapping can be applied to determine which *Edustandaard* subject field the textbook falls under. The resulting data can then be summarised and analysed both as a complete collection and as collections by subject area. This, in brief, is the methodology applied by the working group.

We discuss this process in more detail in this chapter. Section 1.2 starts with a description of the sources consulted to produce the inventory of open textbooks. Sections 1.3 to 1.5 then go into further detail on the steps taken to process the metadata obtained from the platforms mentioned above and to enrich this metadata with LCC class. Section 1.6 presents the results of the data processing steps in the various analyses, and comments and recommendations are made in Section 1.7.

1.2 Online sources of open textbooks and metadata

The analysis in this report is restricted to a few generally accessible platforms that provide open access to their metadata. Platforms that do not provide access to their metadata or that cannot be accessed by the general public, such as platforms that require users to register or subscribe, are briefly discussed but further disregarded in this analysis. Metafinders such as Merlot, OER Metafinder and OASIS, which search other platforms but do not themselves

offer books, are also not included in this inventory, as their metadata is not generally accessible.

A total of 95% of the metadata of the inventoried books in the dataset is from DOAB, while the remaining 5% is from BCOpen, OTL and Unglue.it.

Directory of Open Access Books (DOAB)

Directory of Open Access Books[1] (DOAB) offers ~31,400 unique titles, making it the largest collection of open access books on the internet by far. DOAB mainly offers titles from academic publishers and open access titles from commercial publishers (Springer Nature ~2,400 titles, DeGruyter ~1,700 titles, Brill ~640 titles, Taylor & Francis ~880 titles). DOAB also offers titles from other platforms such as IntechOpen[2] (~7,000 titles) andn OpenEdition[3] (~15,800 titles).

British Columbia Open Textbooks (BCOpen)

The British Columbia Electronic Library Network supports the British Columbia Open Textbooks Collection (BCOpen)[4] and offers approximately 320 titles. BCOpen offers approximately 70 titles from its own collection plus titles from a number of Canadian government agencies, smaller platforms such as OpenStax[5] (Rice University, 41 titles), a limited number of titles from Milne Open Textbooks[6] (SUNY Geneseo, seven titles of a total of ~100), some titles from Saylor Academy[7] (nine of ~100 titles) and titles published by other universities.

Open Textbook Library (OTL)

The Open Textbook Library[8] is supported by the Open Education Network of the University of Minnesota and offers approximately 800 titles. It also offers titles from the collections of BCOpen (~37 titles), OpenStax (~58 titles), Milne Textbooks (~25 titles) and Saylor Academy (~54 titles). Furthermore, this platform offers titles published by various universities.

1 www.doabooks.org
2 www.intechopen.com
3 books.openedition.org/
4 open.bccampus.ca
5 openstax.org
6 milneopentextbooks.org
7 www.saylor.org/books/
8 open.umn.edu/opentextbooks/

Unglue.IT

Unglue.it[9] is a platform that authors and publishers can use to distribute their books under free licences. The platform offers approximately 800 titles, mostly from academic publishers. A small number of books in the collection are fiction.

Other (unused) sources

The four platforms named above provide free access to one or more forms of metadata of their collections. There are other platforms, but these either do not offer any metadata at all, or only on the condition that the user signs up or subscribes. As the metadata for these collections was unavailable during this project, or because an inventory of these platforms could only be made manually, or semi-manually, they were not included in the analysis. For the sake of completeness, however, these platforms are briefly described below.

JSTOR[10] offers libraries a metadata file of its collection via OCLC Worldshare Collection Manager (~6,500 titles). An Excel file with limited metadata such as title, author and ISBN is also published. A random sample of these titles shows that most of them are also available in the DOAB collection. The JSTOR metadata file was not requested from OCLC for this project.

LibreTexts[11] is a platform with ~400 titles in 13 different categories. The platform also offers titles from OpenStax. LibreTexts offers downloads in various formats, such as PDF, IMSCC (Canvas/Brightspace) and as web pages. LibreTexts also offers a 'remixer', which is a tool that lecturers can use to compile and remix content from different books.

BookBoon[12] is a platform that offers more than 1,000 open textbooks for subscribers. Most books are free for students, although they need to create an account.

Open Culture[13] has a web page with 200 links to the original sources of open textbooks. The list also includes several OpenStax books.

9 unglue.it/
10 guides.jstor.org/oabooks
11 libretexts.org/
12 bookboon.com
13 www.openculture.com/free_textbooks

The **American Institute of Mathematics**[14] provides a list of links to the original sources of approximately 60 open mathematics textbooks that have been reviewed by the institute. Most titles are available in PDF as well as HTML, XML or LaTeX format. Most of these books appear not to be available on other open access platforms.

1.3 Step-by-step plan for processing metadata

The platforms that do provide metadata for their collections use different file formats. BCOpen and OTL use MARC21 (.mrc), while DOAB and Unglue.it use MARCXML (see Section 1.4 for further details). DOAB also provides a .csv file with metadata. While the content of this file is almost the same as the MARCXML file, there are a few differences in the data fields.

Figure 1 shows the processing of metadata.

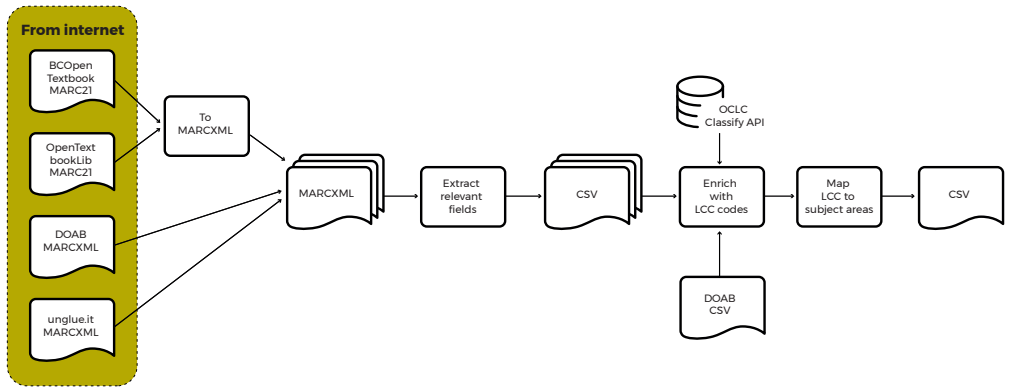


Figure 1 Schematic overview of the steps taken to process the metadata.

The metadata processing steps are explained further below:

- The four source files were downloaded from the internet in MARC21 (.mrc) and MARCXML (.xml) formats. The source file from BCOpen is generated quarterly; the source files from the other three platforms were generated automatically when they were downloaded. The URLs for these files are provided in Table 1.
- The MARC21 files were converted to MARCXML, a file format that is both human and machine readable.
- The data was extracted from the fields relevant to this project in the files in the hierarchical MARCXML format and processed in tabular form in .csv format, with one row per record. The metadata was then ready for enrichment.
- Because the required fields were missing for some books, the .csv files were enriched with extra data from other sources.
- For open textbooks from DOAB, the LCC classes provided in the online .csv file were added to the .csv file obtained from the MARCXML file.
- For metadata without LCC from other platforms, an application programming interface (API) was used, the OCLC Classify API. Using this API, the information was obtained from the ISBN that is needed to determine the LCC class.
- The LCC fields were then used to determine the subject area or areas to which the book belongs, according to *Edustandaard* categories. A book may be categorised under more than one *Edustandaard* subject area, as it may have more than one LCC class. Furthermore, some LCC classes fall under several *Edustandaard* subject areas.

Table 1 URLs of online metadata sources

File	URL
DOAB (.csv)	www.doabooks.org/doab?func=csv
BCOpen (.mrc)	bceln.ca/sites/default/files/marc_records
OTL (.mrc)	open.umn.edu/opentextbooks/download.mrc
DOAB (.xml)	www.doabooks.org/oai
Unglue.IT (.xml)	unglue.it/marc/all

1.4 Conversion of MARCXML to .csv format

The files downloaded from the internet in MARC format also contained fields with information that is not relevant to this project. For this reason, not all fields were included in the conversion to a .csv file in tabular format. The other MARC fields were therefore not used. As data is sometimes displayed in different ways and across several fields in MARC format, an extra step was applied during the conversion to a .csv file. The MARCXML data was then exported and combined if necessary. Table 2 shows the relevant .csv columns and the corresponding MARC fields.

[14] aimath.org/textbooks/approved-textbooks/

Table 2 .csv columns and corresponding MARC21 fields

.csv kolommen	MARC21 veld(en)
ID	001
Title	245a (title) + 245b (subtitle)
Edition	250a 250b
Author	100a or 700a (or 245c)
Language	546a or 041a
ISBN	020a (possibly several times)
LCC	050a + 050b
DOI	024a (only as 024_2 == “doi”)
Subject	650a (possibly several times)
Keywords	653a (possibly several times)
Form	650v (possibly several times)
Summary	520a
Contents	505a
Series title	490a
Series ISSN	490x
Publisher	260b
Year of publication	260c or 542g
Copyright licence	540a or 540f or 542f
File format	347b
URL of source file or web page	856u

When converting an MARCXML file to a .csv file, a number of processing steps are required to ensure that the data is displayed uniformly in some fields. MARC does not apply any conventions for formatting the data fields, so that providers of the various sources are free to record the same type of information in slightly different ways. To ensure a more uniform display of the data in this project, the following processing steps were therefore carried out:

- repair swapped fields in files;
- uniform display of titles, author names, keywords, etc;
- recognise Creative Commons licence formats and save in same format;
- recognise languages in language field and translate if necessary.

1.5 Data enrichment

In Section 1.3, we mentioned that not all of the fields in the MARC/MARCXML files contain data. In this section, we explain how the data was enriched with extra information, obtained using the LCC class. Further steps were then taken to classify textbooks by subject area.

DOAB offers a .csv format export of the LCC class of most records, but this information is missing in the XML version of the DOAB data. Using this LCC data, the missing data in the .xml file can be enriched with the missing classes, after the .xml file has been converted to a .csv file.

The OCLC Classify v2 API can also be used, to query the class of a book using an identifying characteristic (e.g. ISBN, title or OCLC number).

Using the LCC code, which is either already contained in the metadata of the open textbooks or is added later using the steps described above, the open textbooks were categorised by subject area. The LCC code of a textbook includes a prefix, and this describes the subject area of the textbook.

Mapping was then carried out to link the LCC prefixes (the first or first two letters of the code) to the subject areas in the *Edustandaard* taxonomy. The mapping applied is shown in Table 3. Note that this mapping system was developed for the purpose of this study. It has therefore not been subject to expert review and should not be regarded as definitive or absolute.

Table 3 Possible mapping of LCC prefixes to *Edustandaard* subject areas

Edustandaard subject areas	LCC prefixes
Earth and environmental sciences	G, GOES, GB, GC, GE, GF, QE, QH, QK, QR, S, SB, SD, SF, SK
Economics and business	GV, HB, HC, HD, HE, HF, HG, HJ, TX
Physical and information sciences	BC, QA, QB, QC, QD, QE
Behavioural and social sciences	BF, GN, GR, GT, GV, H, HA, HM, HN, HQ, H, HT, HV, HX
Health sciences	QL, QM, QP, R, RA, RB, RC, RD, RE, RF, RG, RJ, RK, RL, RM, RS, RT, RV, RX, RZ, SH
Interdisciplinary	U, UA, UB, UC, UD, UE, UF, UG, UH, V, VA, VB, VC, VD, VE, VF, VG, VK, VM
Art and culture	AZ, B, BC, BD, BH, BJ, BL, BM, BP, BQ, BR, B, BT, LTD, BX, C, CB, CC, CD, CE, CJ, CN, CR, CS, CT, D, DA, DB, DC, DD, THE DF, DG, DISTRICT HEAD, DJ, DK, DL, DP, DQ, DR, D, DT, DU, DX, E, F, M., ML, MT, N, AFTER, NB, NC, ND, NE, NK, NX
Education	L, LA, LB, LC, LD, LE, LF, LG, LH, LJ, LT
Law and political science	J, JA, JC, JF, JJ, JK, JL, JN, JQ, JS, JV, JZ, K, KB, D, KE, KF, KG, KH, KJ, KK, KL, KZ
Language and communication	P, PA, PB, PC, PD, PE, PF, PG, PH, PJ, PK, PL, PM, PN, PQ, PR, PS, PT, PZ
Engineering and technology	NA, T, TA, TC, TD, TE, TF, TG, TH, TJ, TK, TL, TN, TP, TR, TS, TT, VM

1.6 Summary of relevant information in metadata

IIn this section, we present and discuss the results of the data processing steps described in the previous sections. The analysis is split into two parts: the full set of metadata of all books in this section, and an analysis per *Edustandaard* subject area in Appendix A.

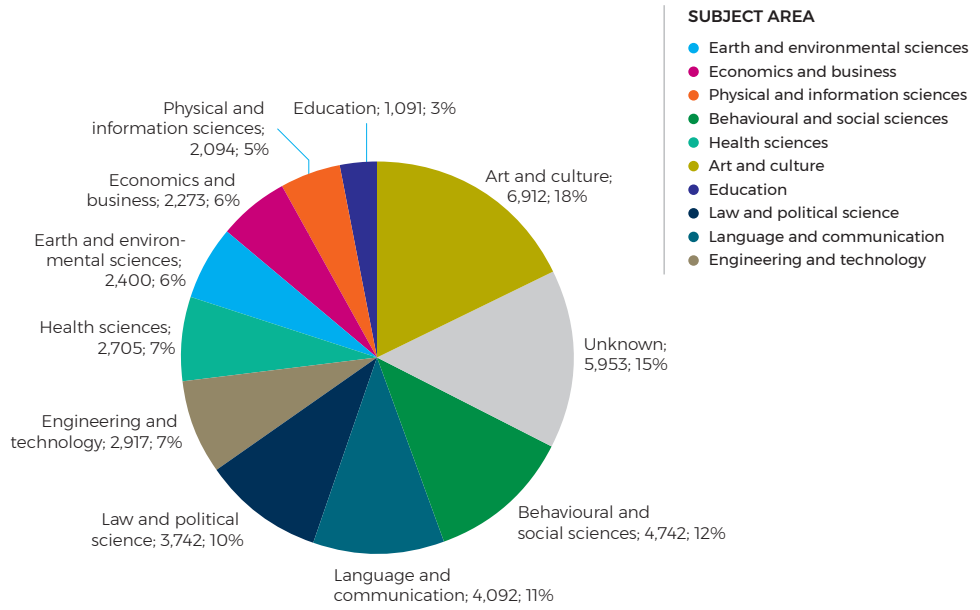
Overview of all titles

The figures on the following pages are derived from the analysis of the full set of metadata of all books. The open textbooks were analysed according to the following categories:

- Subject area
- Source portal
- Publisher
- Language
- Licence type
- File format

Subject area

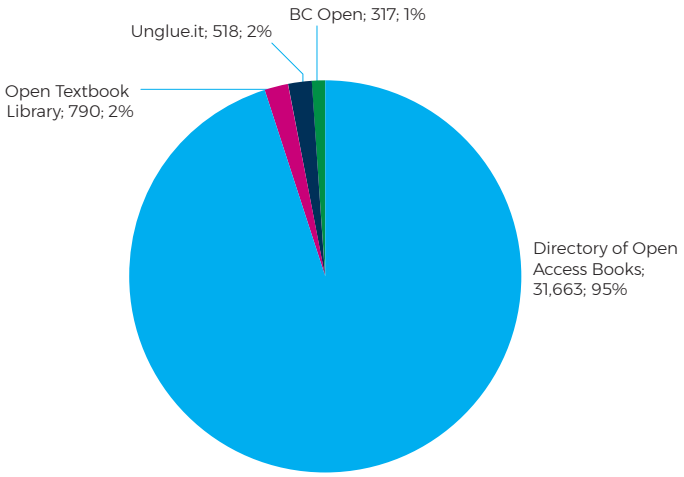
The figure below shows the number of open textbooks in each of the 11 *Edustandaard* subject areas. Following enrichment of the metadata, an LCC class was determined for approximately 27,500 books. Almost 6,000 books could not be assigned a subject area and were therefore labelled 'unknown'. Note that a book may be assigned more than one LCC class and therefore fall under more than one subject area. As the figure shows the number of books per subject area, the sum of the books in each subject area is therefore greater than the total number of books.



Source portal

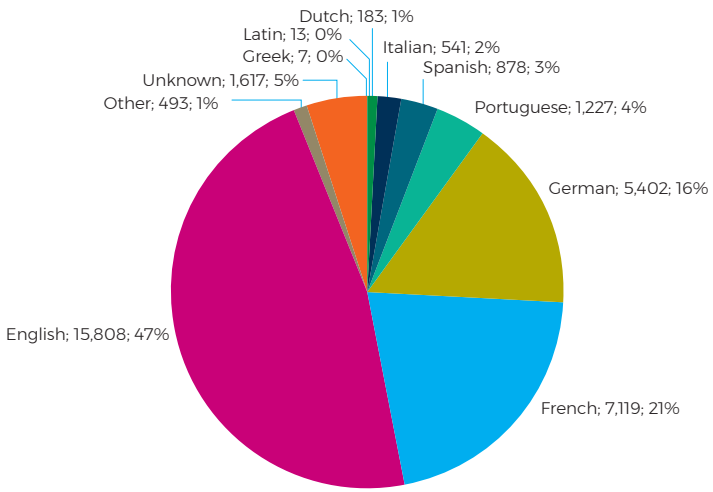
The source portals are the platforms from which the metadata was obtained for this study. Most of the titles (~32,000, or 95%) were obtained from DOAB. This was as expected, because DOAB offers academic open access literature in several different formats. DOAB is not limited to open textbooks for higher education, but also offers other collections. The other

platforms focus much more on educational textbooks. Although the absolute numbers in these repositories are therefore much lower, the textbooks that they do offer are more targeted to the education sector.



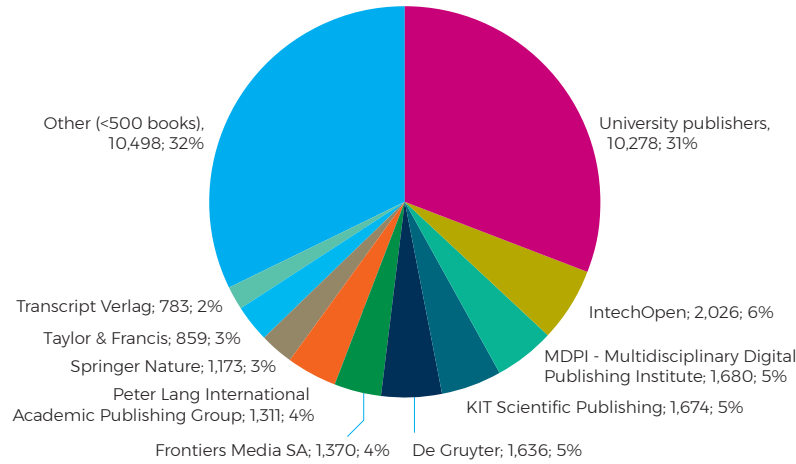
Language

The figure below shows which languages the open textbooks are written in. Books are labelled as 'unknown' if the language is not provided in the metadata. Languages with a small number of books are collectively labelled 'other'. For some books, the language field in the metadata contains several languages, making it difficult to determine which one is the correct language. These books are also classified as 'other'.



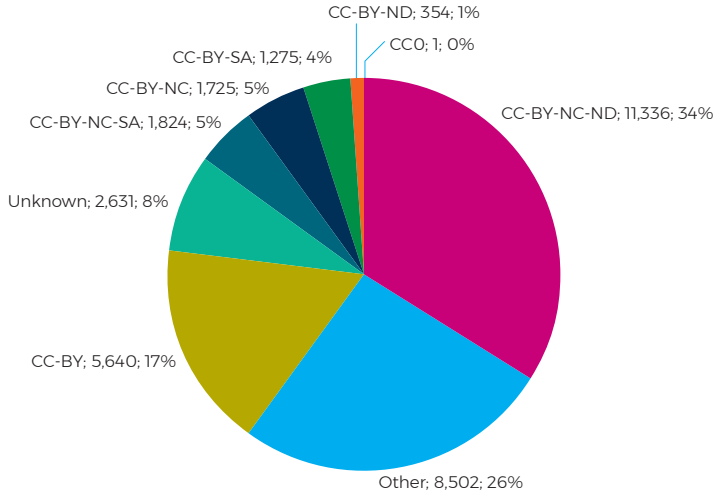
Publisher

The figure below provides an overview of the publishers of the open textbooks. Many books are published by university publishers, mainly in Europe and North America. Because there are dozens of 'university presses' that each publish dozens to hundreds of books, these have been collectively labelled 'university publishers'. All other types of publishers that publish fewer than 500 books each are labelled 'other'. The open textbooks from these publishers were harvested and listed in DOAB.



Licence type

The figure below shows the type of licence under which the books are published. Most books (66%) were published under a Creative Commons licence (shown in the figure as CC-..). Books that were published under a variety of publisher-specific licences are labelled 'other'.



File format

The table below provides an overview of the file formats in which the titles are available, according to the metadata. For almost three quarters of the titles, this information is not provided in the metadata files and the file format is therefore unknown. Often, the metadata for these books simply contains a web page URL from which the file can be downloaded. In some cases, the metadata includes the file format, but inspection of the web page then shows that the book is available in different formats. This applies for example to many titles in the BCOpen dataset.

Table 4 Overview of file formats and their numbers in the metadata

File format	Number
PDF	9,774
XML	39
ePub	33
Unknown	23,442

Overview by subject area

A more concise overview of the data was made for 10 of the 11 *Edustandaard* subject areas. As none of the books in the dataset were classified as 'Interdisciplinary', no overview was made of this subject area.

The following categories are provided for each *Edustandaard* subject area:

- Language
- Year of publication
- Licence type
- Keywords

Word clouds were generated based on the keywords in the 'keywords' field of the tabular .csv metadata file. These word clouds provide a clear summary of the most common keywords in each subject area. The 200 most common keywords are shown, and the larger the keyword, the more common it is. Note that only keywords from English language books were used. The word clouds were generated using the Python software package Word Cloud[15]. As the differences between subject areas were relatively small for the other categories, no word clouds were made for these. The results can be seen in Appendix A.

1.7 Comments and recommendations

Various steps were taken to produce the inventory of open textbooks. Based on the observations made during this process, we would like to make the following comments and recommendations:

- For many books, there is no LCC code in the metadata, even though it is often provided on the front flyleaf of the book itself. This can be solved by extracting the LCC codes from the full text.
- We recommend that the mapping of LCC prefixes to *Edustandaard* subject areas as carried out in this study is validated by experts in this field. Mapping to *Edustandaard* degree programme can also be achieved by using more information than just the first letter or letters of the LCC code.
- The file formats are generally absent or incomplete in the metadata (for example, many titles on the BCOpen website are offered in several formats). A possible solution to this is to add this information using web crawlers that search the URL of the source.
- There is some uncertainty concerning the accuracy of the licence types in the metadata. For example, DOAB frequently names the licence provided in the book, but this may differ from the licence under which the book is offered on a website.

[15] pypi.org/project/wordcloud/

- The language of a book is also sometimes unclear in the metadata, or can be interpreted in different ways. It may therefore be better to determine the language using the extracted text from the document itself.
- The .csv file provided by DOAB, with LCC codes and other information, contains errors that need to be repaired manually by the user before the file can be used to enrich the metadata. The working group has reported this problem to the organisation behind the DOAB platform, and hopes that they will repair these errors.
- Now that this inventory is complete, methods need to be developed to search these lists of textbooks easily. The findability of textbooks using this inventory is currently restricted to the disciplines and subject areas of the textbooks described in the metadata. Manually searching the platforms is not an option, as it is too time-consuming for lecturers and students.
- We tried to find a way to automatically analyse the content of open textbooks, with the aim to generate much more descriptive information. This information could be used to link open textbooks to the degree programmes taught in the various higher education institutions. For this, a proof of concept was carried out using the software Elasticsearch.

ElasticSearch[16] is a search engine developed by Elasticsearch B.V. and is generally available as open source software under the Apache software licence[17]. Elasticsearch is suitable for searching the full text of a large number of documents, making it highly suited to fully indexing and searching a collection of textbooks. There are people in SURF who are acquainted with the use of Elasticsearch, and we therefore expect that it should be relatively easy to integrate a search function for open textbooks into existing platforms such as Edusources. This proof of concept has been made possible thanks to the expertise and technical advice provided by Jelmer de Ronde at SURF. Please send an email to leermaterialen@versnellingsplan.nl for a more detailed description of the technical implementation of this proof of concept.

[16] www.elastic.co/elasticsearch/
[17] www.apache.org/licenses/

2 Phase II: Adaptability of open textbooks

2.1 Introduction

An important reason for using open textbooks is the opportunity that they provide for reusing the material. The adaptability of open textbooks is therefore an important property of these books. However, many open textbooks are published in a format that makes it difficult or impossible to modify the material, for example in a PDF or an ePub format.

In the second part of this project, the working group carried out a technical review of software that can be used to quickly and easily convert files that are difficult to modify into easily adaptable formats, taking into account the applicable Creative Commons licence conditions.

The following activities were carried out:

1. An overview was made of the most common file formats of the open textbooks found in the repositories in the first phase of the project. The open textbooks that are of interest to universities in the Netherlands are provided by various international publishers and platforms.
2. A technical review was then conducted of software that can be used to convert open textbooks that are difficult to adapt, such as PDF and ePub, into files with editable formats, so that lecturers can easily modify these textbooks.
3. An overview was then made of possible solution pathways for modifying open textbooks that are published in formats that are difficult to adapt. This is provided in the form of an overview of scenarios for books containing different types of content, such as text, images, tables, vector diagrams, and so on.

2.2 Overview of file formats

The results of phase I of this project (see Section 1.6) indicate that it is difficult to determine the formats of open textbooks. To obtain a better idea of this, we conducted a random sample of the 17 largest publishers, who together account for approximately half of the titles in the dataset, and examined the formats of the books on their websites. The smaller BCCampus was also included, as most of the titles on this platform are offered in several different formats.

Table 5 Overview of file formats per publisher, based on random sample

Publisher	Number of titles	File formats
IntechOpen	2,026	Only PDF and HTML web page
MDPI - Multidisciplinary Publishing Institute	1,680	Only PDF
KIT Scientific Publishing	1,674	Only PDF
De Gruyter	1,636	Always PDF, sometimes also ePub
Frontiers Media SA	1,370	Always PDF and ePub
Peter Lang International Academic Publishing Group	1,311	Always PDF, sometimes also ePub
Springer Nature	1,173	Always PDF, often also ePub
Taylor & Francis	859	Only PDF
Transcript Verlag	783	Always PDF, sometimes also ePub
Amsterdam University Press	638	Only PDF
Australian National University Press	584	Always PDF, sometimes ePub, HTML, MOBI
Coimbra University Press	543	Only PDF
Universitätsverlag Göttingen	517	Always PDF, sometimes ePub
Brill	447	Only PDF
Böhlau	347	Only PDF
Open Book Publishers	312	Always PDF, sometimes XML (TEI[18]) or online HTML; often ePub and MOBI on payment
Manchester University Press	294	Always PDF, sometimes HTML download
BCCampus	71	Always PDF, often also HTML, ePub, MOBI and XML (Pressbooks[19]), sometimes DOCX/ODT

[18] tei-c.org/guidelines/
[19] pressbooks.com/

The random sample shows that almost all titles are offered in PDF format, and that the second most popular format is ePub. MOBI, the various XML formats and HTML web pages or DOX/ODT documents are much less popular. These file formats are described in more detail in Sections 2.3–2.7, including possibilities for modifying these formats.

2.3 Documents in PDF format

The design of the PDF standard ensures that every document is displayed with the same layout, on every device and in every PDF reader application. This is achieved by using elements to construct documents in PDF format that each have specific coordinates on a page. In contrast to a Microsoft Word document, for example, a PDF document does not have a structure containing information on the 'roles' of certain text, such as main text, title and page number. Once a document is saved as a PDF, this information is therefore lost, making it almost impossible to modify a PDF.

Complete pages or chapters can be copied without modification using a tool such as PDFsam[20] Or, a complete PDF can be converted to a different format using Adobe Acrobat Pro[21]. However, the limitations of the PDF format mentioned above still apply. When converting to .docx format, for example, the new document has no concept of the role of different blocks of text in the document, and text blocks cannot extend beyond a single page. This makes it almost impossible to modify large blocks of text without creating a new document.

The best option for PDF documents is therefore to extract as many useful elements as possible and to process these manually to create a new file in a format that can be modified. However, the way in which a PDF is built means that there are unfortunately no tools that can do this fully automatically.

For extracting all the text elements in a document, the tools Apache Tika and the Poppler package[22] produce reasonable results. However, Poppler is mainly intended for use under one of the Linux distributions, but with a few extra steps it can also be used in Windows.

The Poppler project also includes several 'utilities', each of which have a specific task. For example, the programme *pdftotext* (which is similar to Tika) only extracts the text elements from a PDF document, which it stores in a text file. The *pdftoxml* utility makes

[20] pdfsam.org/
[21] acrobat.adobe.com/nl/nl/acrobat/acrobat-pro.html
[22] poppler.freedesktop.org/

a similar extraction, but also extracts information such as font and page coordinates for each text block in the PDF. Similar is the pdftohtml utility, which creates HTML files from a PDF and attempts to retain as much of the original layout as possible (e.g. including raster images). These raster images can then be extracted using *pdftoimage*.

It is almost impossible to edit tables, diagrams and equations obtained from a PDF document. While various web sites claim to be able to do so, a random survey that we conducted showed clearly that the results are inadequate, in addition to the licencing issues. A better option is therefore to make a screenshot and insert this into the new document.

Because of these formatting issues with PDF documents, we recommend using the source document that was used to create the PDF wherever possible. This may require contacting the author or publisher, as these documents are not usually made available to the general public.

2.4 Documents in ePub format

Files with an ePub extension are actually zip archives 'in disguise' that contain HTML files. These files are primarily intended for eReaders. Simply by renaming 'abc.epub' to 'abc.zip', the archive can be unzipped and the HTML files edited using a text or HTML editor. The documents can then be viewed in a web browser. There are also software packages for editing ePub files directly, such as Calibre[23] and Sigil[24].

2.5 Documents in MOBI format

Files with the .mobi extension are primarily intended for Kindle eReaders. They can be converted into ePub format for further editing using the Calibre software package.

2.6 Documents in HTML format

Documents in HTML format were briefly mentioned in Section 2.4, which dealt with documents in ePub format. HTML documents can be edited using a text or HTML editor in combination with a web browser.

[23] calibre-ebook.com/
[24] sigil-ebook.com/

2.7 Documents in XML format

For documents in this format, it is important know which type of XML was used. For the open textbooks in this inventory, the types TEI and Pressbooks were found. Although all files in XML format can be opened and edited in a text editor, it is also easier in this case to copy the relevant elements to a new document in a format that is easier to edit.

2.8 Recommendations

It is relatively difficult, or in the case of PDF documents almost impossible, to modify documents in the file formats in which most open textbooks are published. The concise technical analysis carried out for this project has not resulted in new insights concerning the adaptability of these documents. We therefore make the following recommendations:

- If possible, contact the original authors to obtain the source files and use these to edit the document.
- Use PDF editing software to modify content from a PDF and create a new document, making sure to comply with the conditions of the Creative Commons licence.
- For the reuse of visual content such as diagrams, images, photographs and tables, one solution is to make screenshots of these. Note, however, that the quality of a screenshot is usually poorer than that of the original.
- Authors of textbooks must consider how they expect their textbook to be reused. It is not enough to publish a textbook with an open licence that allows users to adapt and reuse the textbook if the format of the published file means that it is difficult to modify.

3 Conclusions and follow-up actions

The ‘Inventory of open textbooks’ project was carried out by the Reusing open educational resources working group of the Towards digital (open) educational resources zone. This is the first time that an attempt has been made in the Netherlands to produce a comprehensive overview of open textbooks and to determine the subject areas and disciplines in which open textbooks have been published. The results of the inventory show that many textbooks are published under open licences, and that this number is growing.

When making its inventory of open textbooks, the working group found that it was very dependent on the availability of metadata for the textbooks in order to be able to produce an exhaustive overview. Textbooks were only included in the inventory if they were available on platforms that allow open access to their metadata. The inventory could therefore be made more complete if access was obtained to the metadata of the unused sources named in Section 1.2.

The metadata that was available was not always provided in a standard format or was of a poor quality, making it impossible to achieve a clear classification of textbooks in terms of subject area, keywords and/or discipline. Using sophisticated tools, however, the metadata can be enriched, making it possible, for example, to analyse the key concepts in a textbook. Elasticsearch was found to be a promising tool for this.

A brief review in phase II of the adaptability of the inventoried open textbooks showed that this is poor for many textbooks. Although the licences that apply to the books usually permit the copying or adaptation of content and its application in another context, the format in which the content is published generally does not allow this. We therefore recommend that adequate attention is given to resolving this issue when developing services to support authors in the publication of open educational resources. In this way, lecturers will be able to copy and adapt future open educational resources and mix them with other content in order to create new content.

These recommendations form our advice for the Infrastructure & content working group of the Towards digital (open) educational resources zone. Our suggestions to further enrich metadata and develop services for the publication of open textbooks have meanwhile been incorporated into the two-year project proposal of this working group. The activities of the Reusing open educational resources working group are therefore complete.

Would you like to be kept informed of the follow-up to this project, or would you like to receive further information on the technical review conducted in this project? If so, please send an email to leermaterialen@versnellingsplan.nl, making sure to mention the inventory of open textbooks project.

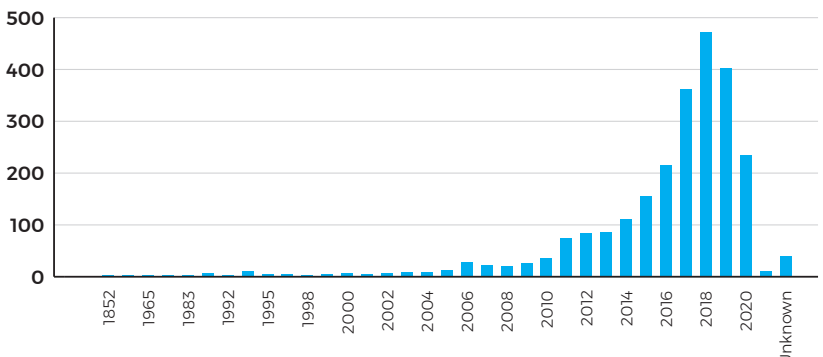
Appendix A

Results of inventory of open textbooks by discipline

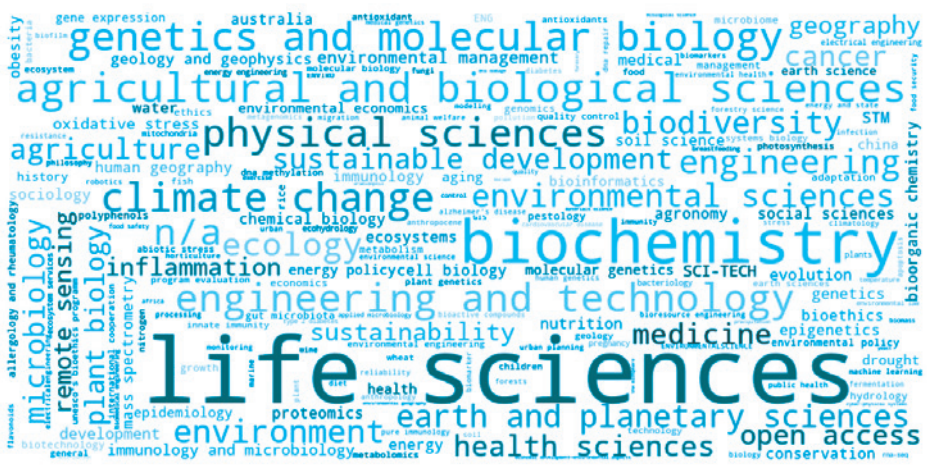
Earth and environmental sciences

Number of publications

2,400

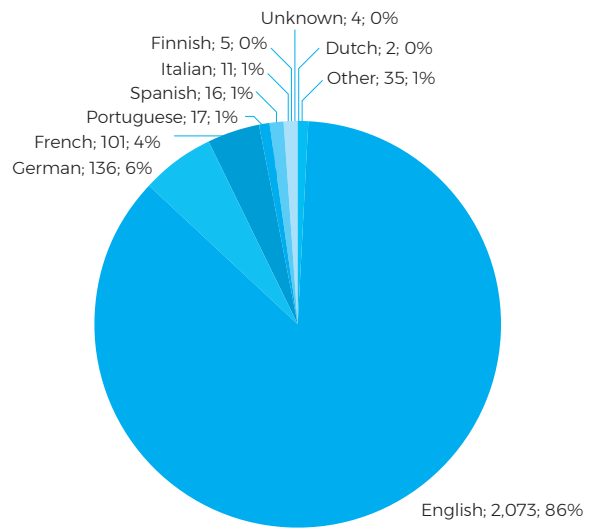


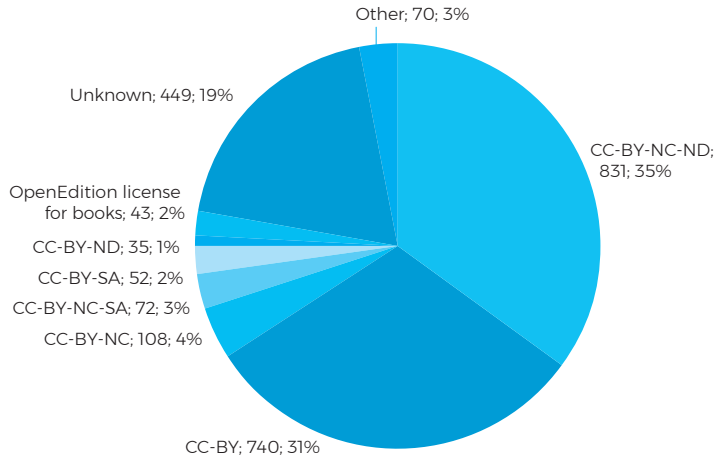
Number of books in year of publication



Subjects

Languages





Licence type

Economics and business

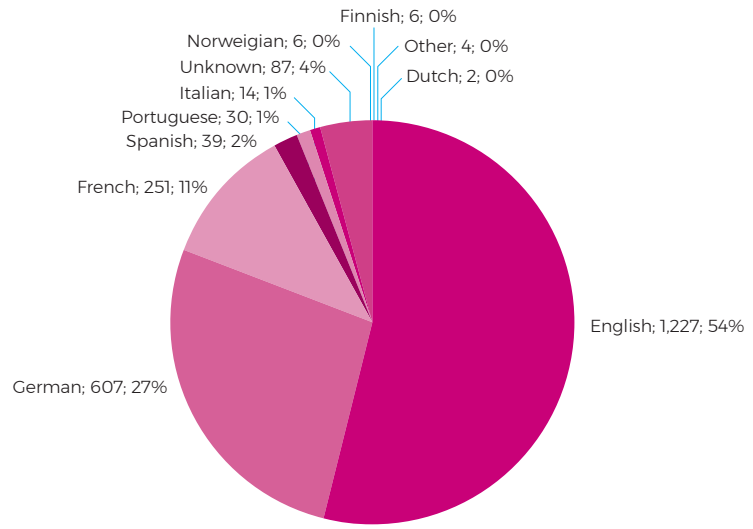
Number of publications

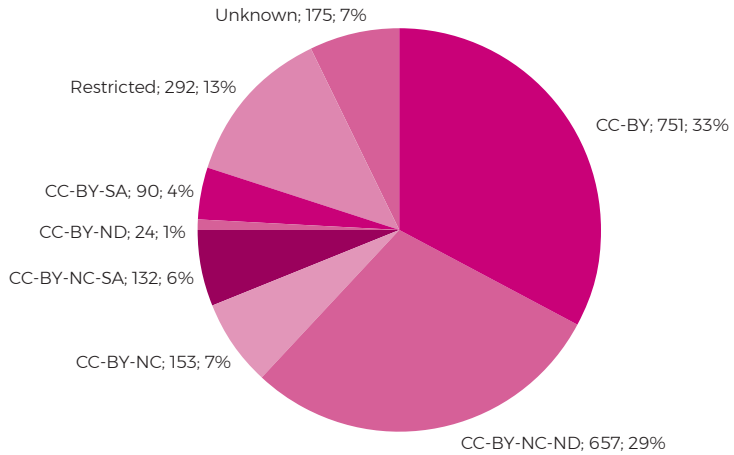
2,273

Number of books in year of publication

Subjects

Languages



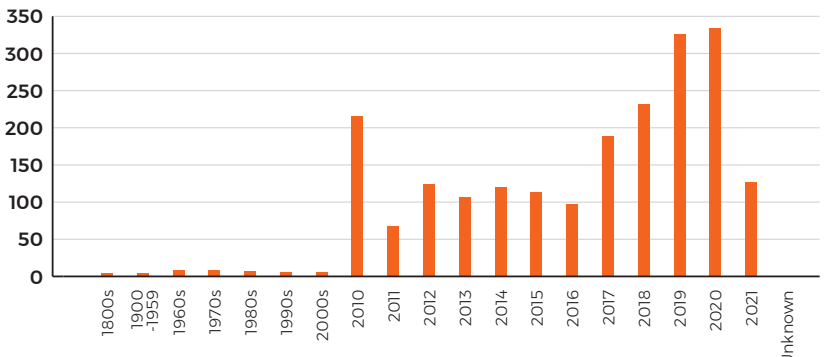


Licence type

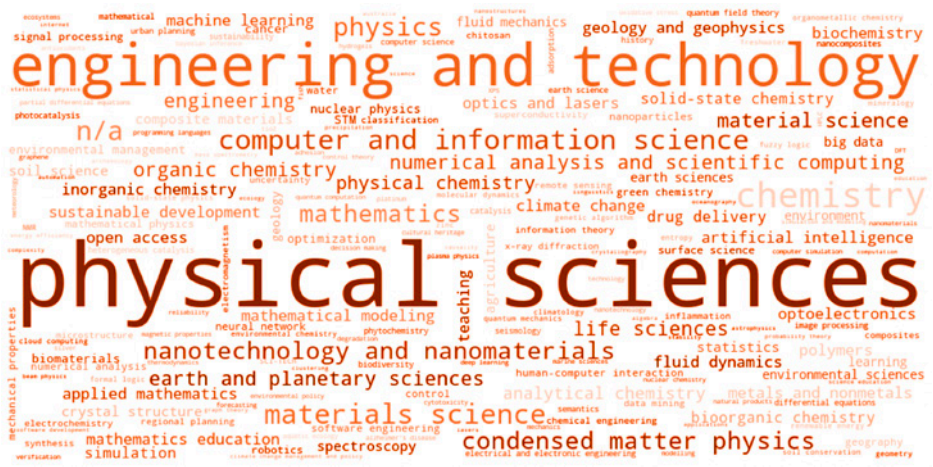
Physical and information sciences

Number of publications

2,094

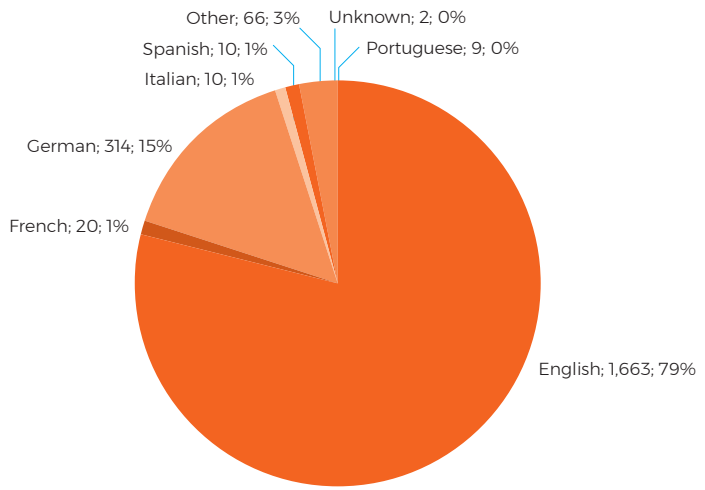


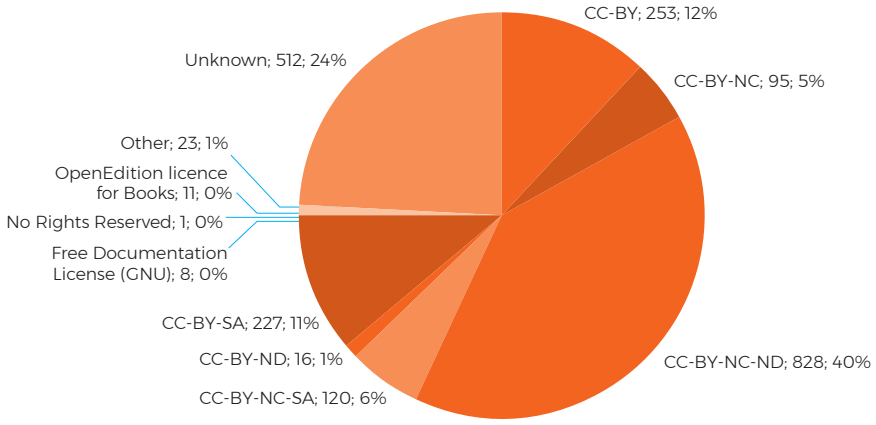
Number of books in year of publication



Subjects

Languages



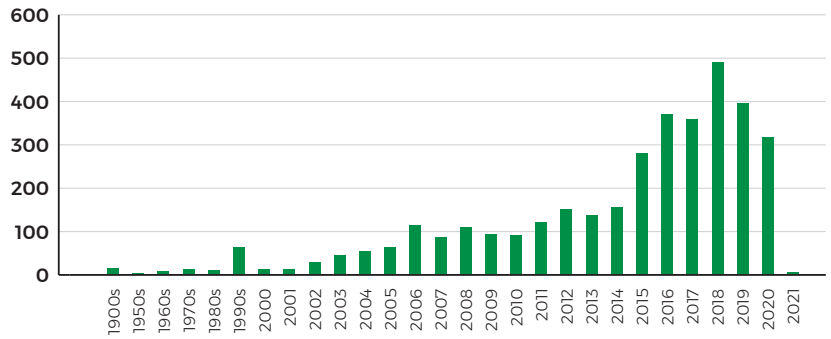


Licence type

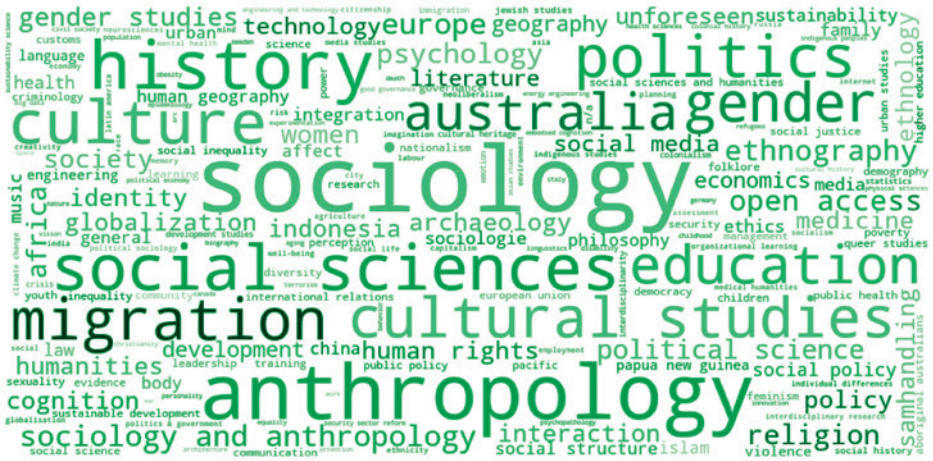
Behavioural and social sciences

Number of publications

4.742

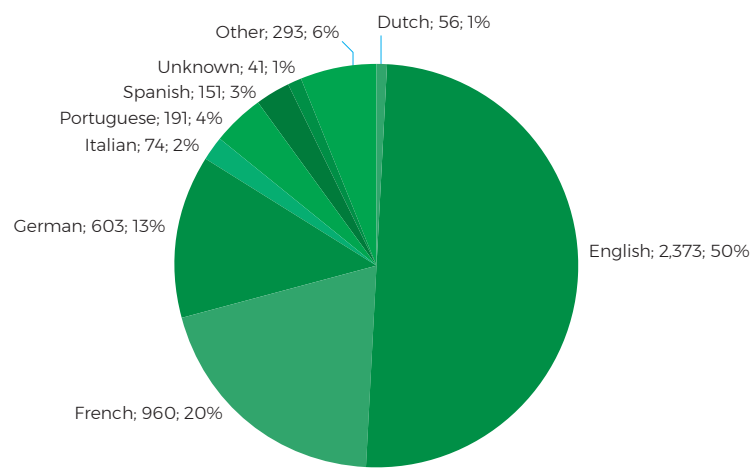


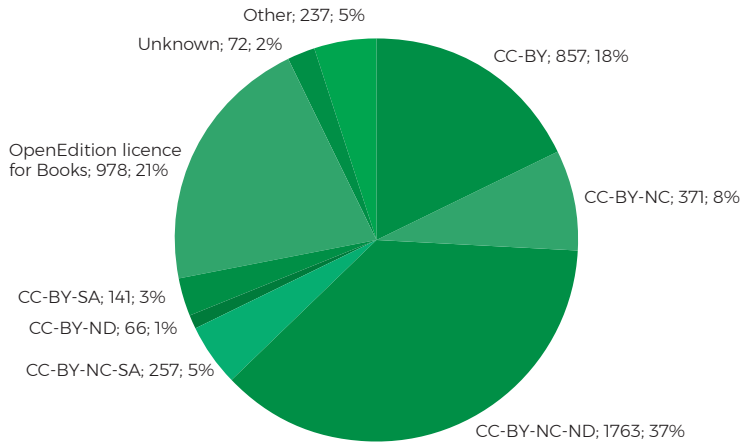
Number of books in year of publication



Subjects

Languages



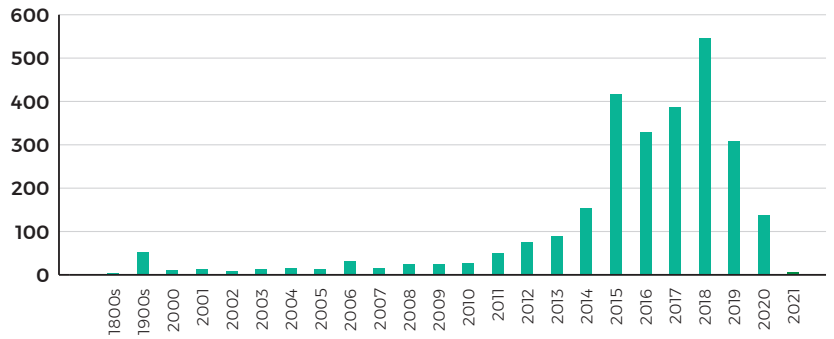


Licence type

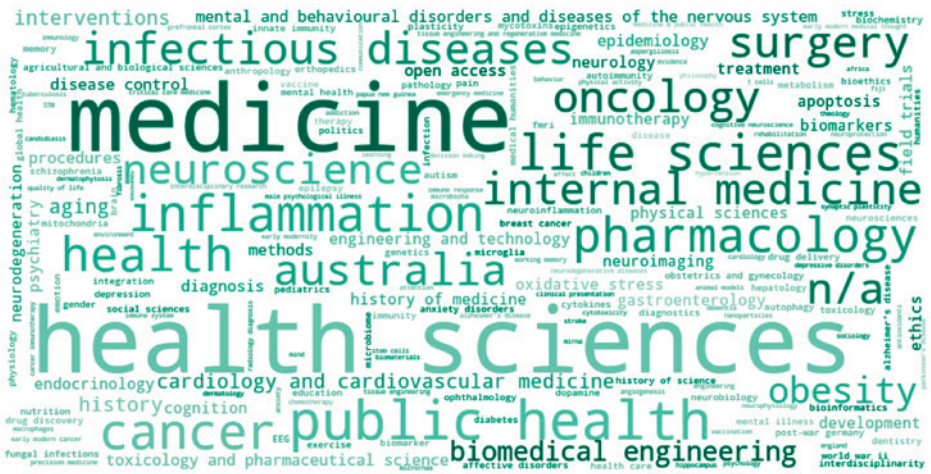
Health sciences

Number of publications

2,705

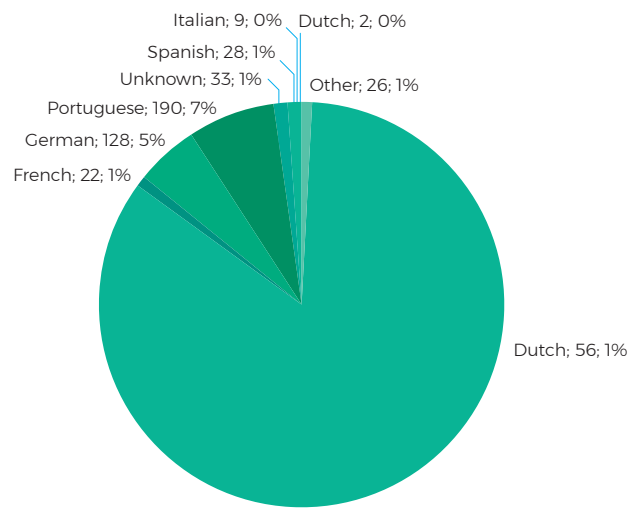


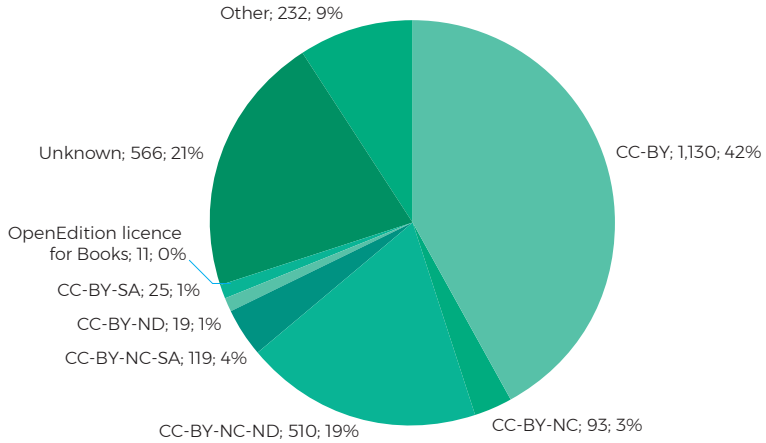
Number of books in year of publication



Subjects

Languages



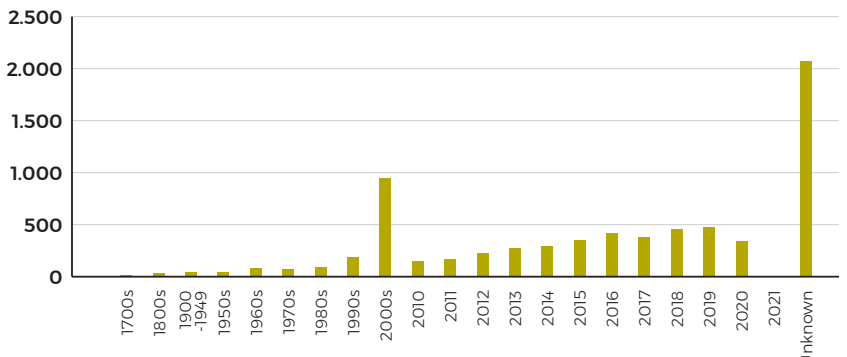


Licence type

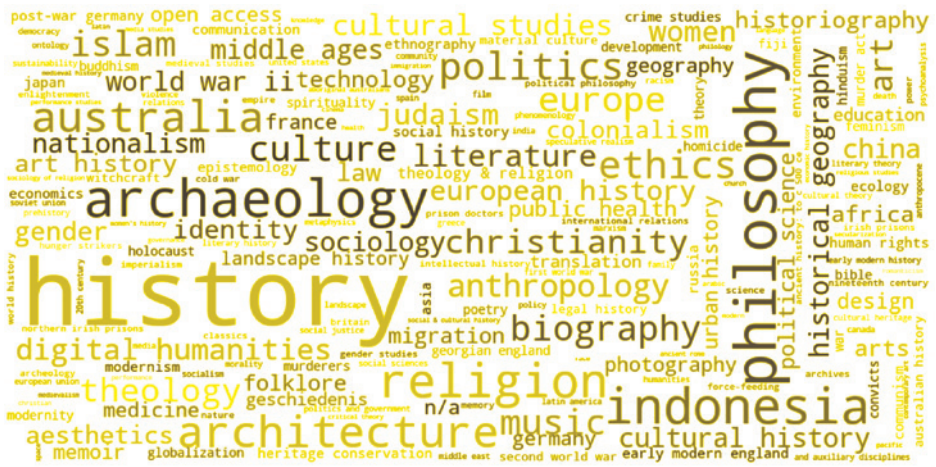
Art and cultural studies

Number of publications

6,912

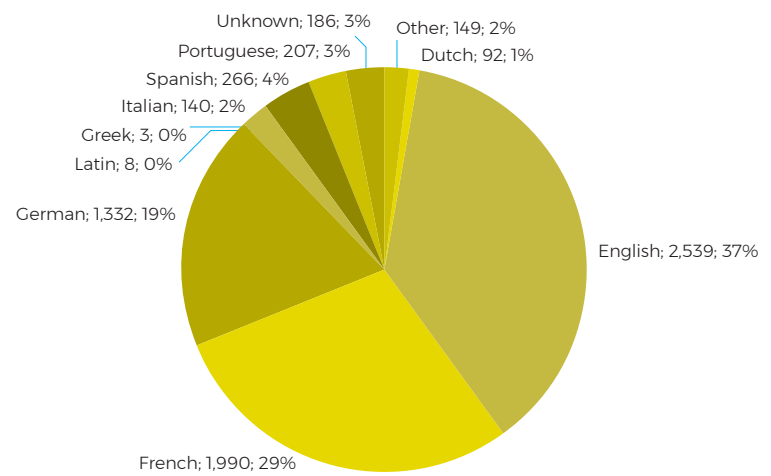


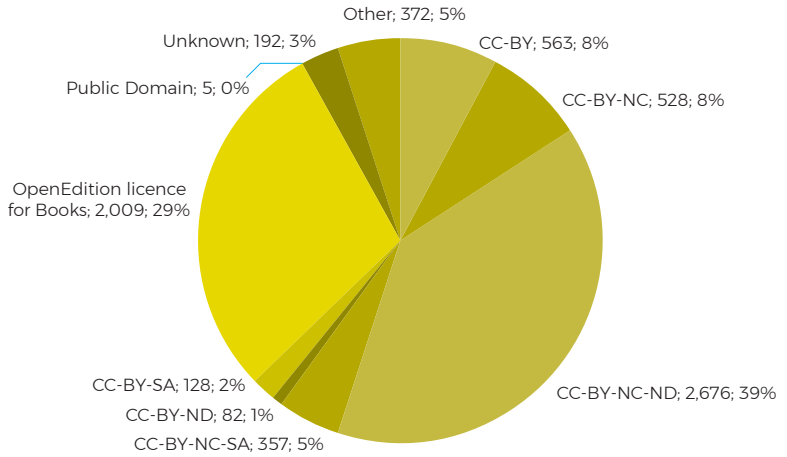
Number of books in year of publication



Subjects

Languages



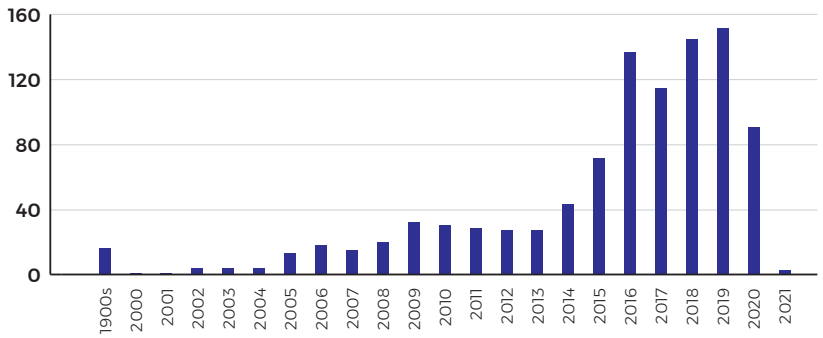


Licence type

Education

Number of publications

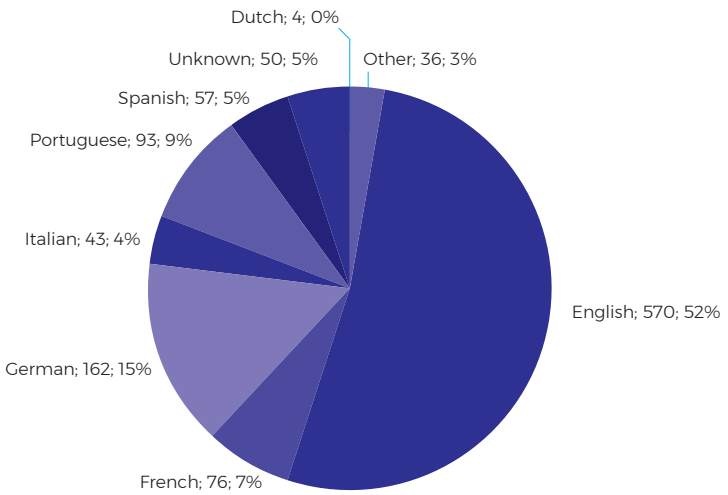
1,091

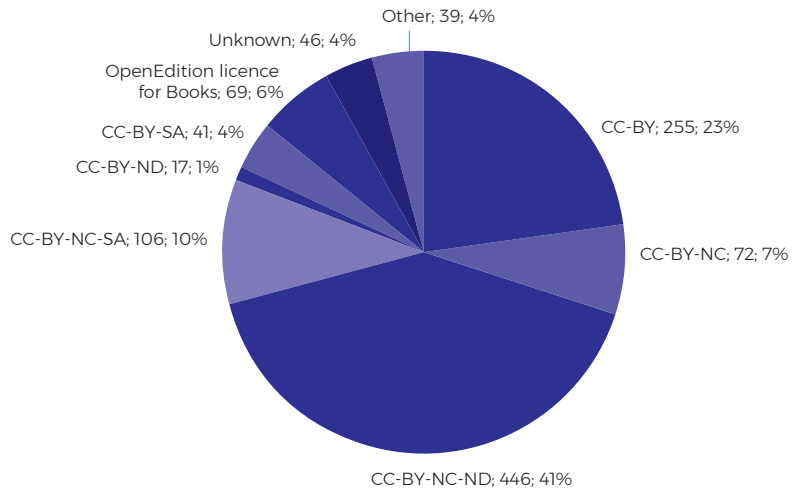


Number of books in year of publication

Subjects

Languages



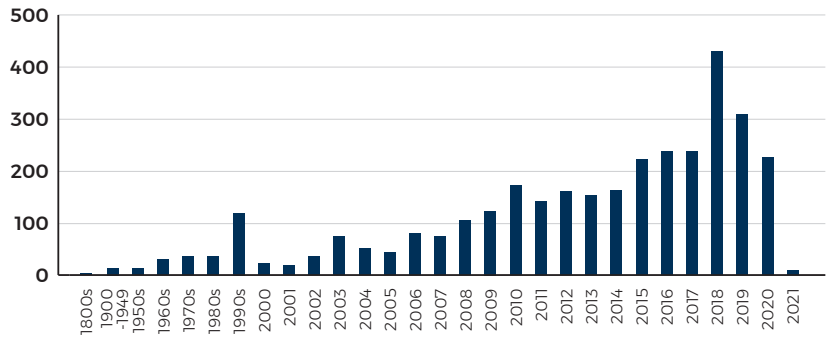


Licence type

Law and political science

Number of publications

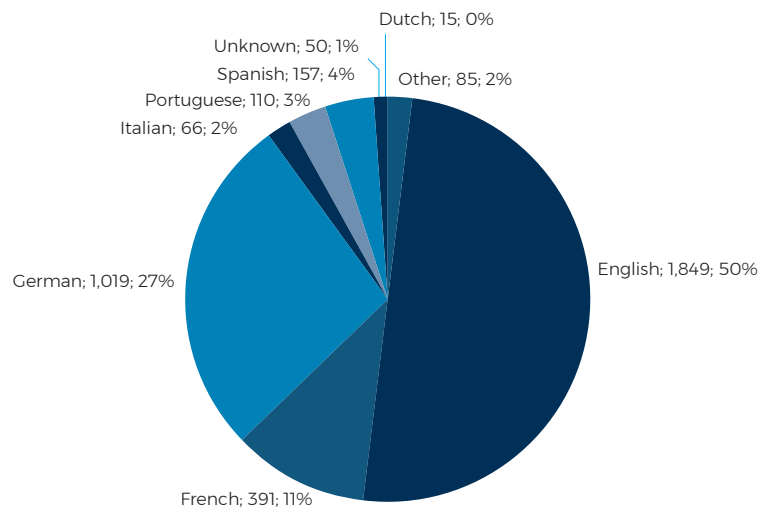
3,742

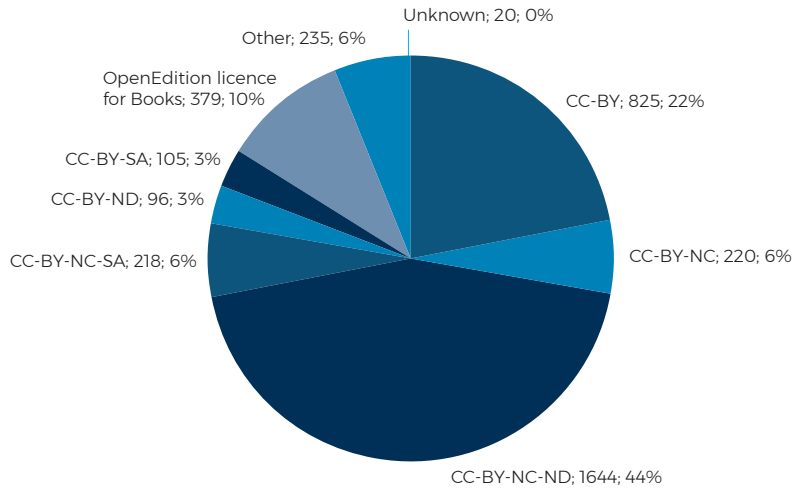


Number of books in year of publication

Subjects

Languages



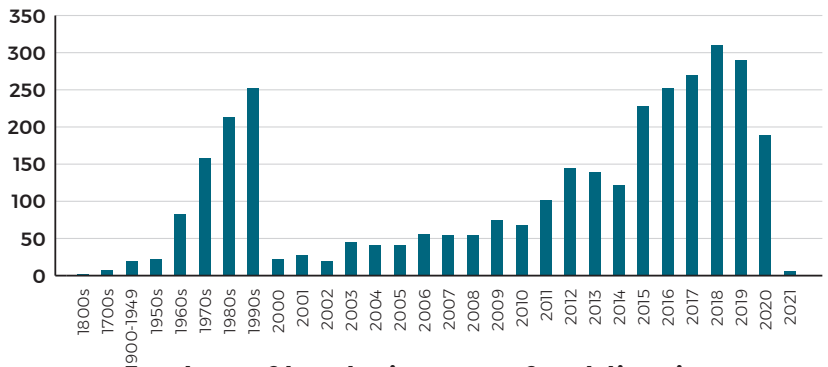


Licence type

Language and communication

Number of publications

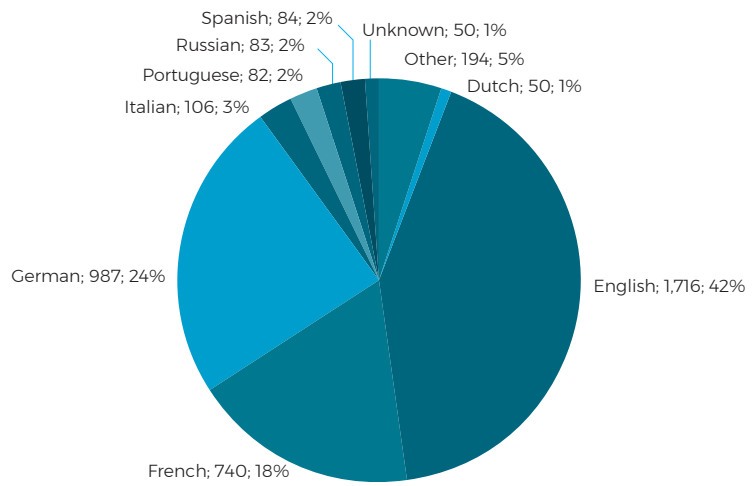
4,092

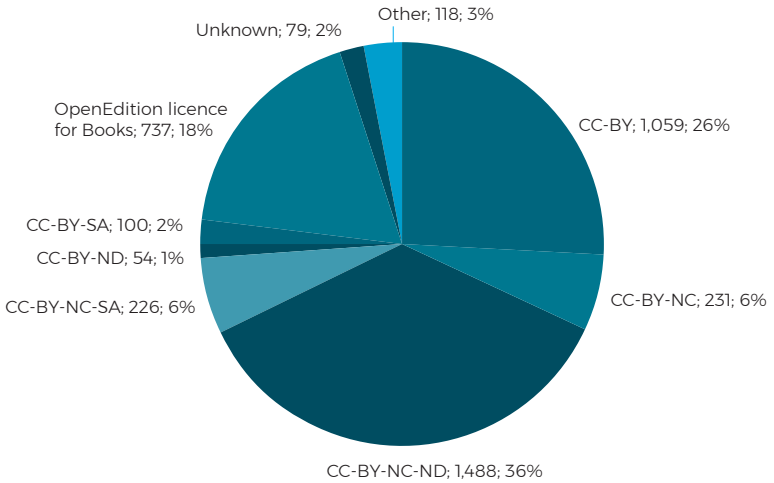


Number of books in year of publication

Subjects

Languages



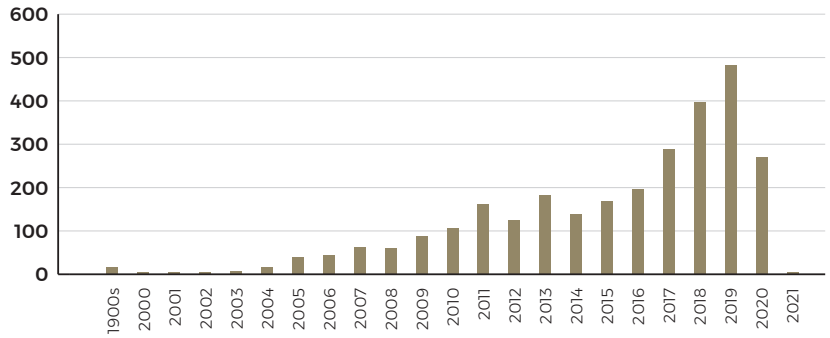


Licence type

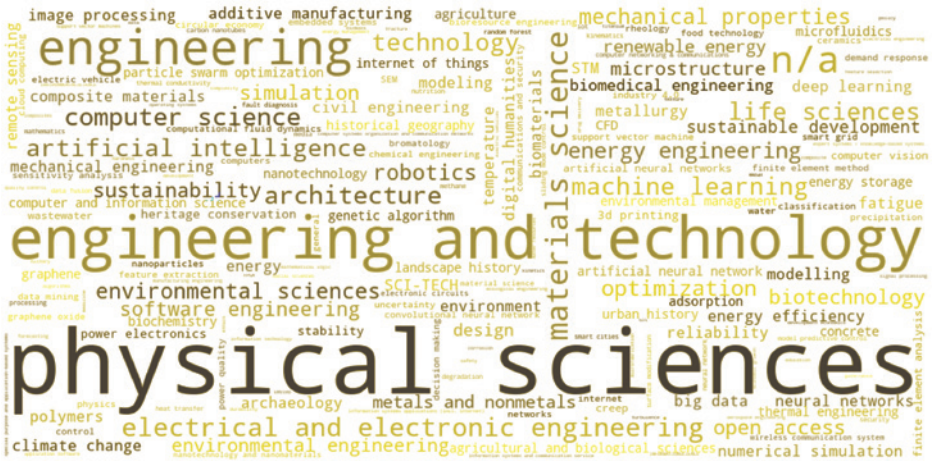
Engineering and technology

Aantal publicaties

2,917

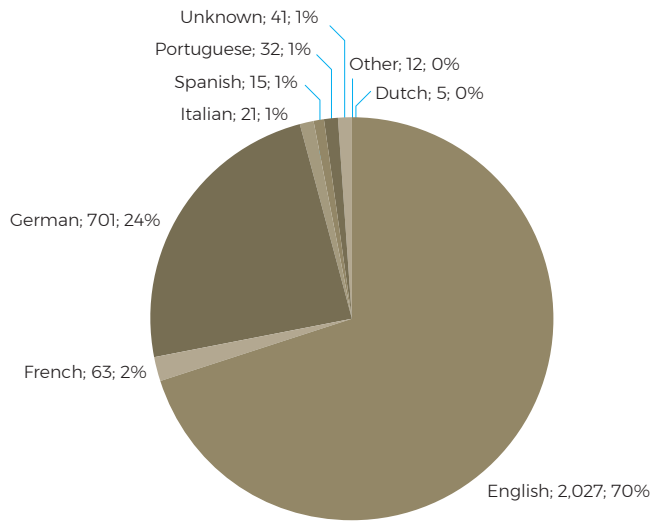


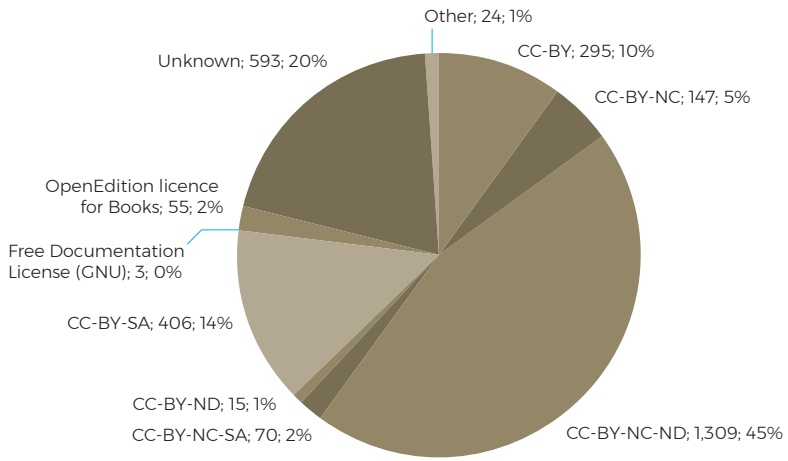
Number of books in year of publication



Subjects

Languages





Licence type

The Acceleration Plan for Educational Innovation with IT is a four-year programme from SURF, the Netherlands Association of Universities of Applied Sciences and the Association of Universities in the Netherlands, that aims to bring together initiatives, knowledge, and experiences to realise ambitions for higher education at an accelerated pace. This is taking place in eight 'acceleration zones'. In the Towards digital (open) educational resources acceleration zone, eight universities are working to ensure that students and lecturers have the opportunity to compile and use an optimal mix of educational resources.

For more information and our publications, please visit
www.versnellingsplan.nl