CWTS BIBLIOMETRIC REPORT

Meaningful metrics

Bibliometric study the EPFL (2009-2020/2021)

August 22, 2022

Universiteit Leiden

Bibliometric study
the EPFL (2009-2020/2021)

Report for the ETH Board

Nicolas Leclaire
Häldeliweg 15
8092 Zurich, Switzerland
E-mail nicolas.leclaire@ethrat.ch
Webpage https://www.ethrat.ch/en

CWTS

Ed Noyons, Clara Calero, Rodrigo Costas, Jeroen van Honk

CWTS B.V.
P.O. Box 905
2300 AX Leiden, The Netherlands
Tel +31 71 527 5806
E-mail info@cwts.nl
Webpage http://www.cwtsbv.nl/

General parameters of the bibliometric report

Parameters

Database	:	Web of Science (Articles, Reviews and Proceedings papers in the SCIE, SSCI, AHCI, and CPCI)
Version	:	2152 (CWTS)
Classification system	:	Publication-level classification system (about 4000 fields, referred to as research areas)
Publication window	:	2009–2020
Citation window	:	Maximum 4 years (and until 2021)
Counting Method	:	Fractional counting at the level of organisation for citation impact measurement
Self-citations	:	Excluded
Top indicators	:	Top 10%

Contents

List of indicators

Avg Reads Average number of reads per DOI. A *read* is defined by saving a publication in a Mendeley user account.

IntCov Internal coverage. Estimated Web of Science coverage of a set of publications. A description of the calculation is provided in Annex C.1.

IntDisc Measure of *interdisciplinary* research, defined by the proportion of references in a publication assigned to other fields. Fields are defined by journal categories. In addition, the cognitive distance of fields to each other is also considered (more info at Section 2.2 (p. 16) and Annex D).

MCS Mean citation score. The average number of citations received by a publication (TCS/P[full]).

MNCS The mean normalised citation score. This represents average citation score per publication, normalised by research area and publication year. Research areas are defined by a detailed publication classification system of CWTS, consisting of about 4000 areas. The average MNCS in the entire database is 1. Scores higher than 1 reflect a citation-based impact that is higher than the world average.

MNJS The mean normalised journal score. This represents the normalised average citation impact of journals. The MNJS is an average score for all publications in the same journals in which an institution published. The normalisation is based on the same principles as the MNCS. The average MNJS in the entire database is 1. Scores higher than 1 reflect a journal citation impact that is higher than the world average.

P[full] The number of publications, full counting. Each publication is counted in full (i.e. as 1).

P[fract] The number of publications, fractionally counted. The fraction is determined based on the number of co-authoring organisations.

P[OA] Number of publications, full counting, in Open Access(OA). In addition, we provide the number for the different kinds of OA: Gold, Hybrid, and Green. A publication is tagged by one type only. Gold and Hybrid overrule Green. Information is based on Unpaywall data (July 2021).

PP[OA] The proportion of publications in Gold, Hybrid or Green OA, while publications without a DOI are discarded (OA unknown).

PP[collab] Proportion of publications, full counting, involving collaboration (at least two institutions co-authoring).

PP[int collab] Proportion of publications, full counting, involving international collaboration (co-authorship of organisations from more than one country).

PP[industry] Proportion of publications, full counting, involving industry (co-authorship with companies).

PP[uncited] Proportion of publications, full counting, that are not cited.

PP[self cits] The average number of author-self citations per publication. A self-citation is defined as any of the authors of a cited publication is the same as any of the authors of the citing publication.

P[top10%] The number of publications, counted in full belonging to the top 10% of their research area. The area is determined on the basis of a detailed publication classification system of CWTS, consisting of about 4000 areas (See Annex B).

PP[top10%] The proportion of publications (P[fract]) belonging to the top 10% most cited of their area and in the same year. The areas are determined using a detailed publication-level classification system , consisting of about 4000 areas. The PP[top10%] in the entire database is 10%. A score above 10% represents impact that is higher than the world average.

PA[F inst] Share of female authors of an institution within a publication.

PA[F pubs] Share of female authors within a publication (institution plus co-authors).

A[M inst] Number of male authors of an institution.

A[FM inst] Number of authors of an institution for which we could define gender male or female.

RPA[F] Proportion of female authors compared to the total of authors for which gender (male or female) was defined (more info at Section 2.2).

TCS The total citation score. This represents the total number of citations accumulated within the citation window, excluding author self-citations.

For more details about the normalised citation indicators, please refer to Waltman et al. (2012). More information about the mentioned publication-level classification is in Annex B.

Definitions, abbreviations and acronyms

CWTS Centre for Science and Technology Studies, Leiden University

A&HCI Arts & Humanities Science Citation Index

SCIE Science Citation Index Expanded

SSCI Social Science Citation Index

CPCI Conference Proceedings Citation Index

DOI Digital Object Identifier (a permanent ID for publications)

JSC Journal Subject Category

OA Open Access

Research area A set of publications on a certain topic, identified by the Leiden Algorithm (Annex B)

Subject A set of publications in journals belonging to a (subject) category

WoS Web of Science

1 Introduction

The ETH Domain consists of two Federal Institutes of Technology, ETH Zurich and EPFL, and four research institutes PSI, WSL, Empa and Eawag. Together, they play a vital role in the Swiss science system for education, research and transfer of knowledge and technology.

The ETH Board commissions an intermediate evaluation every four years. The most recent one took place in 2019. The bibliometric study was executed in 2018. The evaluation is a moment for the Swiss Federal Council, the ETH Board, as well as staff and management of ETH Domain to find out where ETH Domain stands vis-a-vis the ambitions and measures formulated in the strategic planning document. Moreover, the intermediate evaluation should lead to recommendations relating to these ambitions and measures.

Bibliometric studies can provide evidence related to ambitions and measures as part of a self-assessment report. Although we consider that meeting the standards of objectivity for determining the impact of scientific research is important, we believe that decision-making towards the goal of evaluating the quality of institute's research ought to be multi-dimensional rather than overwhelmingly quantitative. Bibliometric measures provide objective evidence about production, collaboration and impact but only for the research that has been published in (international) journals and proceedings. Therefore, we strongly recommend that quantitative evaluations are complemented with qualitative information (for example the mission and the research goals of a department) and expert assessments.

This report includes the bibliometric analysis of the scientific output of EPFL, covering the period 2009-2020, including citations up to 2021. The studies are based on a quantitative analysis of scientific publications in journals and proceedings processed for the Web of Science (WoS) versions of the Science Citation Index and associated citation indices: the Science Citation Index (SCI), the Social Science Citation Index (SSCI), the Arts & Humanities Citation Index (A&HCI) and the Conference Proceedings Citation Index (CPCI).

Although most of the methodology is similar to the study performed four years ago for EPFL, the results may sometimes differ substantially, due to the fact that in the current report conference proceedings papers are included and fully integrated, but that depends on the role conferences play for an institution if this is actually the case. Moreover, new indicators were introduced: RPA[F], IntDisc, P[OA], PP[OA], and Avg Reads.

We introduce each result in brief, while more detailed information about data and method is provided in Section 2 and Annex C) of this report.

In Section 3 the results of our analysis and interpretations are reported. These results are discussed in 5 parts:

1. Section 3.1: Overall output and impact
2. Section 3.2: Research focus in context
3. Section 3.3: Collaboration and partners
4. Section 3.4: Research accessibility
5. Section 3.5: Impact and knowledge use.

In the annexes, we provide more detailed scores for some indicators, more detailed information about specific approaches, as well as information about CWTS infrastructural elements involved in the analyses.

2 Data collection and methodology

2.1 Data collection

EPFL provided CWTS with a list of publications from its own repository infoscience. CWTS used these data to match the publication records with the records in its database (matched results). Simultaneously, CWTS collected EPFL's publication data from its database using the author affiliations in publications. Both data sets were compared to each other.

After EPFL and CWTS compared, checked and corrected these two sets, the final dataset was prepared for the bibliometric analysis.

Additionally, for the Mendeley readership analysis EPFL provided CWTS with any DOI registered in its repository.

2.2 Summary of method

In this section, we discuss the methods underlying the bibliometric analysis developed. We discuss the basic principles of our indicators and the context in which they can (or should not) be used. Additional and more detailed information about methods and data can be found in the annexes.

2.2.1 Indicators

In bibliometric analyses regarding research performance, we usually discern two types of indicators: size-dependent and size-independent, taking into account the different size of institutions under investigation. Larger institutions, for instance, will be involved in more publications than smaller ones. Subsequently, this will affect the absolute number of top 10% publications, as well as all other size-dependent indicators. In Figure 1 we visualise the correlation between the two indicators for the 6 ETH institutions.

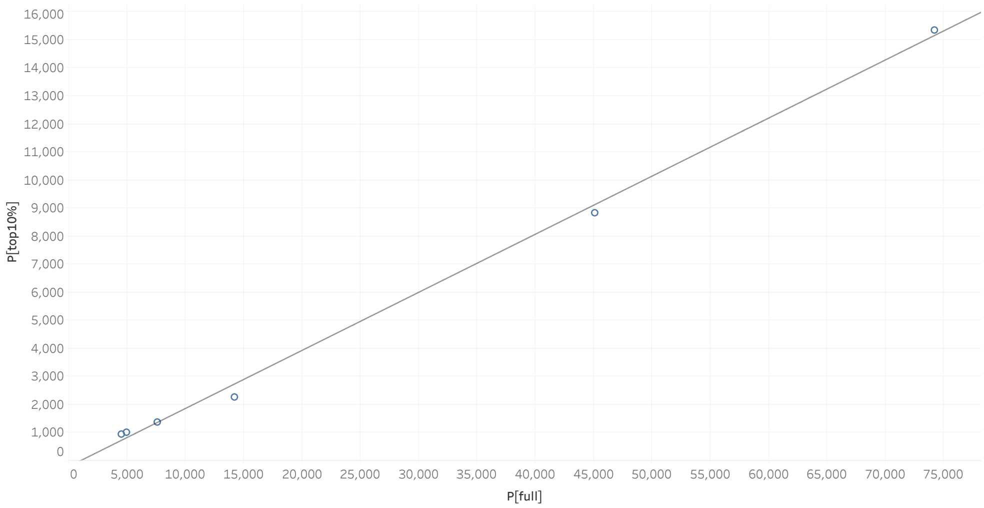

Figure 1: P[full]vs.P[top10%]for 6 ETH institutions

Proportion indicators (e.g., PP[collab], PP[int collab], PP[industry], PP[OA], PP[top10%]) and average indicators (MNCS, MNJS) are size-independent, while others used in this study (e.g., P[full], P[fract], TCS) are size-dependent. In the report we will primarily discuss the results using the size-independent indicators to account for the size differences of the organisations. Moreover, the results for size-independent indicators can, in most cases, be related to the world average.

Output indicators

Size-dependent

The total number of publications in which researchers from an institution were involved (**P[full]**) is the basic output measure. In addition, we provide the indicator **P[fract]** which assesses an institution's contribution to the output P[full]. Each individual publication is divided by the number of organisations co-authoring, regardless of the number of organisations involved. If authors have two affiliations and mention both, both affiliations are counted as fractions. P[fract] is the sum of these fractions of publications in which an institution was involved.

Size-independent

Proportion indicators characterise sets of publications regardless of the number and are therefore size-independent. They are often used to characterise output. For instance, **PP[collab]** indicates the proportion of output with at least two different organisations involved. **PP[int collab]** indicates the proportion of output involving international collaboration. In this report, a publication is tagged as an international collaboration if at least one of the co-authoring organisations is based outside of Switzerland. Furthermore, two other proportion indicators are used: **PP[industry]**, representing the proportion of P[full] co-authored with a company and **PP[OA]**, the proportion of P[full] published in Open Access (OA).

For OA publications, we discern different types: OA Gold, OA Hybrid and OA Green. The definition of the types used in this report are:

- Gold: The publisher makes all articles and related content available for free immediately on the journal's website.
- Hybrid: Publication freely available under an open license in a paid-access journal.
- Green: Published in toll-access journals, self-archived by authors (in repositories or researchers' websites), independently from publication by a publisher.

OA publications are counted only as one type at the same time. If a paper is both Green and Gold, it is counted as Gold. Bronze OA publications are free to read only on the publisher page without a license. As such, they were disregarded as OA. These were identified as *Closed Access* publications.

Impact indicators

Size-dependent

The scientific impact of an institution's output is measured by citations. We provide the total number of citations received (**TCS**) in the period of maximum 4 years after publication, up to 2021. For more recent years the citation window is shorter than 4 years. We exclude author self-citations. Another size-dependent indicator of impact is P[top10%], i.e. the absolute number of publications belonging to the top 10% most cited publications (in their area and from the same year).

It should be noted that all citation-based indicators (including TCS) are calculated using a limited and fixed time-window. The total amount of citations for early publications may therefore be higher than processed for this report.

Size-independent

The **MNCS** is the indicator to measure citation impact after normalising by research area and publication year. The research area to which a publication belongs is defined by a publication-level classification (for details, see Annex B). In this classification each publication is uniquely assigned to a research area. Areas are defined by their citation environment (cited and citing publications). This classification is more fine-grained and is considered more accurate than a journal classification (Ruiz-Castillo and Waltman, 2015). In a journal classification all publications from one journal are in the same class. Similar journals are in the same class and journals may belong to more than one class. We use this journal classification to characterise an institution's output in its research profiles but not to normalise impact. The journal classification is less fine-grained and as such easier to relate to the main subjects addressed.

In addition, we provide the proportion of publications in the top 10% most cited publications (within their research area, i.e. class, and in the same year, **PP[top10%]**).

This indicator correlates strongly with the MNCS but is not sensitive to outliers. The MNCS can sometimes be biased by one paper being cited many times. The PP[top10%] is not influenced by this one paper, as it is 'just' one of the top 10% or not. An MNCS that is relatively much higher than the PP[top10%] points to a highly skewed distribution of impact across publications. In other words, a few publications receive a huge number of citations, compared to the other publications.

Finally, we also use an indicator measuring the impact of journals, the Mean Normalised Journal Score (**MNJS**). This indicator assesses the impact in term of citations of the journals (aggregated), in which the institution has published, using the same normalisation as we use for measuring the impact (MNCS). As such, the MNJS does not measure the (average) impact of an institution's publications, but rather the impact of the journals in which its researchers publish.

2.2.2 Additional indicators

In this study we introduce indicators that relate to the context of the published research. We will discuss them in brief in the next subsections.

Worldwide growth of research fields

An indicator to position an institution's research activities in the context of what happens at a larger scale is the [**Field growth**]. We use the science landscape (see Annex B) to reflect what happens worldwide, by calculating a growth indicator for each area (the [**Area Growth**]).

The [Field growth] relates the output of an institution to these area growth values ([Area Growth]) as follows. First, we calculate for each of the 4000 research areas in the science landscape, the share output of the most recent two years (2019-2020) as compared to the total in 2009-2020 (the period under study). This share of output in the most recent years is normalised by a reference value, which is the result of the number of recent years (2) and the number of years of the total period considered (12): 0.17. Areas in which the share of output in the recent years is higher than 0.17, have a [Area Growth] above 1, a positive growth.

Any value above 1 means a positive growth, while values below 1 indicate a negative growth. In Figure 2, we plotted the [Area Growth] in the landscape of all science, by color-coding. Green areas show a positive growth (>1) in the most recent two years, while red areas show a negative growth (<1). The size of a circle proportionally reflects the number of ETH Domain publications published in 2009-2020 worldwide, ranging from 1 up to 1,400.

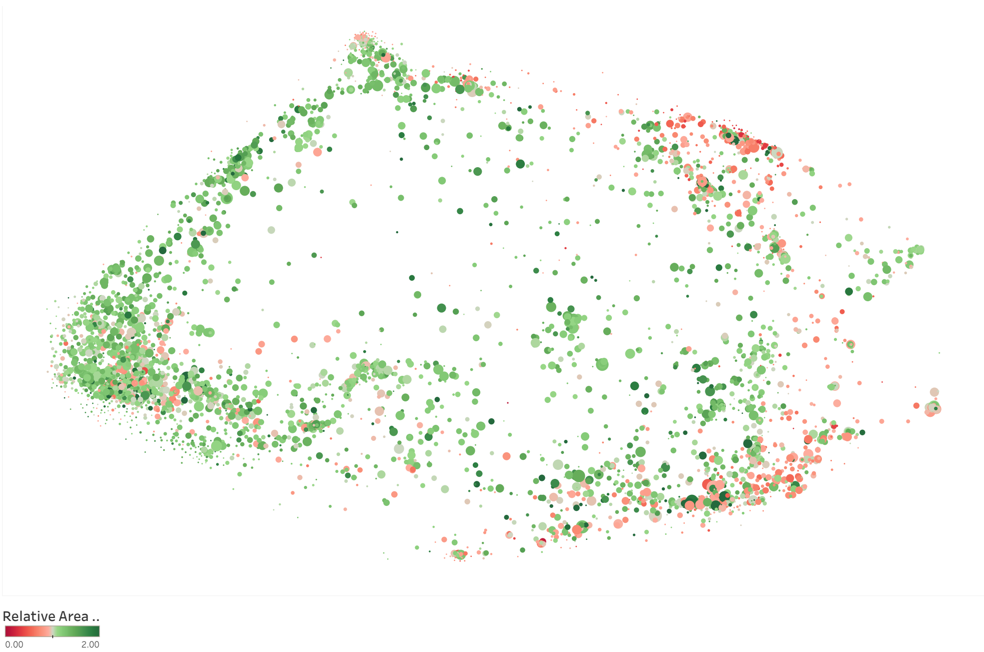

Figure 2: Landscape of all science, color-coded by [Area Growth]

[Field growth]

We use the [Area Growth] to characterise the fields in which EPFL researchers are active. Thus we contribute to the answer to the question: is EPFL's research positioned in fields with an increasing interest worldwide or not?

The [Field growth] is the average of [Area Growth] values of the areas in which an institution's publications can be found. Consider the output of an institution X, with 100 publications. These 100 publications may be in 20 different areas. Depending on the [Area Growth] values of these areas, these 100 publications relate to 20 different [Area Growth] scores. The average [Area Growth] values of the 100 publications, then indicates the estimated growth of fields in which X is active: the [Field growth] of institution X.

Interdisciplinary research

We introduce a measure related to the interdisciplinary character of the published research. Being more or less interdisciplinary is defined by the knowledge base (the prior art that is being cited) of the published research. The content of cited publications is defined by the journal subject categories.

If a publication cites research from one (and most likely its own) subject category only, it is defined as mono-disciplinary (measure close to 0). If a publication cites research from different subjects, we consider it as interdisciplinary. If the subjects are cognitively at a long distance from each other, the measure of interdisciplinarity is even higher, with a maximum of 1.

The cognitive distance between subject categories is determined by the density of the citation traffic between them. If a publication (A) cites output in subject X and Y, while X and Y are remote from each other (little citation traffic between them), it is considered more interdisciplinary than publication B, which cites publications from Y and Z, which are cognitively closely related (i.e., in subject categories frequently citing each other).

For each publication we calculate an interdisciplinary value and for sets of publications we then calculate their average (**IntDisc**), which is a value between 0 and 1, where 0 indicates mono-disciplinary and 1 means maximum interdisciplinarity.

In summary, interdisciplinarity is:

1. Defined by cited references in a publication;
2. On the basis of the variety of journal categories of cited publications;
3. Considering cognitive distance between these categories;
4. While this distance between categories is based on mutual citation traffic.

The above leads to the definition of interdisciplinarity we use in this report:

> The interdisciplinarity indicator (IntDisc) relates to the diversity of research supporting the current research.

In order to be able to interpret the IntDisc measure in a broader context, we calculated a reference value (**Ref Intdisc**), which is the IntDisc for the journal category at large in 2020. In this way interdisciplinarity can be assessed within each journal subject category by relating it to the world average. We integrated both scores (IntDisc and Ref Intdisc) in profiles, where interdisciplinarity is included. More info can be found in Annex D.

Share of female authors

We also introduce an indicator related to gender diversity of research staff. We calculated the probability of an author name to be male or female, by looking at the first name. If first names (or nicknames) point to a gender within a specific country, the gender is set using the following four-step procedure (also described at CWTS Leiden Ranking):

1. Author disambiguation. Using an author disambiguation algorithm developed by CWTS (Caron and van Eck, 2014), authorships are linked to authors. If there is sufficient evidence to assume that different publications have been authored by the same individual, the algorithm links the corresponding authorships to the same author.
2. Author-country linking. Each author is linked to one or more countries. If the country of the author's first publication is the same as the country occurring most often in the author's publications, the author is linked to this country. Otherwise, the author is linked to all countries occurring in his or her publications.
3. Retrieval of gender statistics. For each author, gender statistics are collected from three sources: Gender API, Genderize.io , and Gender Guesser. Gender statistics are obtained based on the first name of an author and the countries to which the author is linked.
4. Gender assignment. For each author, a gender (male or female) is assigned if Gender API is able to determine the gender with a reported accuracy of at least 90%. If Gender API does not recognize the first name of an author, Gender Guesser and Genderize.io are used. If none of these sources are able to determine the gender of an author with sufficient accuracy, the gender is considered unknown. For authors from Russia and a number of other countries, the last name is also used to determine the gender of the author. Using the above procedure, the gender can be determined for about 70% of all authorships of major universities. For the remaining authorships, the gender is unknown.

For each publication, we counted the *number* of female authors at the level of the institution (A[F **inst**]) as well as at the level of the entire publication (A[F **pubs**]). In addition we counted those for male authors. We disregarded authors for which the gender cannot be defined or is ambiguous. The total amount of authors which we defined female or male is indicated by A[FM **inst**] and A[FM **pubs**].

Hence, for each publication in which EPFL authors were involved, there is a share of female EPFL authors (PA[F **inst**]), and a share of female authors for the publication at large (PA[F **pubs**]). The latter is used as a benchmark for the former. RPA[F] indicates the EPFL share, normalised by the share of the benchmark. A value higher than 1 for an institution X, indicates a higher proportion of female authors at X than for its community at large (X plus co-authoring partners).

2.2.3 Profiles

In the report we use two types of profiles:

1. A research profile in which we look at performance of an institution on the level of journal categories; and

2. A collaboration profile in which we look at performance of an institute of three collaboration types of publications.

In a research profile, we breakdown the EPFL output into Journal Subject Categories (JSC) to add content to the general statistics. It gives a general impression of all the broad subjects in which EPFL is involved. We include categories that cover at least 1% of the total output (P[full]).

For collaboration profiles, we classify all publications by their author affiliation information. The different types of collaboration are: (1) Single institution, in which only the institution under study is involved, (2) National collaboration for publications with co-authors from at least two different institutions from the same country, and (3) International collaboration for publications co-authored by institutions from at least two countries.

Output

By breaking down the output over categories, we provide a broad overview of activities and focus, by subject. In each profile we include both P[full] and P[fract], i.e. the number of publications in which an institution was involved (P[full]) and the number of publications normalised by the number of institutions involved as co-author (P[fract]). Moreover, if a publication is in a journal that belongs to two categories, it is assigned 0.5 to each category. In addition, we include an estimated growth factor for each subject [Field growth]. This growth factor is calculated on the basis of developments of research areas (see Section 2.2.2). A [Field growth] above 1 means a growth of output worldwide in the most recent two years.

By breaking down an institution's output over collaboration types, we provide insight into the publication strategy, as well as the integration of an institution into the national or international research community. Large shares of international collaboration output (P[full] and P[fract]) point to a strong international network.

Impact

In both types of profiles, the impact of individual publications is measured in the same way as for the entire institution (PP[top10%], MNCS and MNJS) and broken down over subjects and collaboration types. In the research profile, we rank subject categories on the basis of P[full] (using full counting). In this way we depict an institution's main focus by the number of publications in which its researchers are involved, while the impact is measured by the proportion to which it contributes, hence consistent with the overall impact measurement.

Research profiles in other contexts

We also used the breakdown over subject categories to provide more detailed information on the context in which research is executed and published. The main indicators we provide by subject are:

- RPA[F]: the share of Female authors relative to a benchmark
- P[OA], PP[OA]: the number and share of publications in OA
- IntDisc: the measure to which research is interdisciplinary
- PP[collab]: the proportion of output involving collaboration
- PP[int collab]: the proportion of output involving international collaboration

3 Results

3.1 Overall output and impact

Main findings

> The overall output of EPFL amounts up to 45,073 publications in which its researchers were involved, with the overall number of publications increasing over time. EPFL exhibits an overall high citation impact, with field-normalised impact substantially above the international reference values (MNCS values always above 1.56 and PP[top10%] above 18%). EPFL's publications are predominantly performed in collaboration (76%), with a predominant role of international collaboration (65%), and about 10% involving collaboration with industry. The scientific production of EPFL is mostly published Open Access (64%), showing an increasing pattern over time towards more openness. EPFL contributes substantially to research areas of all the 5 main disciplines of the science landscape, with some focus on topics related to Physical Sciences & Engineering.

3.1.1 Overall performance

In Table 1 the overall bibliometric statistics for EPFL are presented. Overall EPFL has produced a total of 45,073 publications, with 35,474 journal papers and 9,599 proceeding papers. The overall internal coverage (IntCov) is 0.78, literally meaning that about 78% of EPFL cited references are themselves also covered in the Web of Science database. It points out that we estimate that 78% of the scientific output is covered by WoS. This implies more that the topics researched by EPFL are well covered by the database chosen for this bibliometric study. We consider more than three quarters of the output representative for the total.

EPFL publications have received a total of 595,204 citations (excluding self-citations – which roughly represent 26% of all citations). The vast majority of citations are concentrated around journal papers, with a mean citation impact (MCS) of 16.15. The mean overall citation impact of the proceeding papers is much lower (MCS=2.32) which can be explained by the shorter nature of proceeding papers, making them less prone to receive citations, which is also supported by the rather high percentage of uncited proceeding papers (PP[uncited]=53%).

When it comes to field-normalised citation impact, the MNCS value of EPFL is very high with a value of 1.63, meaning that EPFL field-normalised impact is 63% higher than it would be expected by its international expected baseline. Proceeding papers have a particularly high normalised impact (MNCS=1.76), indicating that although this document type is not especially prone to accrue citations, EPFL is still achieving a high citation impact in its set of proceeding papers.

When analysing the production of highly cited outputs, EPFL has produced a total of 8,843 top 10% highly cited publications (P[top10%]=7,048 of journal papers and P[top10%]=1,795 of proceeding papers), meaning that in proportion EPFL has produced about 19% of its publications with high impact (PP[top10%]=19%).

About 64% of EPFL publications are Open Access (PP[OA]=64%), with a similar presence of OA between journal papers and proceeding papers (64% of OA in journal papers vs. 63% in proceeding papers).

EPFL publications are mostly performed in collaboration, with about 76% of its outputs with some degree of institutional collaboration (PP[collab]=76%), and 65% of all EPFL publications involving co-authors from more than one country (PP[int collab]=65%). In the case of collaboration with industry (indicator PP[industry]), about 10% of all EPFL publications are performed in co-authorship with industrial partners. In the case of proceeding papers, they tend to exhibit a lower presence of institutional collaboration (PP[collab]=63% in contrast with 80% of journal papers) as well as international collaboration (PP[int collab]=50% vs. 69% in journal papers). However, proceeding papers tend to be more often done in collaboration with industrial partners (PP[industry]=12% vs. 9% of journal papers). This may suggest a potential role of proceeding papers at EPFL as conveyors of more local and industry-related research.

Finally, EPFL's publications' level of interdisciplinarity is captured by the indicator IntDisc(0.33). Compared to the overall value of the ETH Domain (IntDisc=0.35), it can be argued that EPFL has a similar degree of interdisciplinary as the domain at large. In Section 3.2 we will discuss the IntDisc values in more detail.

Most of the bibliometric results in Table 1 are provided by document type (proceedings and journals). Readership and author gender statistics are presented at the overall level only. Readership results are based on provided DOIs which were not classified by these types, while author gender could be defined in journal papers only. The results for these indicators are in their proper section (Section 3.2 and 3.5).

Overall, 19% of the EPFL authors is female (13,517 vs 57,778 male, PA[F inst]: 0.19), which is just above the benchmark (all co-authors in the EPFL output, PA[F pubs]: 0.18). The share of female author for the ETH Domain is 20%. The average number of reads (Avg Reads) is 5.26, while the Avg Reads for ETH Domain is 5.09.

Table 1: Overall bibliometric performance statistics EPFL

Indicator	Journals	Proceedings	Overall
Output			
P[full]	35,474	9,599	45,073
P[fract]	16,012	5,971	21,983
Int Cov	0.83	0.60	0.78
InterDisc	0.34	0.32	0.33
P OA [Gold, Hybrid, Green]	22,396	2,207	24,603
PP [OA]	64%	63%	64%
Collaboration			
PP[collab]	80%	63%	76%
PP[industry]	9%	12%	10%
PP[int collab]	69%	50%	65%
Citedness			
TCS	572,950	22,254	595,204
MCS	16.15	2.32	13.21
P[top10%]	7,048	1,795	8,843
PP[top10%]	19%	18%	19%
MNCS	1.59	1.76	1.63
MNJS	1.49	1.50	1.49
PP[self cits]	27%	21%	26%
PP[uncited]	10%	53%	19%
Author gender			
A[F inst]			13,517
A[M inst]			57,778
PA[F inst]			0.19
PA[F pubs]			0.18
RPA[F]			1.03
Readership			
N reads			113,592
N pubs read			21,610
Avg Reads			5.26

The landscape in Figure 3 is a two-dimensional representation of all science (covered by WoS) with an overlay of the output by EPFL researchers in the different research areas. In Annex B we provide a more detailed description of the landscape and the way it is created. The size of a circle reflects the relative number of publications in which EPFL researchers were involved. The colors in the landscape point to 5 main disciplines we use to support the interpretation of the landscape.

Figure 3 captures the topical distribution of EPFL publications across all the research areas of the publication-level classification system of CWTS. As can be seen EPFL has contributed to areas of all the 5 main disciplines of the classification system, although it presents a large concentration of publications in the areas of Physical Sciences & Engineering, while also having a visible publication activity in Maths & Computer sciences and Biomedical & Health sciences. Via this link you can open a web-based version of the landscape in your browser. By opening the menu on the left, you can change the perspective to any of the six ETH institutions.

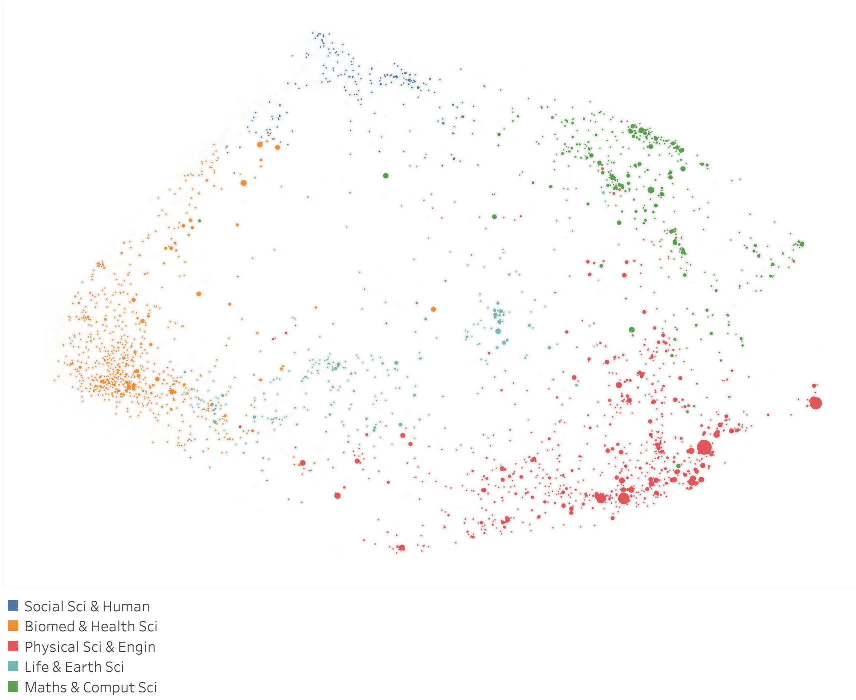

Figure 3: Distribution of EPFL's output across landscape of science (interactive version via this link)

3.1.2 Trends

Table 2 below presents the trend analysis of EPFL by overlapping four-year period of the indicators previously considered. Figure 4 captures the trend evolution of the Journal papers of EPFL, while Figure 5 captures the trend of proceeding papers.

In general, a sustained increasing trend in the number of journal papers published by EPFL is observable in Figure 4. Proceeding papers also exhibit a generally increasing trend overt time (Figure 5), although there is a somewhat decline in the most recent years, particularly from the period 2014-2017 onward. Note that this period also includes the first year of the pandemic (2020).

In addition to the number of publications, EPFL also exhibits patterns of increase in indicators such as IntCov, suggesting an increasing focus on research covered in Web of Science. The growth in the indicator IntDisc indicates an increasing trend in the interdisciplinarity of the research of the institute. The proportion of OA publications (PP[OA]) has also substantially increased from 57% in the period 2009-2012 to about 70% in the most recent period 2017-2020.

The overall impact of the institute as measured by the TCS indicator shows a sustained increase from the initial period 2009-2012 up to the period 2015-2018. There is a decline in the overall TCS impact of EPFL in the more recent periods (2016-2019, 2017-2020). This decline could be partly attributed to the time lag indexing of publications and citations in Web of Science.

The share of female authors at EPFL (RPA[F]) steadily increases from below to above the benchmark over time. Readership is not included in the trend analyses due to missing proper publication year information in DOIs.

Table 2: Trends of EPFL's bibliometric performance

Indicator	2009-2012	2010-2013	2011-2014	2012-2015	2013-2016	2014-2017	2015-2018	2016-2019	2017-2020
P[full]	12,214	13,214	14,240	15,302	16,148	16,653	16,956	17,165	16,711
P[fract]	6,649	7,036	7,437	7,818	7,947	7,906	7,837	7,739	7,387
Int Cov	0.76	0.77	0.78	0.78	0.78	0.78	0.79	0.79	0.80
InterDisc	0.31	0.32	0.32	0.33	0.33	0.33	0.34	0.35	0.35
P [OA]	5,778	6,410	7,042	7,787	8,446	9,066	9,587	10,146	10,379
PP [OA]	57%	59%	60%	61%	63%	64%	65%	68%	70%
PP[collab]	71%	72%	73%	74%	76%	78%	79%	80%	81%
PP[industry]	8%	7%	8%	8%	9%	10%	11%	12%	11%
PP[int collab]	59%	60%	61%	62%	64%	67%	68%	69%	70%
TCS	151,828	175,179	211,260	220,176	240,464	252,555	256,575	245,261	202,912
MCS	12.43	13.26	14.84	14.39	14.89	15.17	15.13	14.29	12.14
P[top10%]	2,427	2,682	2,921	3,104	3,196	3,210	3,286	3,283	3,220
PP[top10%]	19%	19%	20%	19%	19%	18%	19%	19%	19%
MNCS	1.70	1.71	1.71	1.66	1.64	1.60	1.60	1.59	1.56
MNJS	1.50	1.52	1.50	1.49	1.49	1.50	1.49	1.50	1.48
PP[self cits]	23%	24%	25%	25%	26%	27%	27%	27%	27%
PP[uncited]	22%	20%	19%	19%	19%	18%	17%	17%	18%
RPA[F]	0.93	0.99	1.03	1.06	1.07	1.08	1.07	1.07	1.07

In terms of field-normalised impact (i.e., PP[top10%] and MNCS; see Figures 6 and 7), EPFL shows a sustained very high impact during the entire period for both journal papers and proceeding papers. For example, the MNCS value of EPFL journal papers has never been below 1.50, and the unit has systematically published more than 18% of highly cited journal papers in each of the periods of analysis. A similar argument can be made for proceeding papers, with MNCS values always higher than 1.60, and more than 18% of highly cited proceeding papers during the entire period.

There is still some slight decrease in the field-normalised citation impact of EPFL in its journal papers visible if Figure 6. In the case of proceeding papers (see Figure 7), there is a visible decrease in the MNCS indicator of EPFL, although the value of the indicator of PP[top10%] shows an increase from just below 18% in the period 2009-2012 to slightly above 18% in the most recent period 2017-2020.

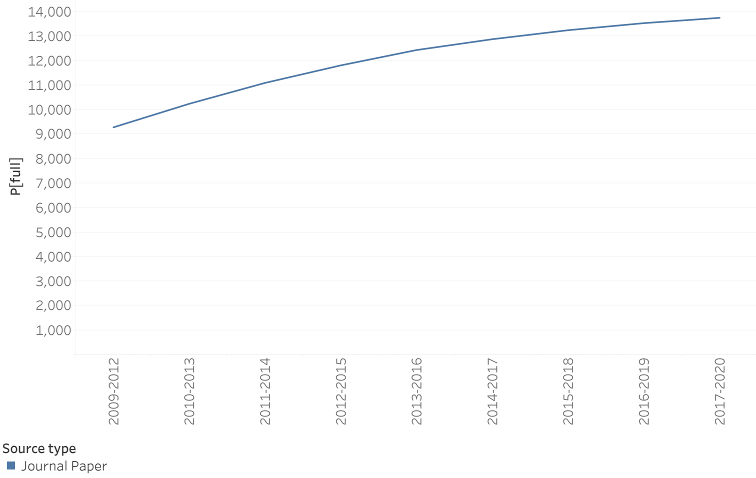

Figure 4: EPFL's journal output trend (P[full]) by overlapping 4-years' period

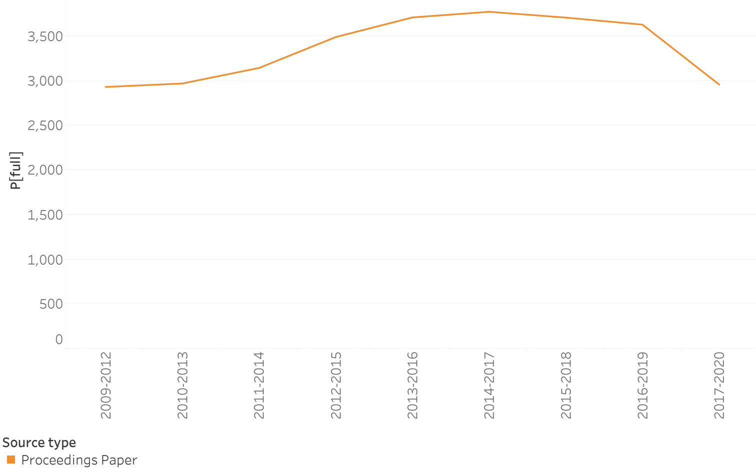

Figure 5: EPFL's proceedings output trend (P[full]) by overlapping 4-years' period

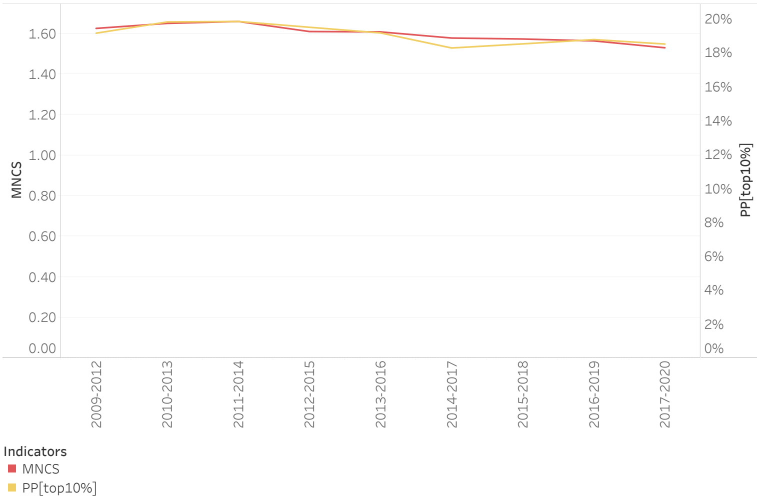

Figure 6: EPFL's journal impact trend (MNCS and PP[top10%]) by overlapping 4-years' period

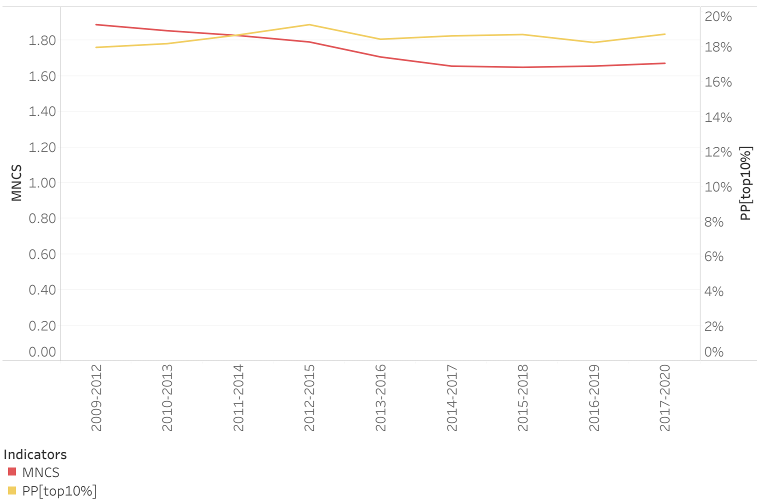

Figure 7: EPFL's proceedings impact trend (MNCS and PP[top10%]) by overlapping 4-years' period

3.2 Research focus in context

Main findings

> The most important subjects for EPFL in terms of output are *Engineering, Electrical & Electronic*; *Physics, Applied*; *Materials Science, Multidisciplinary*; *Chemistry, Multidisciplinary*; *Multidisciplinary Sciences*; *Optics*; *Chemistry, Physical* and *Physics, Multidisciplinary*. The impact of these main subject categories of activity is high. These categories shows also worldwide growth during the last two years. Focusing on share of EPFL's female authors, these categories have a share around benchmark. Finally, EPFL research in these subjects shows lower interdisciplinarity values compared to the benchmark.

3.2.1 Research profile

In this section we break down the output of EPFL into Journal Subject Categories (JSC) to add context to the general statistics. We call this a research profile. It gives a general impression of broad subjects in which EPFL's researchers are involved. The list of categories in the profile is limited to those that represent at least 1% of EPFL's total output.

In each profile we include both P[full] and P[fract], i.e. the number of publications in which EPFL was involved (P[full]) and the number of publications normalised by the number of organisations involved. Note that in such profiles, if a publication is in a journal that belongs to two subject categories, it is assigned half (0.5) to each category. The profile (Figure 8) also shows MNCS, MNJS (second column) and PP[top10%] (third column) per category, to measure impact.

It is important to keep in mind that the indicators displayed in the research profiles are distributed into journal subject categories (since these are well know and recognized discipline categories), while their normalisation has been performed based on the CWTS field categorisation (as these are more fine-tuned, see Annex B).

In addition, we include a growth indicator in Figure 8 for each category: [Field growth] (second column). This value indicates the estimated growth worldwide of a subject category. A [Field growth] above 1 means a positive growth of output worldwide in the most recent two years.

Figure 8 shows that the main subjects of activity for EPFL is *Engineering, Electrical & Electronic*, followed by *Physics, Applied*; *Materials Science, Multidisciplinary*; *Chemistry, Multidisciplinary*; *Multidisciplinary Sciences*; *Optics*; *Chemistry, Physical* and *Physics, Multidisciplinary*. These subjects have at least 3% of the EPFL's total output. The impact scores of these publications are high, with values for PP[top10%] higher than 15%. The impact of *Multidisciplinary Sciences*; *Chemistry, Multidisciplinary* and *Physics, Multidisciplinary* publications stand out with a PP[top10%] of 28% or more.

At the lower part of the profile, we discern also some categories with very high impact: *Computer Science, AI*; *Computer Science, Software Engineering* and *Cell Biology*.

Finally, the [Field growth] indicator shows that most of the subjects present in Figure 8 remained stable or grow worldwide during the last two years, especially *Energy & Fuels*; *Chemistry, Physical* and *Environmental Sciences*. Only one subject is 'shrinking' (*Computer Science, Software Engineering*).

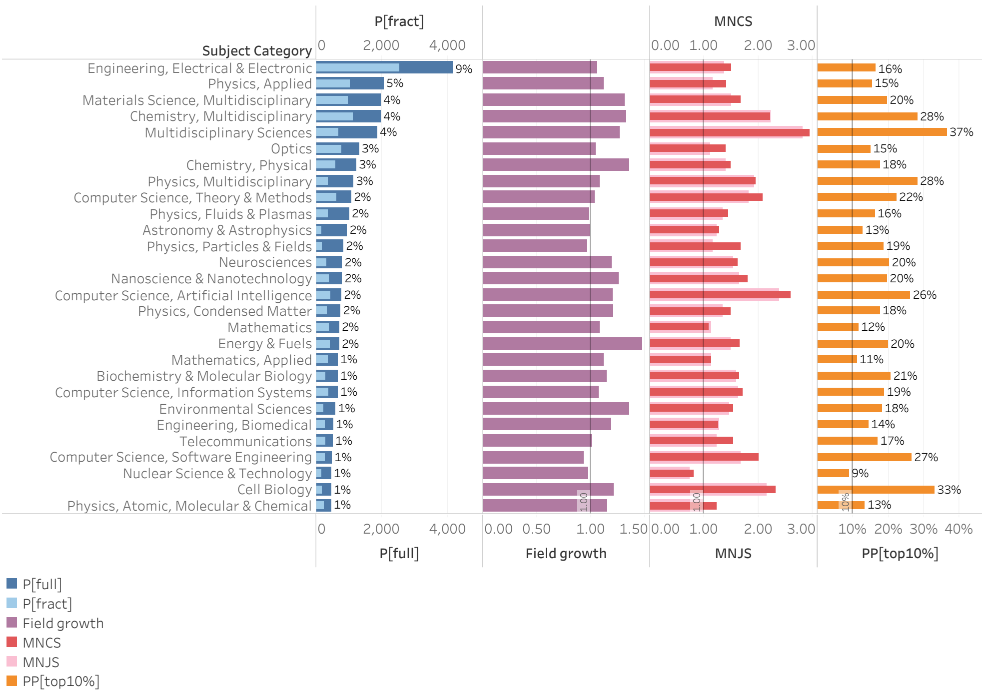

Figure 8: EPFL's research profile (output, impact across subject categories)

3.2.2 Female author contribution across subjects

In Figure 9, we present the same Journal Subject Categories as in Figure 8 and added information related to author gender diversity (RPA[F], third column). EPFL's authors are tagged as male or female using the first name or nickname as it appears on the publication. PA[F inst] indicates EPFL's share of female authors identified for publications (second column). Subsequently, this share is compared with the share of female authors in the publication at large (including all co-authors, PA[F pubs]). The ratio of female authors within EPFL and the share within the publication at large is RPA[F] and visualised in the third column with 1 as a point of reference. A value above 1 means a higher share of EPFL female authors than for all institutions in the same set of publications. For instance, if a publication has 10 authors, of which 3 are female, the PA[F pubs] (reference value) is 0.33. If EPFL is represented by 4 authors, 2 of which are female, the PA[F inst] is 0.5. The RPA[F] would then be 0.5/0.33: 1.52.

A more detailed description of the approach is in Section 2.2. Underlying statistics for EPFL as large can be found in Annex A.

Focusing on the indicator RPA[F], Figure 9 shows that for the majority of the subjects, especially the ones with a higher share of the output, the share of EPFL's female authors is close to the benchmark or slightly below. There is one subject, *Mathematics*, where the share of female authors is 30% higher than the benchmark. On the other side, there are just a few subjects with a value below the benchmark: *Chemistry, Physical*; *Physics, Fluids & Plasmas*; *Physics, Particles & Fields* and *Nuclear Science & Technology*.

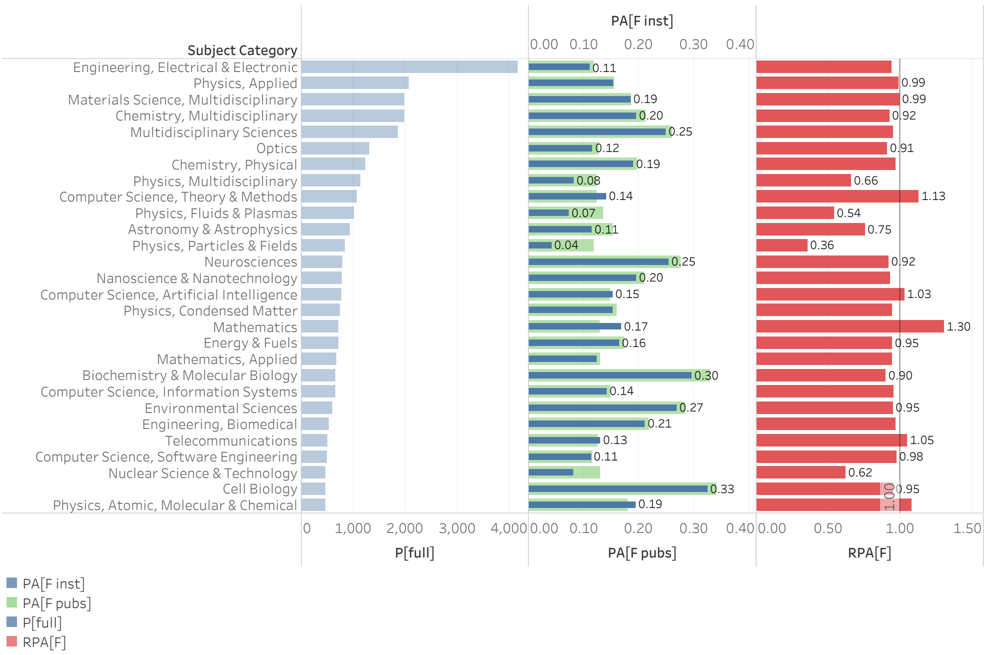

Figure 9: EPFL's share of female authors across subject categories

3.2.3 Interdisciplinary research across subjects

Figure 10 represents interdisciplinarity of EPFL's research output. It uses the same subject categories as in Figure 8 and relies on the publications' references (i.e. other publications cited by the publication of interest). For a more detailed explanation of our definition of interdisciplinary research, see Section 2.2 and Annex D. If a publication cites publications from different subject categories, it is more interdisciplinary than if it cites publications from the same category. In addition, we use a cognitive distance measure to value the diversity of fields being cited. If a paper cites publications from fields that are not closely related (e.g., medical sciences and mathematics) it is more interdisciplinary than if it cites publications from different medical fields. The benchmark we introduce for this indicator is the IntDisc for a subject category at large in 2020.

As Table 1 showed in Section 3.1 the overall value of IntDisc=0.33 for EPFL indicates a relatively low degree of interdisciplinarity, since EPFL research tends to rely on a small set of cognitively nearby disciplines. From a comparative point of view, the degree of interdisciplinarity of EPFL is around the average value of ETH Domain (IntDisc=0.35), therefore not specially high or low within the context of the organization.

At the level of subject categories, Figure 10 shows broad values of interdisciplinarity compared to the overall for specific subjects. Subjects with much lower degree of interdisciplinarity compared to the overall EPFL are *Mathematics* and *Physics, Condensed Matter*. Subjects with much higher degree of interdisciplinarity compared to the overall EPFL are *Engineering, Biomedical* and *Environmental Sciences*.

Figure 10 also shows the overall value of IntDisc per subject categories (grey color). This value is used as the benchmark for the interdisciplinarity values for EPFL (green color). In general, most of the categories show lower IntDisc values compared to the benchmark. Only *Physics, Particles & Fields* and *Neurosciences* are the ones with higher interdisciplinarity value compared to the benchmark.

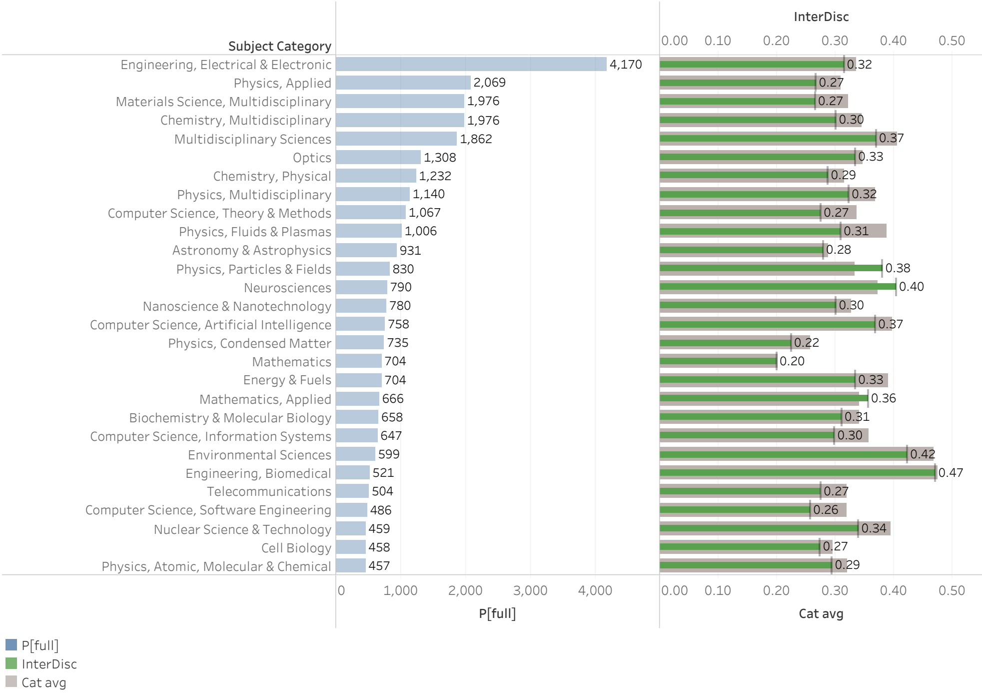

Figure 10: EPFL's interdisciplinarity across subject categories

3.3 Collaboration and partners

Main findings

> For EPFL, proportions of publications with collaboration and international collaboration go up over the analyzed time period, as does the proportion of publications done together with industry. Most publications for EPFL are done in international collaboration, though single institution publications have the highest output. Output done in national collaboration is relatively low. Of all the ETH institutions, EPFL collaborates most with ETH Zurich (1,894 publications), and also has the highest output with this institution. On a country level, affiliations from the United States are collaborated with most frequently.

3.3.1 Collaboration profile

This section includes a trend analysis for the collaboration indicators as well as a collaboration profile.

The trend analysis in Table 3 breaks EPFL's output and collaboration indicators down over time, using overlapping four-year publication windows.

In the collaboration profile in Figure 11, we break down EPFL's output and impact by collaboration type, distinguishing between 'no collaboration' (single author or all authors affiliated with EPFL), national collaboration (all authors having a Swiss affiliation from different institutions) and international collaboration.

Table 3: EPFL's trend collaboration statistics

Indicator	2009–2012	2010–2013	2011–2014	2012–2015	2013–2016	2014–2017	2015–2018	2016–2019	2017–2020
P[full]	12,214	13,214	14,240	15,302	16,148	16,653	16,956	17,165	16,711
PP[collab]	71%	72%	73%	74%	76%	78%	79%	80%	81%
PP[int collab]	59%	60%	61%	62%	64%	67%	68%	69%	70%
PP[industry]	8%	7%	8%	8%	9%	10%	11%	12%	11%

In Table 3, we can observe an upward trend for both PP[collab] (from 71 up to 81%) and PP[int collab] (from 59 to 70%), which is something that is common among the various ETH Domain institutions. The PP[industry] also increases over time though (with some fluctuation), which is less common (for most institutions it remains stable). It seems that from 2013-2016 onward, there is a slight uptake in industry collaboration.

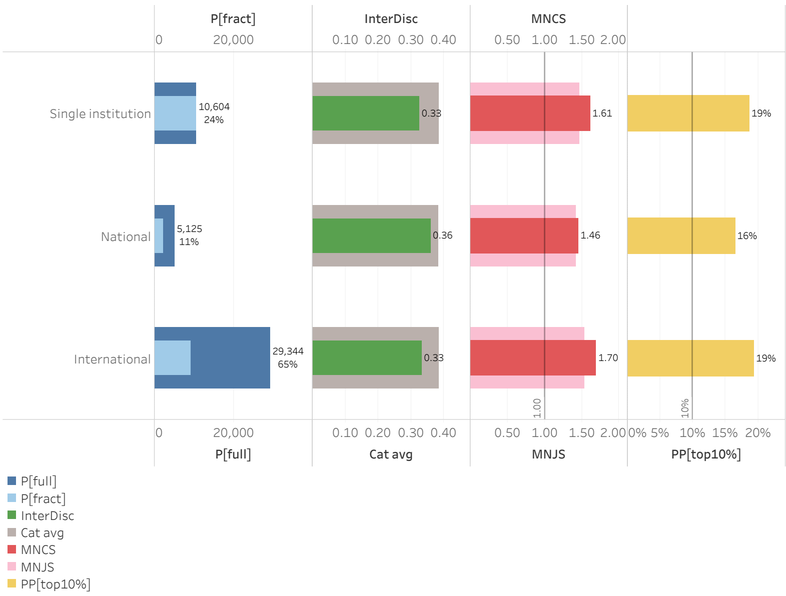

Figure 11: Collaboration profile (output, impact) of EPFL

Immediately notable in Figure 11 is not just the dominance of international collaboration when it comes to research output (the blue columns), but also the low number of (exclusively) national collaboration output. We should note here that publications involving international collaboration may also include national collaborators. It should be noted, too, that while international is dominant for full-counting output (P[full]), for fractional-counting output (P[fract]) single institution is actually higher.

The green bars indicate the interdisciplinarity (IntDisc) measures for the different collaboration types (for more information on how this is calculated, please refer to Annex D). Here it is interesting to see that national collaboration stands out, scoring three decimal points higher than the other two categories. Yet differences are small and do not seem to point to any pattern regarding collaboration type. See section 3.2 for more detailed analysis of the interdisciplinary aspect.

Yet when we look at the impact indicators, the pattern reverses again, with national collaboration lagging behind the other two. Here, international collaboration really stands out with an MNCS of 1.71 and an MNJS of 1.53. We can also see that MNCS (dark-red bars) is consistently higher than MNJS (light-red bars), which means that EPFL's publications outperform the average impact of the journal.

However, for PP[top10%] (in the orange column), single institution and international actually perform on the same level (19%, or almost double the average). Since the PP[top10%] indicator is not influenced by outliers like the MNCS is, the difference between the two suggest that the international collaboration impact number is heightened by a smaller number of very impactful publications.

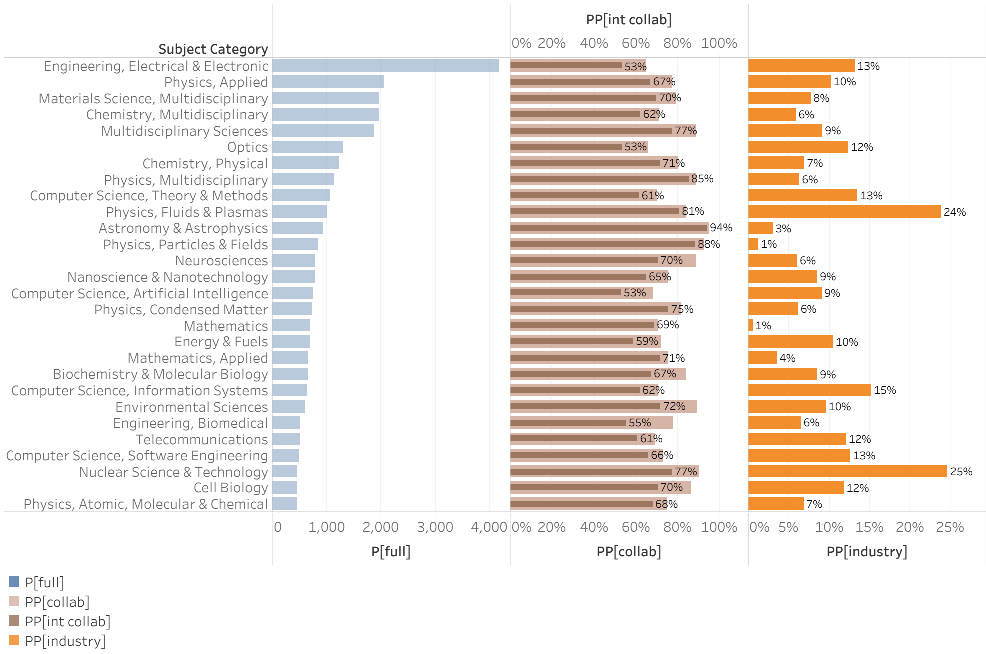

Figure 12: EPFL's output and collaboration types across subject categories

In Figure 12, the collaboration indicators PP[collab], PP[int collab] and PP[industry] are calculated by Web of Science subject category for EPFL publications.

First of all, *Engineering, Electrical & Electronic* is a clear front-runner where output is concerned, having more than double the number of publications of the second-highest category (4,170 vs. 2,069). The next four categories, of which three are to some extent multidisciplinary, also stand out regarding production.

For PP[collab], it is interesting to observe that by far the largest output category, *Engineering, Electrical & Electronic*, also has the lowest collaboration proportion (65%) and shared-lowest international collaboration proportion (PP[int collab]). Also notable is that the highest PP[collab], for *Astronomy & Astrophysics*, is almost entirely made up of international collaboration, given that its PP[collab] and PP[int collab] are almost the same (95 vs. 94%).

For the last column, PP[industry], we see clear differences between the subject categories. Two stand out on top: *Physics, Fluids & Plasmas* and *Nuclear Science & Technology* both feature industry collaboration on roughly a quarter of their publications (24 and 25% respectively). On the low end, two categories stand out too, both with 1%: *Physics, Particles & Fields* and *Mathematics*.

3.3.2 Collaboration within the ETH Domain

Table 4: Co-authorship and impact within the ETH Domain

Indicator	ETH Zurich	EPFL	PSI	WSL	Empa	Eawag
P[full]	1,894	45,073	1,279	390	591	528
MNCS	2.02	1.63	1.45	1.46	1.42	1.58

Table 4 shows EPFL's output and impact (highlighted column), as well as the number of co-publications and impact of EPFL with other ETH institutions.

We can see that most collaboration is done with ETH Zurich and PSI, both above 1,000 publications in co-authorship between the institutions. Given the fact that PSI is a much smaller institution than ETH Zurich, this means a very large share of PSI's output. Collaboration with ETH Zurich also has the highest impact, with an MNCS of 2.02 (or more than 100% above world average). Empa, PSI and WSL constitute the lowest-impact collaborations, with 1.42, 1.45 and 1.46 respectively.

3.3.3 Collaboration outside the ETH Domain

This section seeks to delve deeper into EPFL's collaboration partners outside of the ETH Domain, categorising them first by country and then by institution. Tables 5 and 6 highlight the top collaborators in terms of output. For the results at country level, we used full counting. The output numbers reflect the number and share of output in which countries were involved. For the analysis of co-authoring institutions (Table 6), we used fractional counting. The output numbers indicate the contribution of partnership compared to the total.

The map in Figure 13 highlights countries with more intensive collaboration, with the darkness or intensity of the red indicating the relative level of co-authorship.

In this section we exclude collaborations within the ETH Domain. However, if a publication involves a ETH Domain member and also an external member, it is included.

Country-level

Table 5: Top 12 countries co-authoring with EPFL researchers, excluding ETH Domain internal co-authorship. P[full] and % to EPFL's total

Country	Co-pubs	% to total
United States	8,639	19%
Switzerland	7,617	17%
Germany	6,153	14%
France	5,959	13%
Italy	5,000	11%
United Kingdom	4,801	11%
Spain	3,208	7%
China	3,090	7%
Netherlands	2,578	6%
Russian Federation	2,039	5%
Poland	1,845	4%
Japan	1,816	4%

Table 5 shows the United States as the top collaborating country, followed by Switzerland (here representing all non-ETH Domain collaboration) which in itself is ahead of Germany and the rest of the world.

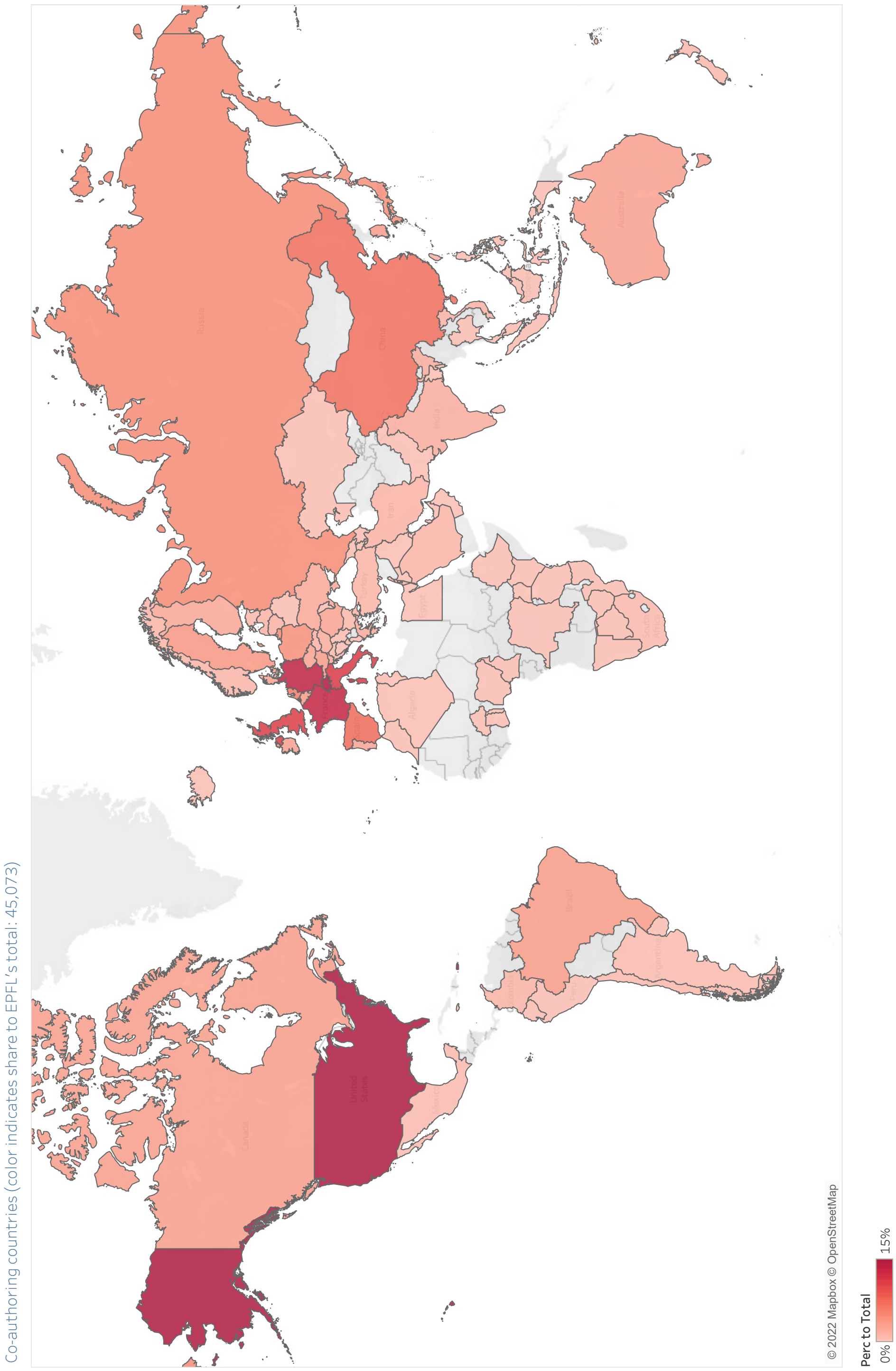

Figure 13: Map of countries co-authoring with EPFL

Institutions

Table 6: Top 20 collaborating institutions of EPFL, excluding ETH Domain internal co-authorship (fractional output and impact)

Inst	Country	Co-pubs	MNCS
University of Lausanne	CH	757	1.50
University of Geneva	CH	532	1.51
Max Planck Society for the Advancement of Science	DE	420	2.07
Centre National de la Recherche Scientifique	FR	282	1.66
Idiap Research Institute	CH	173	1.79
CERN European Organization for Nuclear Research	CH	163	1.59
Massachusetts Institute of Technology	US	155	2.27
University of California, Berkeley	US	150	2.32
Russian Academy of Science	RU	142	1.46
Harvard University	US	139	2.23
University of Zurich	CH	137	2.05
Politecnico di Milano	IT	131	1.44
University of Bern	CH	130	1.50
Chinese Academy of Sciences	CN	120	2.52
Consiglio Nazionale delle Ricerche	IT	119	1.36
Universidade de Lisboa	PT	117	1.33
Istituto Nazionale di Fisica Nucleare	IT	116	2.11
Stanford University	US	109	2.87
Delft University of Technology	NL	105	2.37
University of Cambridge	GB	104	2.35

Table 6 is dominated by two Swiss universities. While the United States was on top of Table 5, we only find American universities lower down this list, suggesting that that number for the United States is spread out among more institutions. Out of the institutions at the top of the list, the Max Planck Society for the Advancement of Science stands out in terms of impact, with an MNCS of 2.07. If we go down to the bottom of the list, there are some high-impact collaborations, with Stanford being particularly notable (2.87).

3.4 Research accessibility

Main findings

EPFL's research is published increasingly in Open Access. The number (and share) of Gold and Hybrid OA publications has grown steadily during the period 2009 up to 2020, while the number of Closed Access publications drops in the most recent years. Moreover, the impact of OA publications remains at a high level throughout, while the impact of Closed Access publications is structurally lower and decreases somewhat.

3.4.1 OA publishing and impact

In this section we discuss the accessibility of EPFL's research output. For publications with a DOI we could define whether it was published Open Access (OA) or not based on Unpaywall data (version July 2021). Therefore, the below statistics only include publications for which we could define OA or not. In addition, we could also determine the type of OA (Gold, Hybrid or Green). The trend analyses allow us to monitor the evolution of EPFL regarding OA publishing.

Using OA information we assess the overall accessibility of EPFL's OA output as well as its citation-based impact, by benchmarking it to non-OA output.

Table 7: EPFL's Open Access (OA) performance statistics by type, excluding publications for which no OA info available

Indicator	OA Gold	OA Hybrid	OA Green	Closed Access	Total
P[full]	5,458	3,935	15,210	13,945	38,548
P[top10%]	1,046	984	3,422	2,360	7,812
PP[top10%]	20%	23%	22%	17%	20%
PP[int collab]	70%	76%	69%	64%	68%

In Table 7, we provide an overview of main performance statistics for three types of OA (Gold, Hybrid and Green) together with their overall performance. P[full] reflects the total number of publications, P[top10%] the number belonging to the top 10% most cited (within its year and field). PP[top10%] assesses the impact of each type, while PP[int collab] reflects the share of output involving international collaboration.

Looking at the entire period (2009-2020), we see a preference for Green OA publications (P[full]). The impact is particularly high for Hybrid and Green OA publications (PP[top10%]). The share of output involving international collaboration is highest for Hybrid OA output (PP[int collab]: 76%). Both PP[top10%] and PP[int collab] are higher for all types of OA publications, compared to Closed Access publications.

Table 8: EPFL's performance statistics trend, Closed vs. Open Access publications

	Indicator	2009–2012	2010–2013	2011–2014	2012–2015	2013–2016	2014–2017	2015–2018	2016–2019	2017–2020
Closed	P[full]	4,368	4,533	4,689	4,877	5,036	5,104	5,056	4,817	4,541
	P[top10%]	790	822	858	849	841	809	787	761	729
	PP[top10%]	18%	18%	18%	17%	16%	15%	15%	16%	16%
	PP[int collab]	58%	59%	61%	62%	64%	66%	68%	69%	70%
Open	P[full]	5,778	6,410	7,042	7,787	8,446	9,066	9,587	10,146	10,379
	P[top10%]	1,307	1,480	1,641	1,808	1,937	2,024	2,143	2,195	2,208
	PP[top10%]	21%	22%	22%	22%	22%	22%	22%	21%	21%
	PP[int collab]	66%	66%	67%	68%	69%	71%	72%	73%	73%

In Table 8, we provide trend results for the same indicators as in Table 7, comparing OA publications with non-OA (Closed Access) publications. These results only include publications for which OA information was available (included in Unpaywall, have a DOI). In Figures 14 and 15, P[full] and P[top10%] are depicted by OA type.

The results in Table 8, show the steady increase of OA publications. OA output almost doubles during the studied period. For P[top10%] we see the same. Normalised by the total number of output per year, the PP[top10%] remains at the same high level (PP[top10%] of 22%) throughout the studied period for OA publications. The impact of Closed Access output decreased somewhat down to 16% in the most recent years but is still at a high level.

From the collaboration perspective, we see that the proportion of output involving international collaboration increased for both types of output, albeit that this proportion is somewhat higher for OA output (PP[int collab]).

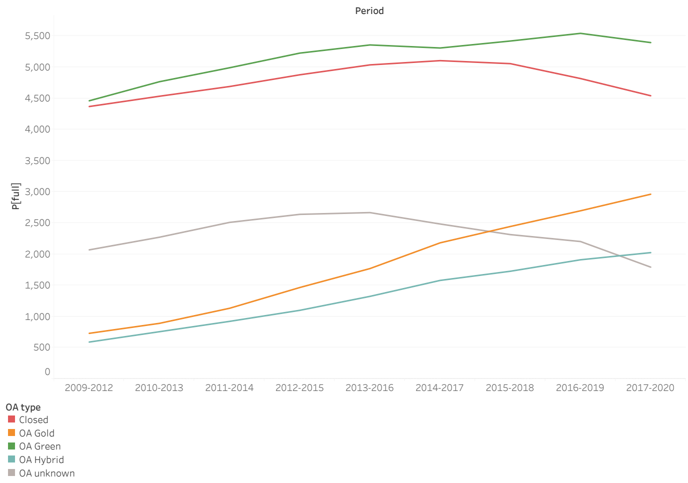

Figure 14: EPFL's output trend by Open Access (OA) type

In Table 2, we already showed the increase of the number and proportion of EPFL's OA publications. In Figure 14, this is visualised in more detail for the different types of OA. In particular Gold and Hybrid OA publications show a steady increase, while Green has been the preferred type since 2009. As Green OA usually does not depend too much on costs and is more easily managed, it implies that EPFL was already proactively publishing OA and gradually adopting more and more Gold and Hybrid OA. The absolute number of Closed Access publications stabilised and even decreased somewhat since 2015.

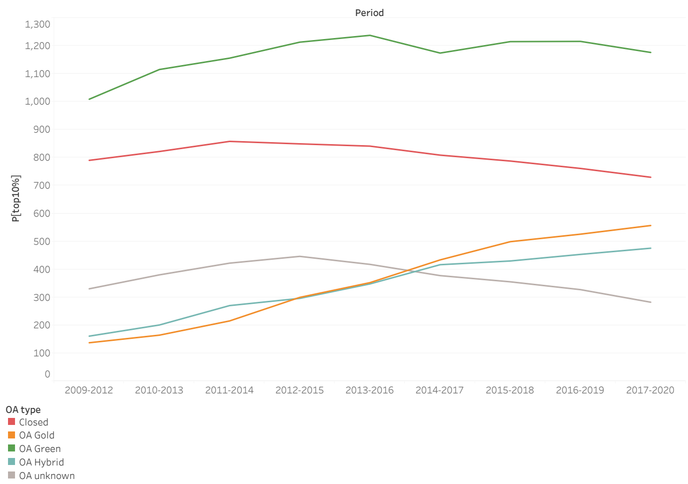

Figure 15: EPFL's trend of top 10% publications by Open Access (OA) type

Along with the number of Gold and Hybrid OA publication, the number of Gold and Hybrid top 10% publications increases since 2009. Moreover, the number of Green OA top 10% publications is by far the highest throughout.

3.4.2 OA publishing and impact by subject

In this section we present EPFL's performance statistics by journal subject category. In Figure 16, we visualise the share of OA publications, related to the overall output (for which access information was available). The bars in the second column of the diagram represent the ratio of the sum of OA publications to the sum of all publications. The light blue bar in the profile in the first column represents the total number of publications. The list of subject categories is limited to those that cover at least 1% of the total output of EPFL.

In Figure 17, the second column visualises the impact of both Closed and Open access publications by PP[top10%] by subject.

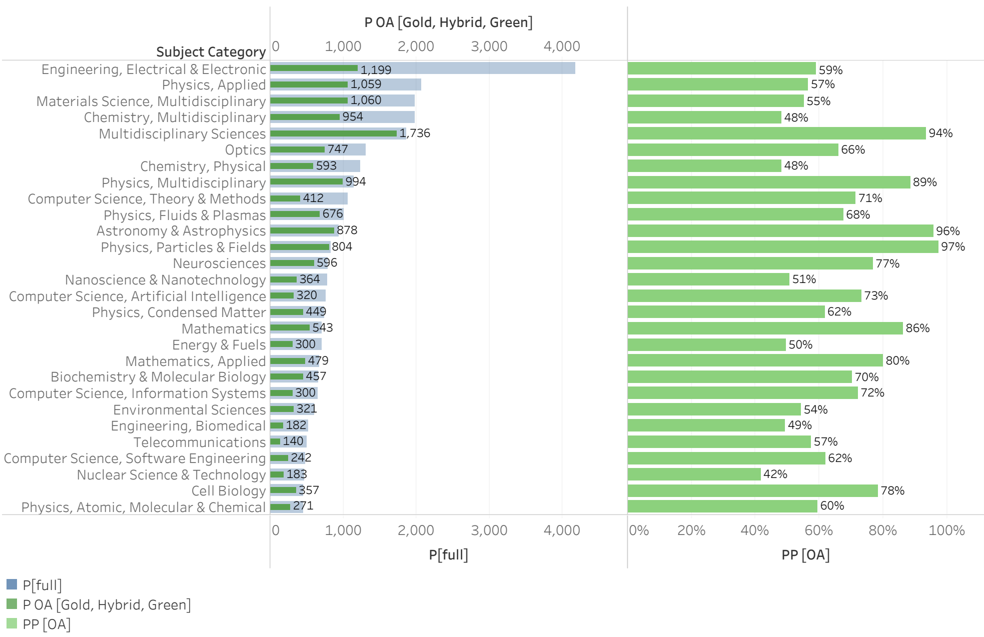

Figure 16: EPFL's output and share of OA publications across subject categories

The profile in Figure 16, show high shares of OA publications in almost all categories, while a few still stand out. For *Multidisciplinary Sciences*; *Astronomy & Astrophysics* and *Physics, Particles & Fields* we found more than 90% published OA. there are also a few with less than 50% OA publications (e.g., *Nuclear Science & Technology* with 42%) but these are exceptions.

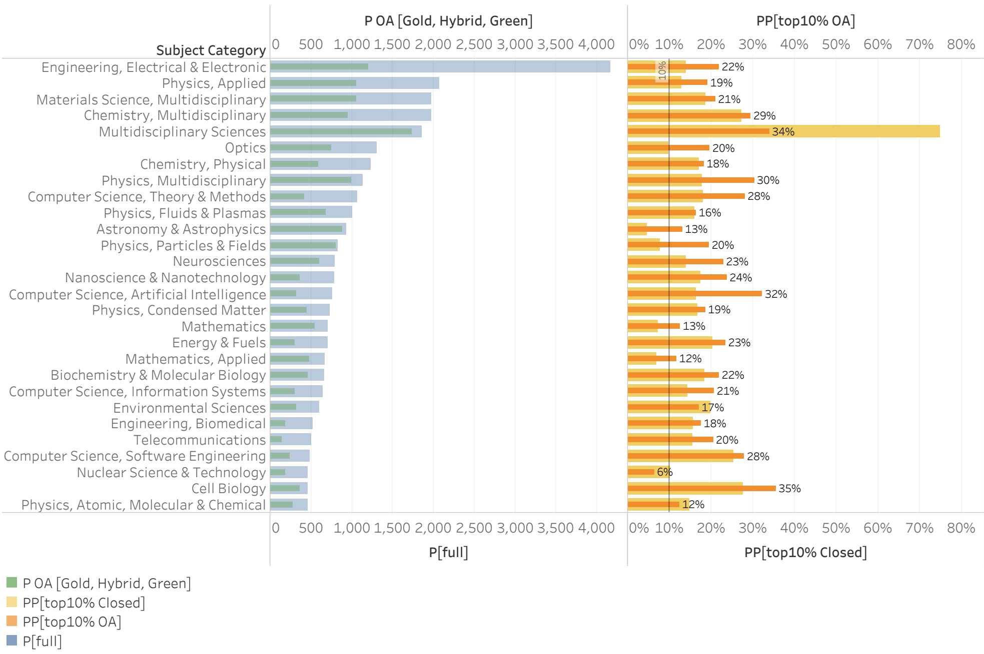

Figure 17: EPFL's impact distribution (PP[top10%]) of Open and Closed output across subject categories

The PP[top10%] of almost all EPFL's categories is higher for OA publication than for Closed Access publications. The usual exception is *Multidisciplinary Sciences*, where publication in journals like *Nature* and *Science* still boost the citation-based impact.

The results in this section show that EPFL has an effective OA policy. In almost all subjects, not only the percentage of OA publications is high but also the number and share of top 10% publications. The large share of Green OA publications throughout the period of analysis suggests that EPFL is an early OA adopter.

3.5 Impact and knowledge use

Main findings

> EPFL's research is read and cited from all over the world. The citation-based impact is primarily determined by institutions located in Europe, Asia and the United States. The readership analysis also shows significant impact of EPFL's research in countries that are not well represented in WoS as these countries (e.g., Brazil).

In this section, we discuss the actors (countries, institutions) that define the impact and use of EPFL's research. We estimate the impact and use by analysing (1) the publications citing EPFL's publications and (2) the country of people reading its publications.

The analysis of publications citing EPFL's output shows the most prominent countries and institutions. Thus we provide an overview of the geographical distribution of EPFL's impact and more specifically the institutions that use EPFL's research.

The readers are analysed using Mendeley data, in which a 'read' is defined by a person (i.e., Mendeley user) saving a publication. The results should be interpreted with that disclaimer in mind. The user information includes the country of origin (if available). In this report, we will present the countries and compare these to the ones citing EPFL's output. Including readership in this study does not show a broader (e.g., societal) impact of EPFL research but merely catches the (potential) scientific impact beyond the WoS data.

3.5.1 Impact and knowledge use at country level

The citation-based impact is defined by publications citing EPFL's output. In these citing publications, we use the affiliations of authors to measure their contribution to the impact of EPFL's research. Table 9 shows the 20 most prominent countries citing EPFL's research output. In the table we include the number of EPFL publications being cited, the number of citations they receive and the average number of citations per publication. The top 20 is defined by the number of citations received (N cits). This list is obviously dominated by countries with many publications in WoS, and we cannot deny their significant role in determining the citation-based impact. By considering the top countries and subsequently looking at the average number of citations given, we normalise to some extent the results.

Table 9: EPFL given citations by country (top 20 most given citations)

Country	N pubs	N cits	Avg cits
United States	24,425	123,001	5.04
China	21,441	108,269	5.05
Germany	16,978	46,607	2.75
United Kingdom	15,238	37,923	2.49
France	13,679	28,260	2.07
Italy	11,863	25,830	2.18
Japan	10,615	23,926	2.25
Spain	10,105	19,925	1.97
Switzerland	10,967	19,288	1.76
Canada	9,755	17,546	1.80
South Korea	8,106	17,238	2.13
India	8,217	16,473	2.00
Australia	8,558	15,720	1.84
Netherlands	7,599	12,912	1.70
Russia	5,589	10,527	1.88
Sweden	5,731	9,104	1.59
Brazil	4,896	8,046	1.64
Taiwan	4,693	7,630	1.63
Singapore	4,620	7,352	1.59
Belgium	5,134	7,069	1.38

In Table 9, we clearly see the dominance of the United States and China defining EPFL's impact. Not only by absolute numbers of citations but also by the averages, these two countries attribute great value to EPFL's research. Both countries cite on average a EPFL publication more than 5 times. Next in line are researchers from other European countries, Japan, Canada, South Korea, India, Australia, Russia, Brazil, and Singapore with between 1 and 3 citations per publication on average. The high average of non-European countries Japan (2.25),South Korea (2.13) and India (2.00) stand out.

In Table 10, we introduce a different perspective on the impact EPFL's research has. By looking at the number of reads by Mendeley users from different countries, we get a better view on the geographical distribution beyond the perimeter of the academic debate (as defined by citations). We realise that this distribution is defined primarily by the authors citing EPFL's output but we hope to broaden the view on the impact somewhat. The List in Table 10 shows the top 20 most prominent countries 'reading' EPFL's publications. The list order is defined by the number of reads (second column: N reads). In the table the first column shows the number of publications being read (N pubs). The third column shows the average number per read publication (Avg Reads). We consider the countries that end up in the readership list (Table 10) but not in the citing countries list (Table 9) as the ones showing the impact beyond the WoS.

Table 10: EPFL readership by country (top 20, by most reads)

Country	N pubs	N reads	Avg Reads
United States	9,825	24,095	2.45
United Kingdom	6,236	11,287	1.81
Germany	5,683	10,001	1.76
Switzerland	4,591	6,142	1.34
France	4,049	5,941	1.47
Spain	3,235	4,515	1.40
Japan	3,241	4,482	1.38
Canada	2,690	3,534	1.31
Italy	2,787	3,505	1.26
Brazil	2,313	3,317	1.43
India	2,344	3,027	1.29
Netherlands	2,257	2,993	1.33
China	2,174	2,732	1.26
Belgium	1,708	2,044	1.20
Portugal	1,280	1,644	1.28
Sweden	1,233	1,505	1.22
Australia	1,260	1,494	1.19
Denmark	1,142	1,436	1.26
Russia	1,114	1,245	1.12
Austria	856	1,050	1.23

From the reader perspective we see some interesting results, comparing them to Table 9. First of all, lower numbers for China, which is an artefact of the data being used. Chinese researchers and academics do not use Mendeley to manage their literature (Fairclough and Thelwall, 2015; Zahedi and Costas, 2020). A similar argument applies to South Korea (absent in this list). In addition, we see a more prominent position of Brazil in this list, in absolute numbers but also on average. In this list of top 20 countries, Brazil is second with 1.89 reads per publication after the US.

3.5.2 Impact by citing institution

In Table 11, we list the top 20 most prominent citing institutions of EPFL's publications. This list provides more insight in the actual research actors being impacted by EPFL. As the list is based on the number of citations given (N citing pubs, second column), it will be biased towards large institutions (with many publications). We normalise these large numbers by including the number of publications being cited (N cited pubs, first column), which leads to the average in the third column (Avg cits).

Table 11: EPFL's top 20 most citing institutions (by number of given citations)

Institution	Country	N cited pubs	N citing pubs	Avg cits
CHINESE ACAD SCI	CN	8,068	19,335	2.40
CNRS	FR	9,570	17,093	1.79
MAX PLANCK SOCIETY	DE	4,537	7,570	1.67
HARVARD UNIV	US	3,588	6,530	1.82
UNIV CHINESE ACAD SCI	CN	3,627	6,311	1.74
ECOLE POLYTECN FEDERALE LAUSANNE	CH	5,029	5,798	1.15
RUSSIAN ACAD SCI	RU	3,233	5,115	1.58
MIT	US	3,496	4,900	1.40
UNIV CAMBRIDGE	GB	3,357	4,739	1.41
TSING HUA UNIV	CN	3,169	4,623	1.46
UNIV TOKYO	JP	2,868	4,622	1.61
PEKING UNIV	CN	2,748	4,424	1.61
STANFORD UNIV	US	3,204	4,383	1.37
UNIV CALIF BERKELEY	US	3,036	4,214	1.39
ETH ZURICH	CH	3,337	4,130	1.24
UNIV OXFORD	GB	2,923	4,046	1.38
SHANGHAI JIAO TONG UNIV	CN	2,707	3,707	1.37
UNIV PARIS-SACLAY EPE	FR	2,677	3,661	1.37
ZHEJIANG UNIV	CN	2,609	3,643	1.40
UNIV SCI & TECHNOL CHINA	CN	2,383	3,592	1.51

This table too is dominated by the largest research institutions in the world who produce many WoS publications and are located in the countries in Table 9: the Chinese Academy of Science, CNRS, and Max Planck Society being mega-institutions. EPFL is the sixth institution contributing to its own impact, and we need to emphasise that these citations do not include author self-citations.

References

Emile Caron and Nees Jan van Eck. Large scale author name disambiguation using rule-based scoring and clustering. In *19th International Conference on Science and Technology Indicators*, volume 19, pages 79–86, Leiden, September 2014.

Ruth Fairclough and Mike Thelwall. National research impact indicators from Mendeley readers. *Journal of Informetrics*, 9(4):845–859, October 2015. ISSN 1751-1577. doi: 10.1016/j.joi.2015.08.003. URL https://www.sciencedirect.com/science/article/pii/S1751157715300596.

Javier Ruiz-Castillo and Ludo Waltman. Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, 9(1):102–117, January 2015. ISSN 17511577. doi: 10.1016/j.joi.2014.11.010. URL https://linkinghub.elsevier.com/retrieve/pii/S1751157714001126.

Ludo Waltman and Nees Jan van Eck. A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12):2378–2392, December 2012. ISSN 15322882. doi: 10.1002/asi.22748. URL http://doi.wiley.com/10.1002/asi.22748.

Ludo Waltman, Clara Calero-Medina, Joost Kosten, Ed C. M. Noyons, Robert J. W. Tijssen, Nees Jan van Eck, Thed N. van Leeuwen, Anthony F. J. van Raan, Martijn S. Visser, and Paul Wouters. The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12):2419–2432, December 2012. ISSN 1532-2890. doi: 10.1002/asi.22708. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22708.

Zohreh Zahedi and Rodrigo Costas. Do Online Readerships Offer Useful Assessment Tools? Discussion Around the Practical Applications of Mendeley Readership for Scholarly Assessment. *Scholarly Assessment Reports*, 2(1):14, November 2020. ISSN 2689-5870. doi: 10.29024/sar.20. URL http://www.scholarlyassessmentreports.org/articles/10.29024/sar.20/. Number: 1 Publisher: Levy Library Press.

Annexes

A EPFL's author gender statistics

Table 12: EPFL: Underlying gender diversity statistics

Indicator	Value
A[F inst]	13,517
PA[F inst]	0.19
A[FM inst]	71,295
A[F pubs]	48,911
PA[F pubs]	0.18
A[FM pubs]	265,263
RPA[F]	1.03

The indicators presented in this table are described in Section 2.2, p. 17.

B Publication level classification

The CWTS citation database is a bibliometric version of Web of Science (WoS). One of the special features of this database is the publication-based classification. This classification is an alternative to the WoS journal classification, the WoS subject categories. The reason to have this publication-based classification is the problems we encounter using the journal classification for particular purposes. We discern the following as the most prominent ones.

B.1 Journal scope (including multi-disciplinary journals)

A journal classification introduces sets of journals to represents a class, in this case a subject category. This implies that journals have a similar scope. They do not need to be comparable with regard to volume (number of articles per year) but they should represent a similar specialisation. This is not the case, of course. Journals represent a very broad spectrum. There are very specialist journals (e.g., Scientometrics) and very general ones (e.g., Nature or Science but also British Medical Journal). The classification scheme can therefore not be very specialised. In WoS, a subject category Multi-disciplinary hosts the very general ones so that a bibliometric analysis of, for instance, the Social Sciences or Nanotechnology, using this classification, will not take papers in Nature into consideration.

B.2 Granularity of the WoS subject categories

The WoS journal classification scheme contains 255 elements. As such it is a stable system. In many cases however, it appears that these 255 subject categories are insufficient to be used for proper field analyses. The problem is that the granularity of the system looks somewhat arbitrary. 'Biochemistry & Molecular Biology' on the one hand and 'Ornithology' on the other, for instance, represent rather different aggregates of research. This is illustrated by the number of journals in each of them. Where the 'Biochemistry & Molecular Biology' category contains almost 500 journals, 'Ornithology' has only 27. We acknowledge that there is no perfect granularity, but we argue that in the WoS subject categories the differences are really too big. A classification based on more objective grounds does not solve this problem but is at least transparent.

B.3 Multiple assignment of journals to categories

In journal classifications from multi-disciplinary databases, journals are assigned to more than one category. Journals often have broader scopes than the categories allow. Also here there are large differences between categories. In the example we used before, 'Biochemistry & Molecular Biology,' journals are on average assigned to almost 2 categories. This means that (on average) each journal in this category is also assigned to one other category. For the more specialist category of 'Ornithology', the average is 1. This means that in this category all journals are assigned to this category only. If publications in journals with a multiple assignment would always cover the categories at stake, this should not necessarily be a problem. However, it mostly means that such journals structurally contain publications from the different categories. Therefore, publications may be assigned to two categories although they belong to just one of them.

B.4 The CWTS publication-based classification scheme

CWTS has developed an advanced alternative for the Web of Science journal classification. It counters three major issues:

1. Journal scope (including multi-disciplinary journals)
2. Granularity of the WoS subject categories
3. Multiple assignment of journals to categories

The CWTS publication-based classification is developed as described in Waltman and van Eck (2012). Since the first version there have been yearly updates of the system. The main characteristics of the classification are as follows.

Publication to publication citation clustering

Clusters of publications are created on the basis of citations from one publication to another. Tens of millions of publications have been processed. The clusters contain publications from multiple years (2000–2020). Each publication is assigned to one cluster only at each level. A cluster is considered, and in many cases validated as, representative for disciplines, research areas, fields or sub-fields. For each cluster, we can calculate growth indices pointing at changing research focus over time.

Multi-level clustering

The classification scheme has at present three different levels. The clusters are hierarchically organised. Currently we discern the following levels.

1. A top level of 25 clusters (fields)

2. A second level of around 800 clusters (sub-fields)

3. A third level of more than 4,000 clusters (research areas or micro-fields)

A common way of visualising the landscape of science by the publication clusters is a 2-dimensional map. In such a landscape (see Figure 18), we position publication clusters in relation to each other on the basis of citation traffic. The denser the traffic between two clusters, the closer they are. The two dimensions do not represent anything. The only thing that matters is the distance. Furthermore, the size of a cluster represents the relative volume (number of publications included), while the color coding adds a main clustering labeled by main disciplines.

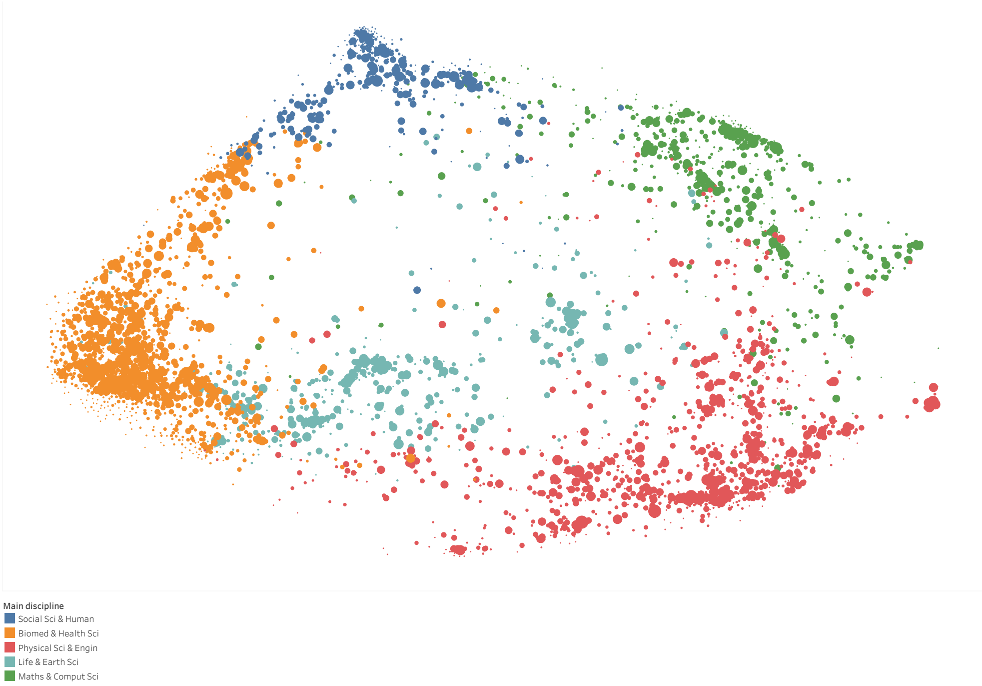

Figure 18: Landscape of all science (around 30 million WoS publications). Circles represent (over 4,000) publication clusters. Position is defined by citation traffic between clusters. Size indicates relative volume. Color reflects 5 main disciplines

C Citation data and analysis

In this annex we provide more detail about the methodology developed at CWTS and applied in this study.

C.1 Database coverage

In a bibliometric study, we base the analyses on publication data. To relate counting and measuring to standards, we depend on international bibliographic databases, such as Web of Science, Scopus, Dimensions. We realise that by using such databases, we may be missing relevant scientific outputs and achievements. In order to assess how much the database *does* cover we calculate the Internal Coverage (**IntCov**) indicator. This indicator is the ratio of cited references covered by the database, to the total number of cited references. If a publications contains 10 references, five of which are also in the database, the IntCov of this publication is 0.5. For a set of publications the IntCov is defined by the average IntCov per publication. If the IntCov of an institution's output in WoS is 0.8, we estimate the coverage of WoS for this institution at 0.8 (80%).

C.2 Database Structure

At CWTS, we calculate bibliometric indicators based on an in-house version of the Web of Science (WoS) online database, which will be referred to as the CI-system. The WoS is a bibliographic database that covers publications of about 12,000 journals and each of these journals is assigned to one or more Journal Subject Categories (JSC). Each publication in the CI-system has a document type. The most frequently occurring document types are 'articles', 'reviews', 'proceeding papers', 'corrections', 'editorial material', 'letters', 'meeting abstracts' and 'news items'. In this report, we only consider document types 'articles', 'reviews' and 'proceedings papers'. In limiting the analysis to these three types of publications, we consider that these documents reflect most of the original scientific output in a field.

The CI-system is an improved and enhanced version of the WoS database versions of the Science Citation Index (SCI), Social Science Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI). The CI-system implements a publication-based field classification which clusters publications into research areas based solely on citation relations (Waltman and van Eck, 2012) (more detail in Annex B). One important advantage of this publication-level classification system is that it allows for a taxonomy of science that is more detailed and better matches the current structure of scientific research. This not only reduces classification bias but is also essential for calculating field-normalised indicators (Ruiz-Castillo and Waltman, 2015).

Moreover, in this study we include citation data up to 2021. Please note that publications require at least one full year to receive citations in order to make robust calculations of citation impact indicators. For this reason, we will work with publications up to and including 2020, counting citations up to and including 2021. For each publication (and its benchmark publications), we consider 4 years of citations since the year of publication. For a publication from 2010, we count citations in the years 2010-2014.

C.3 Citation Window, Counting Method and Field Normalisation

Citation window

Several indicators are available for measuring the average scientific impact of the publications of a research unit, e,g. and institution. These indicators are all based on the idea of counting the number of times the publications of a unit have been cited. Citations can be counted using either a fixed-length citation window or a variable-length citation window. In the case of a fixed-length citation window, only citations received within a fixed time period (e.g. four years fixed window) are counted. The main advantage of a fixed-length citation window is that it is possible to meaningfully analyse the trend patterns of the non-normalised impact indicators, setting the same criteria for all publications included. A variable-length window, on the other hand, uses all the citations that are available in the database until a fixed point in time, which not only yields higher citation counts (depending on the window length), but also more robust impact measurements. When using a variable-length citation window, impact indicators such as the average impact (MCS) and the total impact score (TCS) may systematically present a decreasing pattern.

In this study, we use a fixed-length window of 4 year (if available) for the overall period of the analysis (2009-2020). The most recent year for receiving citations is 2021.

Self-citations

In the calculation of advanced citation impact indicators, we disregard self-citations. A citation is considered a self-citation if the cited publication and the citing publication have at least one author (i.e. last name and initials) in common. The main reason for excluding self-citations is that they often have a different purpose from ordinary citations. Specifically, self-citations may indicate how different publications of a researcher build on one another, or they may serve as a mechanism for self-promotion rather than for indicating relevant related work. Self-promotion can in turn be used to manipulate the impact of a publication in terms of the number of citations received. Excluding self-citations from the analysis effectively reduces the sensitivity of impact indicators to potential manipulation. In doing so, impact indicators can be interpreted as the impact of researchers' work on other members of the scientific community rather than on his or her own work.

Field Normalisation

There can be quite large differences in citation practices in different scientific fields. Field normalisation is about correcting for differences in citation practices between different scientific fields. The goal of field normalisation is to develop citation-based indicators that allow for valid between-field comparisons.

In this report, we will use our in-house publication-based classification system of science to define the scientific fields that are used in this normalisation process. This system has three major advantages compared to the conventional journal-based classification systems of science: Web of Science Journal Subject Categories:

- Proper granularity in terms of fields.
- Fields are defined at the level of publications citing each other, not on allocating complete journals to field(s) where inaccuracies are introduced.
- Publications from journals like Nature, Science, PLoS ONE (multidisciplinary journals) are allocated to the field they actually belong to and not to the artificial journal field 'Multidisciplinary Sciences'.

The reasons to use this publication-based classification are furthered explained in Annex B.

Counting method

Counting methods are about the way in which co-authored publications are handled. For instance, if a publication is co-authored by researchers from two countries, should the publication be counted as a full publication for each country or should it be counted as half a publication for each of them? In this study, we use both full and fractional counting. Full counting means that if a publication is co-authored by multiple organisations, that publication counts multiple times, once for every organisation, regardless of the weight of their contribution. In this report, we use mainly the full counted publications for output and fractionalised (by number of institutions involved) for impact measures.

D Interdisciplinary research

While there are different understandings of interdisciplinarity, the definition that has gained more consensus is the one provided by the US National Academy of Sciences (2005) that states:

> "Interdisciplinary research (IDR) is a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialised knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or field of research practice."
>
> *https://www.nap.edu/read/11153/chapter/4*

There are two key elements in this definition we consider as basic notions to articulate our proposal: the concept of integration and the idea of combining knowledge from two or more disciplines.

We characterise interdisciplinarity at the level of each individual publication, by analysing the disciplines cited by the publication. This approach will allow us to consider the citations to distinct disciplines by the same citing publication as a proxy of the integration of knowledge from different disciplines. For this analysis we consider the Web of Science Journal Subject Categories as disciplines. We analyse the degree or extent of integration through the concept of diversity. Diversity is based on three concepts: variety, balance and disparity. We operationalise interdisciplinarity using Rao-Stirling diversity, an indicator which captures the three inter-related concepts of diversity, and is computed as follows:

$$\Delta = \sum_{ij} p_i p_j d_{ij}$$

$$(i \neq j)$$

> Where pi is the proportion of cited references in the subject category i, pj is the proportion of cited references in the subject category j, and dij is the cognitive distance between the subject categories i and j

In this formula, disparity refers to the cognitive distance existing between two scientific disciplines (or subject categories, in our case). In order to compute the disparity measure, we will create a similarity matrix Sij for the WoS subject categories based on the of citation flows between them. This will be then transformed into a Salton's cosine similarity matrix in the citing dimension. In this transformed matrix, the Sij represents the similarity between each pair of WoS categories, thus the cognitive distance (d) between two subject categories can be computed as d = 1- Sij.

The indicators of interdisciplinarity will allow us to identify an institution's subject categories of a prepresenting the most interdisciplinary research.

We apply the state of the art in analysing interdisciplinarity using bibliometric techniques. However, current approaches to characterise interdisciplinary research from a bibliometric perspective remain contentious. Like any other methodology suggested so far to measure and characterise interdisciplinarity based on scientific publications, our approach is not free of limitations and therefore results of these analyses need to be interpreted with caution.