CWTS BIBLIOMETRIC REPORT

Meaningful metrics

Bibliometric study of the ETH Domain (2009-2020/2021)

August 22, 2022

Universiteit Leiden

Bibliometric study of the ETH Domain (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 A.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 2.1 (p. 15) and Annex C).

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 2.1).

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 institution'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 results of a concise bibliometric analysis of the scientific output of the ETH Domain, covering the period 2009-2020, with 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). For the first time conference proceedings (as far covered by the WoS database) are included in the study.

Although most of the methodology is similar to the study performed four years ago for ETH Domain, 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 A) of this report. In Section 3 the results of our analysis and interpretations are reported.

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

1. Section 3.1: Overall output and impact
2. Section 3.2: Trends
3. Section 3.3: Collaboration and partners

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

For this report, we used the publication data provided by the individual ETH Domain institutions. Publications were deduplicated at the level of the ETH Domain.

ETH Domain institutions provided us with a list of publications from their repositories. The publications were matched with the data in the CWTS version of the Web of Science (WoS) database.

2.1 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.

2.1.1 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.

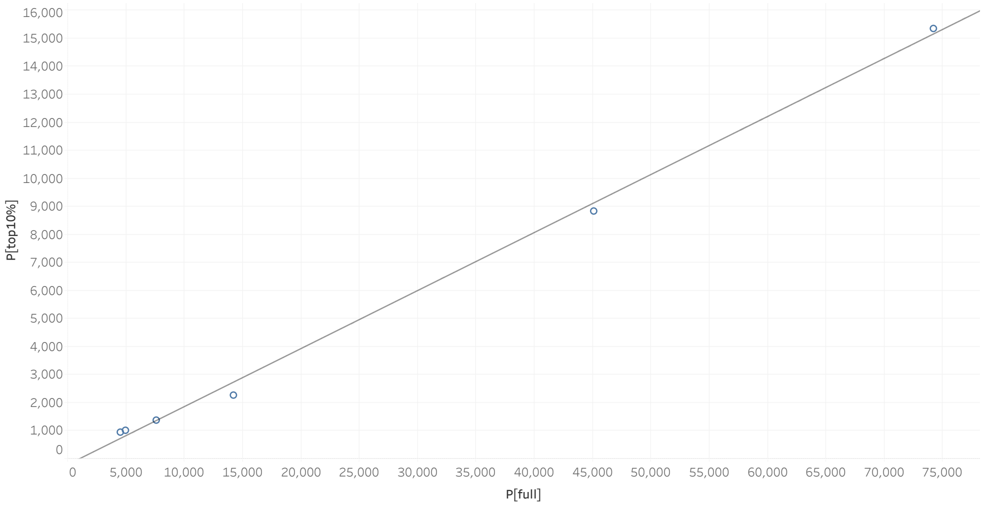


Figure 1: P[full]vs.P[top10%]for 6 ETH institutions

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.

2.1.2 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.

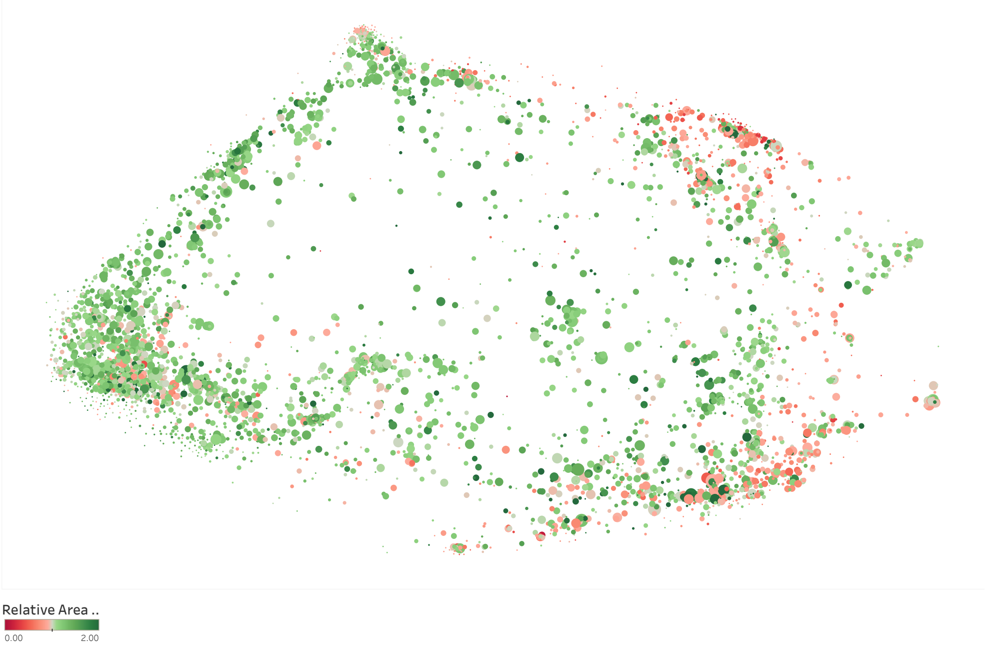


Figure 2: Landscape of all science, color-coded by [Area Growth]

[Field growth]

We use the [Area Growth] to characterise the fields in which ETH Domain researchers are active. Thus we contribute to the answer to the question: is ETH Domain'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 C.

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 ETH Domain authors were involved, there is a share of female ETH Domain 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 ETH Domain 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).

3 Results

In this chapter we discuss the performance of the ETH Domain over the entire period 2009-2020 and in a trend analysis of overlapping 4 year blocks. We discuss the output and impact and collaboration, as well as some indicators relating to the context in which the research is executed, such as gender diversity and Open Access publishing.

3.1 Overall output and impact

Main findings

> ETH Domain researchers were involved in 136,535 WoS publications, which is estimated at 79% of the total scientific output. Almost 60% is published in Open Access. Almost 80% of the publications is co-authored with other organisations, while 67% involves international collaboration. 9% is co-authored with the private sector.
>
> The impact of ETH Domain output is well above the world average, 64% by MNCS and almost twice the world average by PP[top10%].
>
> The share of female authors at the ETH Domain is 9% higher than the benchmark (the co-authoring partners).
>
> The ETH Domain research has a broad variety and covers basically the entire landscape of science. The research foci of the six individual institutions show some mutual overlap but merely points to complementarity.

In this section, we discuss the overall performance of the ETH Domain in the period 2009 up to 2020. It should be noted that these results are often heavily biased towards the larger institutions (ETH Zurich and EPFL), especially for the size dependent indicators.

Nevertheless, the results should provide a proper general overview of the bibliometric performance of the ETH Domain at large. This section contains the overall statistics as well as a positioning of ETH Domain research in the landscape of all science. By providing such positioning of all six institutions next to each other, we visualise their overlap and complementarity.

It should be noted that the provided overview (covering the entire period 2009-2020) allows only little opportunity for interpretation or contextualisation. We discuss an analysis over time of the indicators reported in this section, in the next section (see section 3.2). Thus, we provide better insight in the (intended or not) developments.

In Table 1, we list the ETH Domain scores for 5 types of indicators.

Table 1: Overall bibliometric performance statistics ETH Domain

Indicator	Score
Output	
P[full]	136,535
P[fract]	64,052
Int Cov	0.79
InterDisc	0.35
P OA [Gold, Hybrid, Green]	72,007
PP [OA]	59%
Collaboration	
PP[collab]	79%
PP[industry]	9%
PP[int collab]	67%
Citedness	
TCS	1,633,206
MCS	11.96
P[top10%]	26,973
PP[top10%]	19%
MNCS	1.64
MNJS	1.48
PP[self cits]	26%
PP[uncited]	16%
Author gender	
A[F inst]	50,324
A[FM inst]	251,339
PA[F inst]	0.20
PA[F pubs]	0.18
RPA[F]*	1.09
Readership	
N reads	350,985
N pubs read	68,966
Avg Reads	5.09

* RPA[F] may differ from the ratio PA[F inst] to PA[F pubs] due to rounding.

Output

Researchers of the ETH Domain were involved in 136,535 WoS publications from 2009 up to 2020 (P[full]). Normalised by the number of co-authoring institutions, the output adds up to 64,052 (P[fract]). We estimate that almost 80% of the output is covered by WoS (IntCov: 0.79%). The interdisciplinarity (IntDisc) of research at the ETH Domain is estimated at 0.35. 60% of the output (P[OA]: 72,007) in the entire period of 12 years was published in OA.

Collaboration

We found for the ETH Domain that almost 80% of ETH Domain's publications involved collaboration (i.e. co-authored by more than one institutions). 67% involved international collaboration (PP[int collab]). Furthermore, 9% of the output involved a private partner (PP[industry]).

Citedness

Publications by ETH Domain researchers were cited more than 1,6 million times (TCS), which means almost 12 times on average (MCS). The impact of ETH Domain publications, when normalised by research area and year, reaches an MNCS of 1.64, which means 64% above the world average of 1.

An important contribution to this high impact is due to the large number of publications belonging to the 10% most cited worldwide (P[top10%]: 26,973). The proportion of ETH Domain output belonging to the top 10% most cited publications (PP[top10%]) is 19%, which means almost twice the worldwide average value of 10%. We calculated 16% of the output not being cited. Finally, we found 26% of the citations to be author self-citations (PP[self cits]), which were not considered for the impact measurement.

Author gender

We found that 50,324 of the 251,339 of the ETH Domain author names (A[FM inst]) are female, which represents 20% (PA[F inst]: 0.20). This share is 18% (PA[F pubs]: 0.18) for all co-authors of the publications in which ETH Domain researchers were involved (the benchmark). The ratio (RPA[F]: 1.09) indicates a slightly higher involvement of female authors at the ETH Domain as compared to the benchmark.

Readership

We counted 68,966 ETH Domain publications to be saved by 350,985 Mendeley users, which is 5.09 on average per publication, stored in Mendeley (Avg Reads).

ETH Domain research focus

To provide more (general) insight in the foci of ETH Domain's research, we plotted the output across the landscape of science (Figure 3).

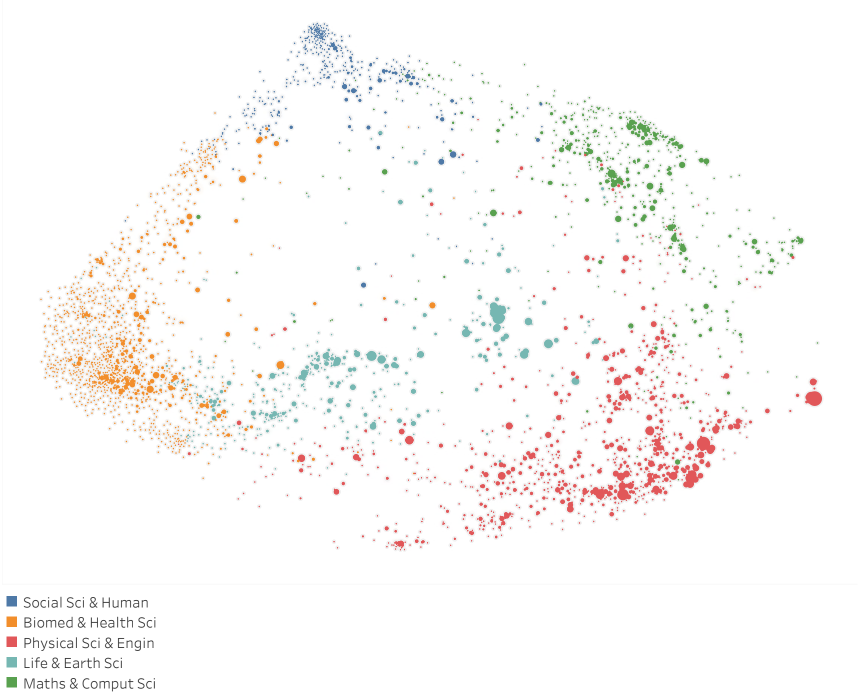


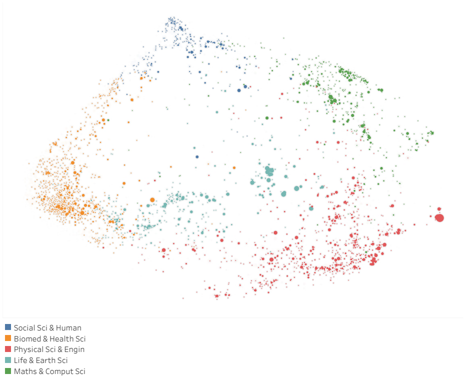
Figure 3: Distribution of ETH Domain's output across landscape of science (interactive version via this link)

The landscape in Figure 3 is a two-dimensional representation of all science (covered by WoS) with an overlay of the output of the six institution of the ETH Domain together 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 ETH Domain researchers were involved. The colors in the landscape point to 5 main disciplines we use to support the interpretation of the landscape. The map of the landscape shows a broad distribution of ETH Domain's research.

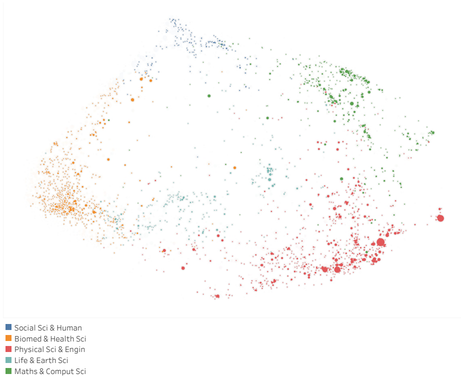
In addition to that, we plotted the output across the landscape of the 6 ETH Domain institutions (Table 2). These landscapes underlay the ETH Domain landscape and show partly the complementarity and overlap of research foci of the six institutes. ETH Zurich and EPFL cover the entire map, being the most general universities. WSL and Eawag show primary interest in life and earth sciences, while Empa and PSI focus primarily on physical sciences and engineering. These landscapes can be explored interactively via this interface. Open the menu on the left to change the perspective to an institution.

Table 2: Landscapes of all ETH Domain institutions (open in browser)

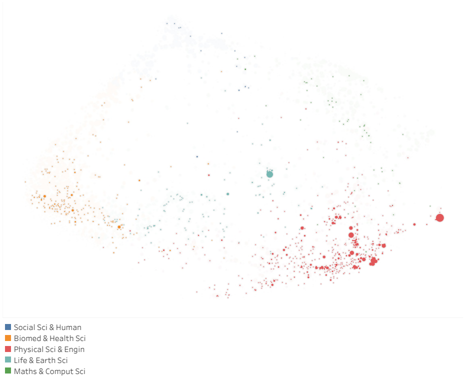
ETH Zurich



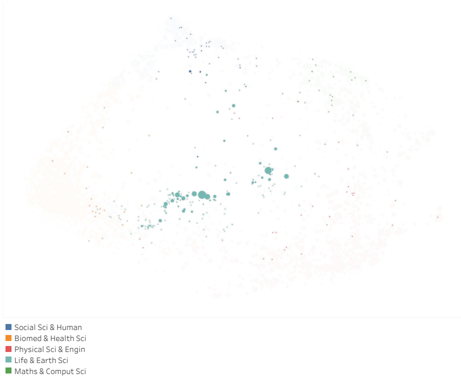
EPFL



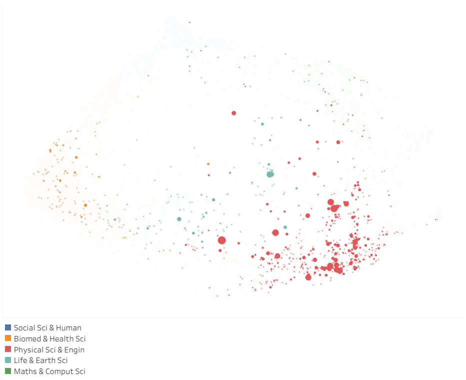
PSI



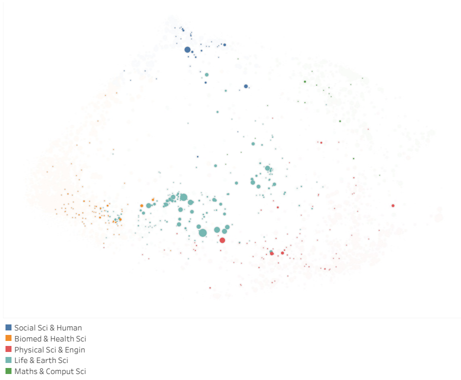
WSL



Empa



Eawag



3.2 Trends

Main findings

ETH Domain researchers were involved in an increasing number of publications, although it seems to saturate somewhat in the most recent period, which relates to the delayed processing and uptake of proceedings in WoS. We found an increasing number and share of OA publications, publications, involving international collaboration and publications co-authored with industry. The impact if the output remains at a high level throughout. Another striking increase is the share of female authors in ETH Domain publications. Finally, we measured an increasing involvement of ETH Domain researchers in growing aeas.

3.2.1 General statistics

In this section we discuss the trend for key indicators related to the performance of the ETH Domain at large. By looking at trends we provide a sense of how the ETH Domain research has developed between 2009 and 2020.

Note that many of the indicators and trends depicted in Table 1 are biased towards the bigger institutions ETH Zurich and EPFL. In other words, these institutions disproportionately influence the trends observed for the ETH Domain at large, especially for size-dependent indicators.

Table 3 shows a steady increase of output (P[full]) in which ETH Domain researchers were involved (from 37,017 up to 52,000, more than 40% increase). This is also the case for the contribution of the ETH Domain (P[fract]), but the increase is relatively less (from 19,630 up to 21,909, 15% increase). This means that researchers of the ETH Domain were involved in larger teams.

Another important trend we can see in these results is the increase of the number of Open Access publications (P[OA]). Relative to the total output in which ETH Domain researchers were involved (PP[OA]), we see a significant positive trend as well, from 49% up to 68%.

Table 3: Trends of ETH Domain's bibliometric performance

Indicator	2009–2012	2010–2013	2011–2014	2012–2015	2013–2016	2014–2017	2015–2018	2016–2019	2017–2020
P[full]	37,017	39,352	42,021	44,875	47,518	49,535	51,092	52,377	52,000
P[fract]	19,630	20,408	21,279	22,127	22,514	22,638	22,664	22,605	21,909
InterDisc	0.33	0.33	0.34	0.34	0.35	0.35	0.36	0.37	0.37
P [OA]	15,671	17,363	19,176	21,344	23,810	26,230	28,677	31,101	32,526
PP [OA]	49%	51%	52%	54%	57%	59%	62%	65%	68%
PP[collab]	72%	73%	75%	76%	78%	80%	81%	83%	83%
PP[int collab]	60%	61%	62%	64%	66%	68%	70%	71%	72%
PP[industry]	8%	8%	8%	8%	9%	10%	10%	10%	10%
P[top10%]	7,295	7,868	8,456	9,129	9,528	9,785	10,142	10,224	10,150
PP[top10%]	19%	19%	19%	19%	19%	19%	19%	18%	18%
MNCS	1.67	1.68	1.67	1.69	1.66	1.64	1.64	1.61	1.59
PA[F inst]	0.18	0.19	0.19	0.20	0.20	0.20	0.21	0.21	0.22
RPA[F]	1.04	1.10	1.12	1.12	1.11	1.11	1.11	1.11	1.12

The share of output in which ETH Domain researchers collaborated with researchers from other institutions (PP[collab]) increased with 11 percent points (from 72% up to 83%), while the share of co-authored publications with foreign partners (PP[int collab]) increased from 60% up to 72%. The share of output in which industry was involved (PP[industry]) increased as well (8% up to 10%). These figures point to an increased international integration together with an increased involvement of industry with ETH Domain's research.

The measure of interdisciplinarity (IntDisc) increased somewhat over time from 0.33 up to 0.37. It is difficult to say whether this is significant, as research is becoming more interdisciplinary in general. We interpret this trend as an indication that research performed at the ETH Domain is increasingly on a broader, more diverse basis.

Regarding the impact we found that on the one hand, the number of publications that are in the Top 10% cited worldwide (P[top10%]) increased along with the total output. Hence, the proportion of publications in the top 10% (PP[top10%]) remains stable over the time period observed. On the other hand the MNCS score slightly decreases.

The fact MNCS and PP[top10%] don't show the same trend may be explained by a few very highly cited publications in the first years of our analysis. These scores influence the MNCS score but not so much the PP[top10%]. The latter indicates that research at the ETH Domain has a high impact throughout (above 18%).

3.2.2 Open Access publishing

Open Access publishing is a major element in the context of executing open science. As shown in Table 3, the number and share of Open Access (OA) publications increases significantly during the period we studied. In this section we will look at this in more detail. First of all, we look at the impact indicators of OA and non-OA publications. In Table 4, we present four indicators by type (Open or Closed Access): P[full], P[top10%], PP[top10%] and PP[int collab].

P[full] for Closed Access publications drops from 2017 onwards, while the number of OA publications doubles during the period we studied. In Figure 4 the increase of all three OA types is visualised. Particularly, Gold and Hybrid publications show a strong increase. Green OA publications may be at a saturation point.

Table 4: ETH Domain'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]	16,615	16,986	17,454	17,919	18,021	17,943	17,471	16,562	15,623
	P[top10%]	3,074	3,155	3,214	3,235	3,151	3,018	2,832	2,633	2,489
	PP[top10%]	18%	18%	18%	18%	17%	17%	16%	16%	16%
	PP[int collab]	57%	58%	60%	60%	62%	64%	65%	67%	68%
Open	P[full]	15,671	17,363	19,176	21,344	23,810	26,230	28,677	31,101	32,526
	P[top10%]	3,442	3,837	4,290	4,866	5,382	5,854	6,444	6,801	6,992
	PP[top10%]	20%	20%	21%	21%	21%	21%	21%	21%	20%
	PP[int collab]	68%	69%	70%	71%	73%	74%	75%	76%	76%

We see a higher impact for OA publications (PP[top10%]) during the entire period. The impact of Closed Access publications (PP[top10%]) drops somewhat, together with the absolute number of top 10% publications in Closed Access (P[top10%]). We also included the PP[int collab] in this table, because international co-authored papers contribute a lot to scientific (citation-based) impact. For both open and Closed Access, the share of publications involving international collaboration (PP[int collab]) increases. The share is always higher for OA publications.

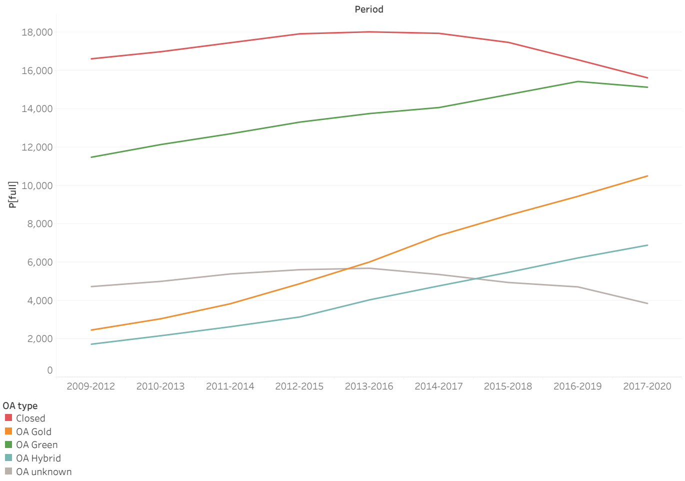


Figure 4: ETH Domain's output trend by Open Access (OA) type

3.2.3 Author gender diversity

In this section we look at gender diversity in publications by calculating the share of female authors.

We estimated that 20% of the ETH Domain authors are female (PA[F inst]: 0.20). In comparison, women amount to 18% of all authors listed on the output published by the ETH Domain (PA[F pubs]: 0.18, i.e. authors from the ETH Domain and co-authors from other institutions altogether). This means that the share of female authors participating to publications is higher within the ETH Domain as compared to co-authoring institutions (RPA[F]: 1.09).

In Figure 5, we depict the trend of female authors (PA[F inst], blue line) and the share of female authors at ETH Domain compared to all co-authoring institutions (RPA[F]) (red line) over time. Looking at these results, we see a steady increase of the share of female ETH Domain authors and a stable 9% above the benchmark from 2011 onwards.

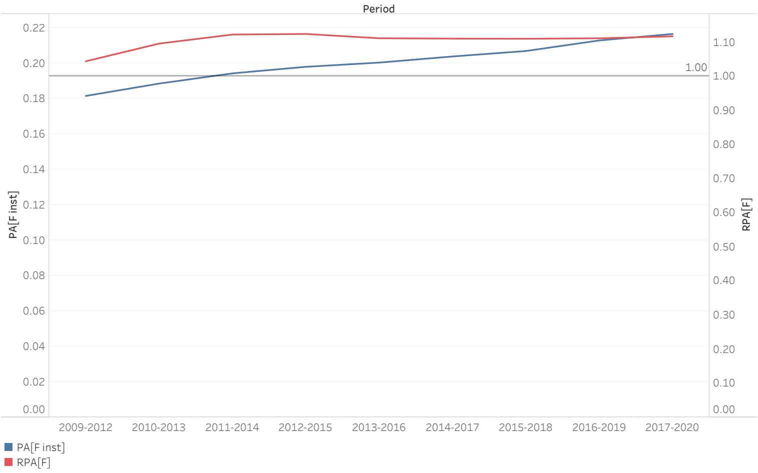


Figure 5: Share of female ETH Domain authors (PA[F inst]), and share of female authors compared to benchmark (RPA[F])

3.2.4 Output in the context of developments worldwide

The final part of results in these trend sections relate to the growth of fields in which researchers at the ETH Domain are active. For this, we combine the activity of the ETH Domain as distributed on the landscape of science, and the growth of the areas in that landscape worldwide.

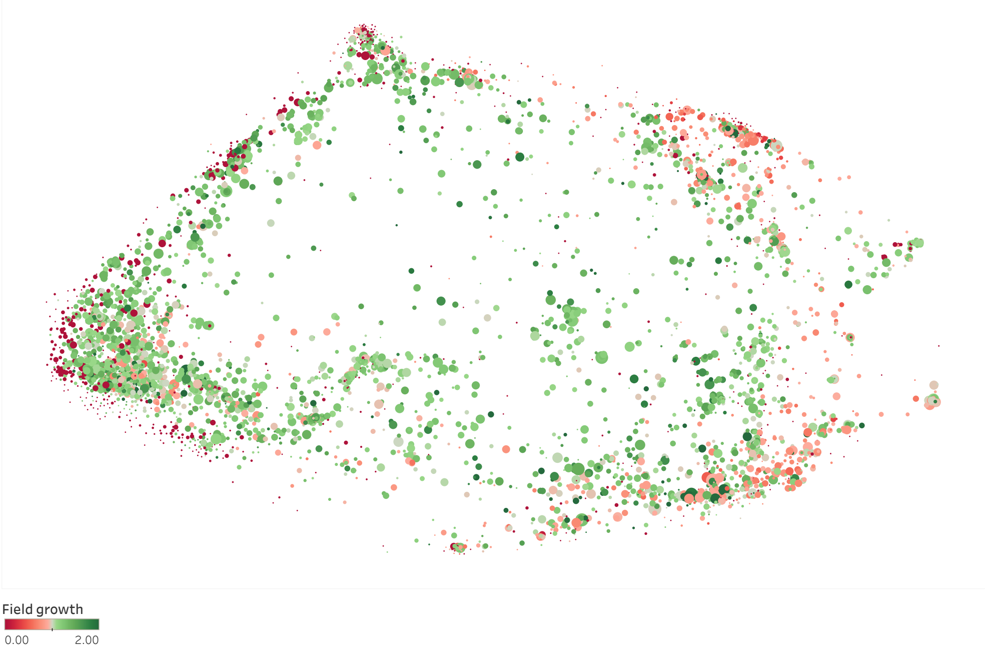


Figure 6: Positioning of ETH Domain research in landscape of all science, color-coded by [Area Growth]

In Figure 6, we depict the distribution of ETH Domain research output (similar to Figure 3) and color coded-each areas by the estimated growth worldwide ([Area Growth]). This map positions ETH Domain's activity and relates it to the developments worldwide. We can see that the vast majority of areas in which ETH Domain researchers publish, is growing (green). Besides that, there are regions in the landscape with substantial ETH Domain output, and stable (grey) activity worldwide or somewhat negative growth in volume (red). In these areas knowledge production has saturated. Worldwide the attention has shifted towards other areas.

Subsequently, we processed this information in a trend analysis, in which we linked publications per 4 years period to the recent growth factor of the area to which they belong ([Area Growth]). The results are plotted in Figure 7. The blue line plots the number of publications per 4-years period (P[full]), while the red line reflects the estimated volume growth of the areas to which the publications in each period belong ([Field growth]). The latter point out that ETH Domain researchers published in growing research areas. Moreover, we found that the number of ETH Domain publications not only increases over the years (P[full], blue line), but also that the research is increasingly published in growing areas (red line).

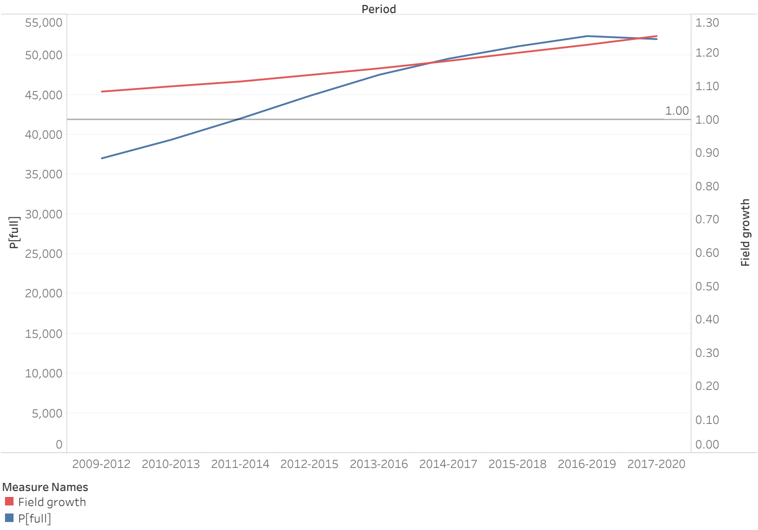


Figure 7: ETH Domain number of publications over time and estimated volume growth of subjects in which ETH Domain researchers are active

3.3 Collaboration and partners

Main findings

To assess collaborative work of ETH Domain researchers, we look at collaborative output within the ETH Domain and output outside the ETH Domain. Within the ETH Domain, we found large differences in number of co-publications, due to the different sizes. The impact is the highest for publications involving ETH Zurich and EPFL. Co-authorship analysis outside the ETH Domain shows a prominent position of other Swiss and German institutions. In terms of impact, co-authorship with US institutions stand out, besides a set of European institutions.

Regarding collaboration, we consider co-authorship and impact within and outside the ETH Domain. Output numbers (co-publications) are size-dependent, while impact measures (MNCS) are size independent. We discuss them as such.

Collaboration within the ETH Domain

The scores in Table 5 show the co-authorship (output and impact) between the institutions of the ETH Domain. The absolute numbers of output are clearly dominated by ETH Zurich and EPFL, but that does not mean that the number of co-authored publications between these two are the highest. In Table 2, we visualised a similar research focus of ETH Zurich and EPFL. In general, we see there is collaboration among all members, with only a few pairs with less than 100 co-publications: Eawag with PSI and WSL with Empa, which can be explained by the different profiles. Another less productive co-authorship between is found (in absolute number) between WSL and Eawag, which can be explained by the lower output overall by these institutions.

It should also be noted that the highest impact is found for the publications in which ETH Zurich and EPFL collaborate (MNCS: 2.02). We also found high impact for co-authored output by ETH Zurich and WSL (MNCS: 1.76).

Table 5: Co-authorship and impact within the ETH Domain

Unit	Indicator	ETH Zurich	EPFL	PSI	WSL	Empa	Eawag
ETH Zurich	P[full]	74,190	1,894	4,294	1,107	2,264	1,832
	MNCS	1.71	2.02	1.54	1.76	1.57	1.54
EPFL	P[full]	1,894	45,073	1,279	390	591	528
	MNCS	2.02	1.63	1.45	1.46	1.42	1.58
PSI	P[full]	4,294	1,279	14,191	125	512	27
	MNCS	1.54	1.45	1.34	1.47	1.64	1.50
WSL	P[full]	1,107	390	125	4,936	20	65
	MNCS	1.76	1.46	1.47	1.42	1.06	1.60
Empa	P[full]	2,264	591	512	20	7,575	121
	MNCS	1.57	1.42	1.64	1.06	1.44	1.54
Eawag	P[full]	1,832	528	27	65	121	4,497
	MNCS	1.54	1.58	1.50	1.60	1.54	1.62

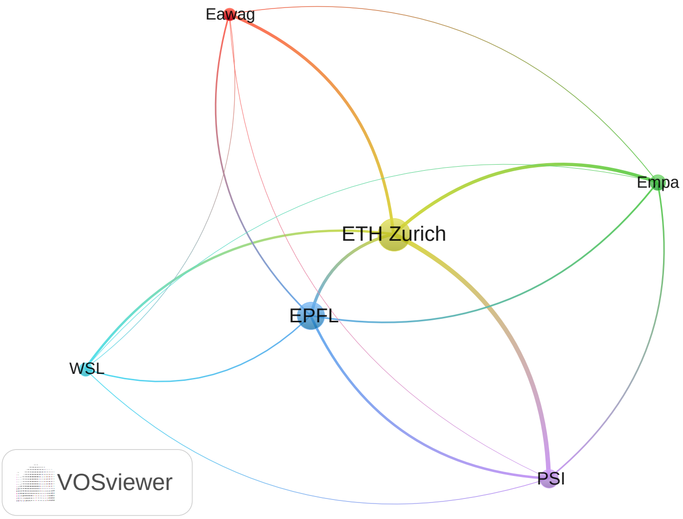


Figure 8: Co-authorship network of ETH Domain's institutions (line width reflects number of co-publications, node size reflects total output)

In Figure 8 we visualise the co-author network of the six ETH Domain institutions, taking all connections into account. This network shows the central position of ETH Zurich and EPFL, with the other institutions taking their own position around them. As all connections are considered, the graph positions ETH Zurich and EPFL close to each other in the center.

Collaboration outside the ETH Domain

The results in Table 6 and 7 show the 40 most prominent partners of the ETH Domain at large and distributed by institution, in terms of number of co-publications. In these results ETH Domain internal collaborations were not considered.

Table 6: Top 40 collaborating ETH Domain's institutions, outside the ETH Domain only (fractional output and impact)

Inst	Country	Co-pubs	MNCS
Univ Zurich	CH	3,129	1.61
Max Planck Soc Advance Sci	DE	1,138	2.08
Univ Lausanne	CH	992	1.57
Univ Bern	CH	898	1.63
Univ Geneva	CH	825	1.48
Ctr Natl Rechr Sci	FR	685	1.82
Univ Basel	CH	637	1.69
Chinese Academy of Sciences	CN	446	1.96
Massachusetts Inst Technol	US	441	2.29
Univ California - Berkeley	US	379	2.44
Tech Univ Munich	DE	373	1.81
Karlsruhe Inst Technol	DE	372	1.86
Harvard Univ	US	341	2.17
Univ Oxford	GB	323	2.32
Univ Cambridge	GB	317	2.02
Russian Academy of Science	RU	311	1.39
Ist Nazl Fis Nuclr	IT	299	1.54
Stanford Univ	US	298	2.48
Katholieke Univ Leuven	BE	296	2.43
Politec Milano	IT	291	1.47
Univ Bologna	IT	283	1.62
California Inst Technol	US	280	2.18
Spanish Natl Res Cncl (CSIC)	ES	264	1.79
CERN Europe Org Nuclr Res	CH	256	1.53
Cons Nazl Ricrc	IT	254	1.34
Delft Univ Technol	NL	251	2.15
Univ Fribourg	CH	237	1.52
Tech Univ Denmark	DK	231	1.94
Univ Freiburg	DE	226	1.93
Agroscope	CH	218	1.36
Princeton Univ	US	208	2.76
Ludwig-Maximilians Univ München	DE	207	1.92
Univ Padova	IT	196	1.66
Universidade Lisboa	PT	194	1.53
Tech Univ Dresden	DE	187	1.92
Heidelberg Univ	DE	185	1.77
Univ Tokyo	JP	182	1.80
Univ Neuchâtel	CH	182	1.55
RWTH Aachen Univ	DE	182	1.87
Imperial Coll London	GB	182	2.25

The list shows the prominent position of the University of Zurich and the Max Planck Society as well as three other Swiss institutions with more than 800 co-publications. Publications with almost all top 40 partners achieve a high impact by MNCS between 1.50 and almost 2.50.

Looking at the distribution across ETH institutions (Table 7), we see that the most productive partnerships are primarily by ETH Zurich, EPFL and PSI. This relates obviously to the size of these institutions. With this list we provide an overview of key partners with the ETH Domain, and refer to the institutions' reports for more detail.

Table 7: Number of co-authored publications (fractional counting) with top 40 collaborators by ETH Domain institution

Institution	Country	ETH Zurich	EPFL	PSI	WSL	Empa	Eawag	Total
Univ Zurich	CH	2,750	137	177	93	101	75	3,129
Max Planck Soc Advance Sci	DE	572	420	135	11	56	20	1,138
Univ Lausanne	CH	201	757	19	45	3	21	992
Univ Bern	CH	470	130	166	76	57	122	898
Univ Geneva	CH	233	532	56	9	16	31	825
Ctr Natl Rechr Sci	FR	308	282	85	25	28	14	685
Univ Basel	CH	408	92	82	25	53	31	637
Chinese Academy of Sciences	CN	143	120	123	56	14	19	446
Massachusetts Inst Technol	US	273	155	17	4	10	4	441
Univ California - Berkeley	US	189	150	35	12	7	10	379
Tech Univ Munich	DE	183	96	70	28	15	8	373
Karlsruhe Inst Technol	DE	173	96	94	10	18	8	372
Harvard Univ	US	191	139	14	3	4	4	341
Univ Oxford	GB	205	81	39	4	6	5	323
Univ Cambridge	GB	170	104	24	16	14	3	317
Russian Academy of Science	RU	119	142	61	10	4	12	311
Ist Nazl Fis Nuclr	IT	137	116	128	0	1	0	299
Stanford Univ	US	179	109	13	1	2	6	298
Katholieke Univ Leuven	BE	220	56	15	2	37	9	296
Politec Milano	IT	134	131	28	1	13	2	291
Univ Bologna	IT	201	78	9	2	5	0	283
California Inst Technol	US	213	58	13	3	3	3	280
Spanish Natl Res Cncl (CSIC)	ES	138	54	39	15	23	16	264
CERN Europe Org Nuclr Res	CH	57	163	52	1	1	0	256
Cons Nazl Ricrc	IT	84	119	38	10	17	3	254
Delft Univ Technol	NL	105	105	18	1	18	17	251
Univ Fribourg	CH	100	78	48	14	30	2	237
Tech Univ Denmark	DK	119	56	42	4	17	14	231
Univ Freiburg	DE	123	41	19	40	20	1	226
Agroscope	CH	183	13	1	23	10	13	218
Princeton Univ	US	121	72	20	3	2	2	208
LM Univ München	DE	155	34	15	5	5	4	207
Univ Padova	IT	75	95	13	14	4	9	196
Universidade Lisboa	PT	57	117	6	8	2	8	194
Tech Univ Dresden	DE	108	25	44	7	13	6	187
Heidelberg Univ	DE	131	43	13	4	2	3	185
Univ Tokyo	JP	102	63	28	0	3	3	182
Univ Neuchâtel	CH	69	89	2	31	2	20	182
RWTH Aachen Univ	DE	116	45	25	0	7	13	182
Imperial Coll London	GB	108	54	14	1	11	4	182

In Table 8, we list the most prominent countries collaborating with the ETH Domain. In 20% of the publications, researchers from the united States. In 19% of the output researchers from Switzerland (outside the ETH Domain) are co-authors. Together with Germany (18%) these are the most important co-authoring countries for the ETH Domain.

Table 8: Top 12 countries co-authoring with ETH Domain researchers, excluding ETH Domain internal co-authorship. P[full] and % to ETH Domain's total

Country	Co-pubs	% to total
United States	27,019	20%
Switzerland	26,576	19%
Germany	24,706	18%
United Kingdom	16,113	12%
France	15,403	11%
Italy	12,392	9%
Spain	8,344	6%
China	8,139	6%
Netherlands	6,992	5%
Austria	5,502	4%
Belgium	5,043	4%
Japan	4,929	4%

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.

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.

A Citation data and analysis

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

A.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%).

A.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.

A.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.

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 9), 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.

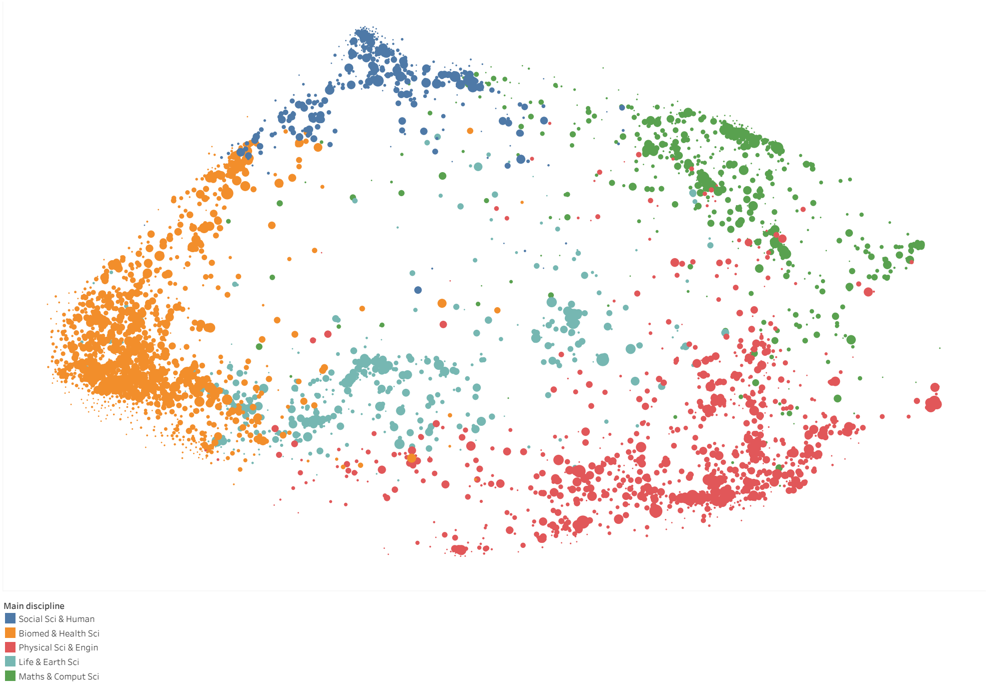


Figure 9: Landscape of all science (around 30 million WoS publications). Circles represent (over 4,000) publication clusters. Position is defined by citation traffic between clusters. Size indicates relative volume. Color reflects 5 main disciplines

C 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.