universitätbonn

Rheinische Friedrich-Wilhelms-Universität Bonn

fiw N50° 43' E7° 7'

Forum Internationale Wissenschaft

FIW WORKING PAPER NO. 08

Jason Chumtong | David Kaldewey

# BEYOND THE GOOGLE NGRAM VIEWER:

## BIBLIOGRAPHIC DATABASES AND JOURNAL ARCHIVES AS TOOLS FOR THE QUANTITATIVE ANALYSIS OF SCIENTIFIC AND META-SCIENTIFIC CONCEPTS

August 2017

# BEYOND THE GOOGLE NGRAM VIEWER:

## BIBLIOGRAPHIC DATABASES AND JOURNAL ARCHIVES AS TOOLS FOR THE QUANTITATIVE ANALYSIS OF SCIENTIFIC AND META-SCIENTIFIC CONCEPTS

JASON CHUMTONG
DAVID KALDEWEY

FIW WORKING PAPER NO. 08

Bonn, 2017 | ISBN 978-3-946306-07-8

# ABSTRACT

In the last years the Google Ngram Viewer has become a popular tool for quick and dirty analyses of how certain concepts emerge and develop over time. So far most scholars using this tool do not reflect on its methodological problems and on whether the corpus presented by Google is actually appropriate for their specific research problem. Therefore, digital humanities experts tend to be both fascinated and sceptical about the Ngram Viewer. In this paper we do not take up the diverse lines of critique, but rather focus on what we deem to be the most important innovation coming with the Ngram Viewer: the very simple idea of tracing relative frequencies of words and phrases over time. We propose to take this basic idea one step further, particularly in regard to scientific and meta-scientific concepts. We show how bibliographic databases and journal archives can be used to analyze how often certain concepts appear in scientific publications. As these databases have not been established with any idea of linguistic analysis in mind, but rather as tools to find and access scientific publications, what we propose can be characterized as "off-label" use of bibliographic databases and journal archives.

# PUBLICATION DETAILS

Rheinische Friedrich-Wilhelms-Universität Bonn
Forum Internationale Wissenschaft
Heussallee 18-24
53113 Bonn

Tel.: +49 228 73 62986
Internet: www.fiw.uni-bonn.de/publikationen
E-Mail: fiw@uni-bonn.de
Layout: roemer und höhmann strategisches design
Satz: Jason Chumtong
ISBN 978-3-946306-07-8

# CONTENTS

# INTRODUCTION

In the last years the *Google Ngram Viewer* has become a popular tool for quick and dirty analyses of how certain concepts and phrases emerge and develop over time. So far most scholars using this tool do not reflect on its methodological problems and on whether the corpus presented by Google is actually appropriate for their specific research problem. Therefore, digital humanities experts tend to be both fascinated and sceptical about the *Ngram Viewer*. In this paper we do not take up the diverse lines of critique, but rather focus on what we deem to be the most important innovation coming with the *Ngram Viewer*: the very simple idea of tracing relative frequencies of words and phrases over time. As we agree with the critics problematizing the quality and relevance of the Google Books corpus, we propose a pragmatic solution, which at the same time takes the basic idea behind the *Ngram Viewer* one step further: we trace the historical development of n-grams in other kinds of corpora.



The starting point of our reflections about the *Ngram Viewer* was a concrete research problem in the history and sociology of science, namely, to trace the development and dissemination of scientific and meta-scientific concepts over time. In light of this specific interest we demonstrate that and how different kinds of databases and archives can be used to analyze how often certain concepts or phrases appear in scientific publications. This paper exemplifies this strategy with two different examples: first, the *Web of Science* search platform, and second, the archive of the *Science* journal. As these databases have not been established with any idea of linguistic analysis in mind, but rather as tools to find and access scientific publications, what we propose can be characterized as "off-label" use of bibliographic databases and journal archives.

# 1. GOOGLE BOOKS NGRAM VIEWER

The *Ngram Viewer* is a free accessible online tool for tracing the relative frequency of words or phrases during a specific time period. The *Ngram Viewer* calculates how often a certain n-gram **[1]** appears in the selected corpus of a given year, relative to the total number of n-grams. The corpus of digitized texts used by the *Ngram Viewer* primarily consists of a subset version of the Google Books collection. The first version of the database, introduced 2009, contained around five million books, the oldest published in the 1500s (Michel et al. 2011) **[2]**. In 2012 a second version was published that incorporated eight million books (Lin et al. 2012). Due to the wide scale of digital archived texts this corpora is not limited to specific genres. It includes all sorts of literature ranging from academic publications to biographies and novels. Though it is not possible to select these different document types as individual categories, the

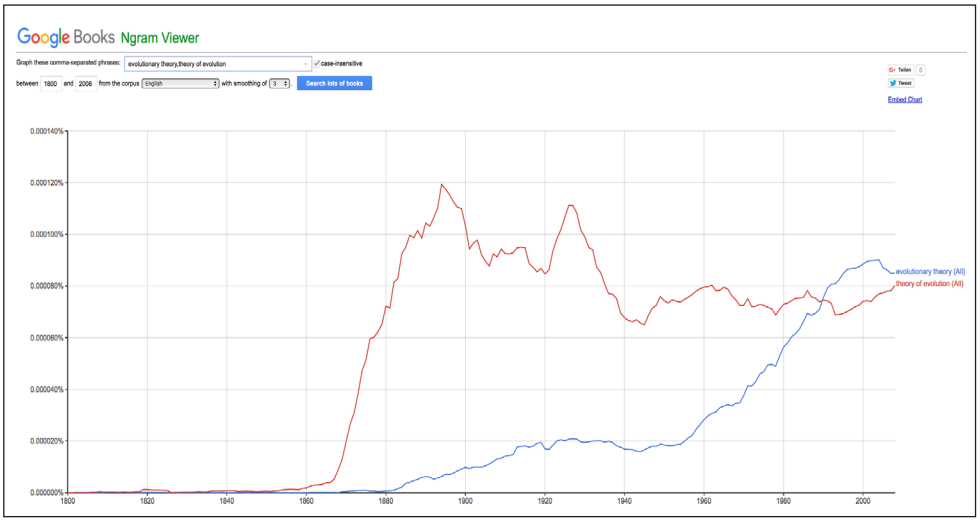
**[1]** N-grams are constant sequences of characters like words, abbreviations and also numbers.

**[2]** In 2004 Google started digitizing literature from over 40 university library holdings around the world. With a collection of up to fifteen million books, the corpus represents twelve percent of all books that were ever published (Michel et al. 2011).

*Ngram Viewer* does offer a differentiation by language. Subcorpora exist for eight languages, with the English corpus being the biggest, containing more than 350 billion words. The corpus covers a time span from 1500 until 2008. However, Michel et al. point out that search inquiries between 1800 and 2000 will deliver the highest data density and quality (culturomics.org) **[3]**. Compared to other big data sets created by various "digital humanties" projects, the *Ngram Viewer* enables fast and easy access to this pool of information without advanced technical knowledge.

Next to a regular search field for the term or phrase of interest, the online tool offers filtering options for the time span, the language, the degree of smoothing **[4]**, and a case insensitive option. It is also possible to search for more than one term or phrase for direct comparison. After a successful search the result is displayed in a diagram with graphs representing the frequency of how often the search terms or phrases are found inside the corpus. The given numbers do not show the concrete counts but the percentage ratio of the search hits compared to all n-grams in the respective year. In short, by using the *Ngram Viewer* interface, one needs only a few clicks to get a first impression of how the usage of certain n-grams developed over time. **Screenshot 1** illustrates this by a figure that points to how the phrase "evolutionary theory" has developed historically compared to the phrase "theory of evolution" (see also Appendix A). Additionally, the linkage with Google Books makes it possible to select certain time periods and manually track which publications during what year are marked in connection to the search.






**Screenshot 1**

**[3]** On the culturomics.org webpage the authors explain that the *Ngram Viewer* corpus does not contain enough books published before 1800 to reliably quantify a respective search, and the corpus after 2000 retains fine changes in regard to the time of the inception of the Google Books project.

**[4]** The smoothing filter effects how the graphs of the search result are displayed. Instead of integrating the exact amount of found items for a single year into the graph, the *Ngram Viewer* adds up the result of several years and uses the calculated average of them. In this way bold deflections that my appear due to significant differences between single years are smoothed to curves.

## 1.1. Working with the Google Books Ngram Viewer

The *Google Ngram Viewer* does not make the search result available for further processing. Even though it is possible to download the raw data, this option only addresses large scale analyses which require technical resources and advanced know-how in computer science **[5]**. However, there is a pragmatic way of extracting data from the HTML source code. In the source code one can directly access the actual data of the latest search:

```
var data = [{„ngram“: „evolution (All)“, „type“: „CASE _ INSENSITIVE“,
„timeseries“: [2.0381485590359461e-06, 2.0233224523380499e-06,
1.9274381070448028e-06, 1.8357519490562273e-06, 1.9181746865167981e-06,
1.8272708423186908e-06, 1.8106076178676339e-06, 1.8216218182130563e-06,
1.7255423309896808e-06, 1.7438908445553482e-06, 1.8456256566220678e-06,[…]
5.3407723991464684e-07, 5.2806392432103171e-07, 5.2500348601824953e-07]}]
```

These numbers are what the *Ngram Viewer* uses to generate the graphs displayed on the screen. After having conducted a search resulting in a diagram displayed directly on the website, the website source code contains the numbers of the calculated percentage ratio between the term or phrase of interest compared to the corpus. Listed chronologically, each number represents the percentage ratio for a specific year. The data generated by specific inquiries can then be exported as a list and processed with alternative software packages, particularly with spreadsheet applications.



This paper uses the *Ngram Viewer* data as the basis for designing own diagrams. So far, many authors that use the *Ngram Viewer* as a research tool present direct screenshots in their papers (e.g., Klein 2013; Ophir 2016; Pettit 2016; Roth et. al 2017), but the functions of the online tool are restricted. For example, the display of compositions of n-grams cannot be combined with the case-insensitive option and the figures produced online are not formatted as printable figures. Therefore, we recommend to generally process the *Ngram Viewer* data with professional software.

## 1.2. Handling Guidelines

Several authors have problematized the Google *Ngram Viewer* corpus and raised doubts about it being representative for natural language and its development over time. There is, for example, a bias in the corpus towards scientific literature (Pechenick et al. 2015). If, as in our examples, an analysis aims at tracing scientific or meta-scientific concepts, this is not a problem. However, the self-stated goal of the Google team responsible for the project, namely, to trace cultural developments more generally, and to establish “culturomics” as a new research field (Michel et al. 2011), is to be handled with care.

In light of this critique we propose that what makes the *Ngram Viewer* a valuable research tool is not primarily its accuracy, but rather is potential for quick-and-dirty heuristic analysis. In view of how academics today use the *Ngram Viewer* in their everyday research and presentation practices, the usefulness of the tool is evident, even

**[5]** For further information: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

if the *Ngram Viewer* rarely yields something like strong data corroborating concrete hypotheses. In the following we take the idea behind the Google *Ngram Viewer* one step further. Or, more precisely, we propose to use the *Ngram Viewer* not as a ready-made tool, but basically as an idea that may guide analysis in regard to very different and more robust corpora. More concretely, we use bibliographic databases and journal archives as alternative corpora for the quantitative analysis of scientific and meta-scientific concepts. As these databases have not been established with any idea of "culturomics" or linguistic analysis in mind, but rather as tools to find and access scientific publications, what we propose can be characterized as "off-label" use.

# 2. WEB OF SCIENCE



In this section we present an example of how the idea of visualizing the appearance and trends of terms inside a specific corpus can be employed on the basis of bibliographic databases. We use the *Web of Science* platform to produce data and figures comparable to the *Google Ngram Viewer*. The "off label" methodology we propose, however, can in principle be applied to other bibliographic databases as well (e.g., Scopus, JStor).

The *Web of Science* search platform manages the prominent citation indexing service developed by the Institute for Scientific Information (ISI) in the 1960s, which was sold to the media and information group Thomson Reuters in 1992 and has recently been taken over by Clarivate Analytics. The platform does not include full texts of scientific publications, but manages over 50 million database entries with relevant information. Thus, indirectly, *Web of Science* functions as one of the worlds largest databases for scientific publications across all disciplines. Since all recorded publications are tagged and linked to each other via a variety of subject categories, the database helps discovering relationships between certain keywords and research areas, but also single articles or specific journals.

A search for the term "evolution" for instance opens up a list of all publications within the *Web of Science* database that are linked to this keyword through its title, topic, or abstract if provided. Depending on one's interest it is possible to filter the search result for additional categories like most cited article or paper, document types, research category or publication years. Due to a variety of filtering combinations users have full control of the criteria determining their search. Moreover, *Web of Science* offers the possibility to download the result list, thus allowing to use the data for own working processes. For this paper the latter is seen as a condition for visualizing historical trends of scientific and meta-scientific concepts within the field of scientific publications.

Comparable to the graphs generated by the *Google Ngram Viewer*, the downloadable data in *Web of Science* can be processed into diagrams showing the appearance of single words, as well as word combinations during a specific time period. By exporting

the search results from *Web of Science* into spreadsheet applications it is possible to display the searched terms in a fully customizable frame. Having the heuristic possibilities of the *Ngram Viewer* in mind, we approach the *Web of Science* data with a similar idea of visualizing the development of terms or phrases on the basis of their relative frequency inside the corpus. In contrast to the *Ngram Viewer*, however, the *Web of Science* corpus is not a full-text database that could be disentangled into n-grams. Instead, the basic unit of *Web of Science* are scientific publications. Each entry contains basic information such as title, author, publication name, year published, and is furthermore linked, among other categories, to research areas, subject categories, languages, and document types. For the semantic analysis intended here two field tags are relevant, namely the title [TI] and the topic [TS] field. The latter does not restrict search results to the title, but also includes keywords and abstracts, which are, however, only accessible for the more recent past – keywords and abstracts are indexed by *Web of Science* since 1991. The difference between "title" and "topic" search is crucial for the data analysis. A title search evidently does not actually include much quantities of text. In contrast, using the topic search, which includes keywords and abstracts in addition to the title, gives more text to analyze. However, as abstracts have been indexed only after 1991 the topic search is restricted to the more recent past. At first sight, and compared to the *Ngram Viewer*, this sounds disappointing. At second glance there are some advantages to this kind of data, because if a given term or phrase appears in the title or abstract of a scientific publication, its appearance is more likely to be significantly associated with the content of the article than any n-gram to be found in any part of a full text book repository. Framed as an instruction, the following examples will explain how to extract data from the *Web of Science* database in order to produce figures that visualize the relative frequencies of certain terms over time.



## 2.1. Working with Web of Science

Before starting a search, *Web of Science* demands to chose which set of databases shall be included in the search process. As the actual set of available databases depends, quite randomly, on what the respective institution has subscribed **[6]**, we strongly recommend to restrict analysis to the *Core Collection* – this is a database accessibly for most users, spanning a time horizon from 1945 to the present. As long as we are interested primarily in the English language and a cross-disciplinary perspective, this *Core Collection* produces reliable data and enables reproduction of the search results. Depending on the concrete research interests, however, the same kind of analysis could also be conducted with more specific databases, for example in regard to specific disciplines (medicine, life sciences) or alternative languages and national contexts. The integration of all databases in the same search procedure would blur a clear picture of how the entire *Web of Science* corpus developed over time. The *Core Collection* covers the longest time horizon and the widest range of research areas compared to the others, which makes it ideal for a use comparable to the *Ngram Viewer*.

The *Core Collection* consists of four different indexes, which have been introduced at different points in time and are not coextensive in regard to the time period they cover: the *Science Citation Index* (SCI) contains publications from 1945 up to the present, the

**[6]** For example, the University of Bonn has subscribed to the following databases accessible via Web of Science: the Russian Science Citation Index (bibliographic information from over 500 articles of Russian researchers) , the SciELO Citation Index (scholarly literature from open access journals published in Latin America, Portugal, Spain and South Africa), the MEDLINE (literature from the U.S. National Library of Medicine), the KCI-Korean Journal Database (multidisciplinary articles from the National Research Foundation of Korea) and the *Core Collection* (Web of Science's own database).

*Social Science Citation Index* (SSCI, 1956–present), the *Arts&Humanities Index* (A&HCI, 1975–present) and the Emerging Sources Citation Index (ESCI, 2015–present). Like their description already hints at, all four indexes characterize a compilation of scientific disciplines. Hence, the possibility to select them individually or in combination becomes a handy solution for adjusting the range of data the search should focus on. When doing so it is important to pay attention to the time period they cover, since the amount of documents added during this time period can have impact on the quality of the search result. **Figure 1** shows the impact of the different indexes on the composition of the *Web of ScienceCore Collection* in relation to the year they were introduced **[7]**. The same applies to the inclusion of abstracts. Documents which contain this additional category of information have a higher chance to get caught by the search request. Since an abstract adds more semantic data, the entries containing abstracts are more valuable for analysis than those which are only categorized by title, journal, publication year etc. However, abstracts and keywords are indexed by *Web of Science* only since 1991. Therefore, it does not make sense to conduct *topic* [TS] searches for the time span between 1945 and 1990. Whenever such a time horizon is necessary, the Advanced Search field tag has to be *title* [TI]. Using the field tag *topic* [TS] in contrast makes sense for the time span between 1991 and today. **Figure 2** shows the comparison between the *topic* [TS] and *title* [TI] search for the term “evolution”.



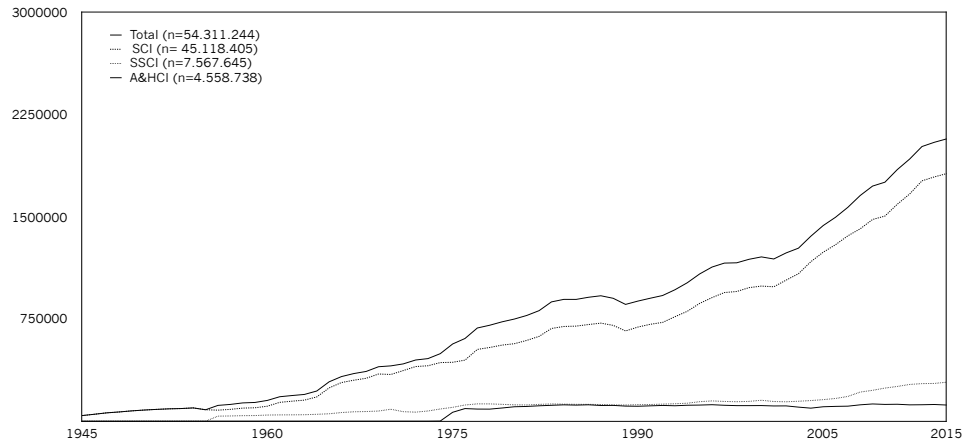



**Figure 1:** Total number of publications in the *Web of Science* core collection, 1945-2015, smoothing=3, data retrieval July 16, 2016.

*Web of Science* primarily serves as a citation tool, which is why some important information does not appear at first sight. When searching for a term the result list will show all relevant publications connected to the term of interest. Yet, the relation between these hits and the amount of documents being searched is not transparent. Without this ratio, transferring the actual numbers of the search into graphs will skew up the result, since trends cannot be interpreted sufficiently. Equal to the *Ngram Viewer* it is indispensable to normalize the number of found publications by calculating their relative frequency in regard to all accessible publications in a given year. Therefore it is a prerequisite to first determine the corpus that will serve as the starting point of the search and then extract its numbers to calculate the ratio.

**[7]** Due to the short time period the ESCI covers, it was excluded from this figure.

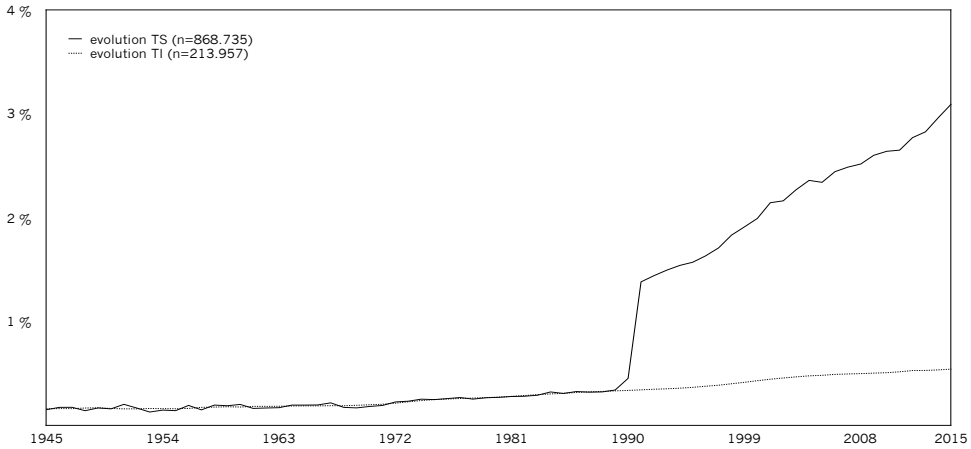



**Figure 2:** Web of Science Topic search [TS] vs. Title search [TI], exemplified by the term „evolution", 1945-2015, smoothing=3, data retrieval September 20, 2016.



## 2.2. Extracting data: Research areas as distinguishing variables

With the following example we illustrate how to determine a corpus in regard to a specific research interest and how to extract data within that corpus. We focus on research areas as a distinguishing variable in order to find out how the term "evolution" developed in the natural sciences as compared to the social sciences **[8]**. To do so, we use the two respective indexes, the SCI (roughly representing the natural sciences), and the SSCI (roughly representing the social sciences), to generate comparable figures.

Starting with a simple search request, we first narrow down all publications for the demanded time period inside the index of interest. To do so, we select the respective index under the *More Settings* options. We choose the SCI to build a corpus of natural sciences publications, and define the time span by typing `PY=1945-2015` in the *Advanced Search* field. Doing so leads to the result of 45.122.084 publications **[9]**, representing every single database entry the SCI tagged by the respective publication year, independently of any further affiliation **[Screenshot 2]**. From here on it is possible to start refining the result.

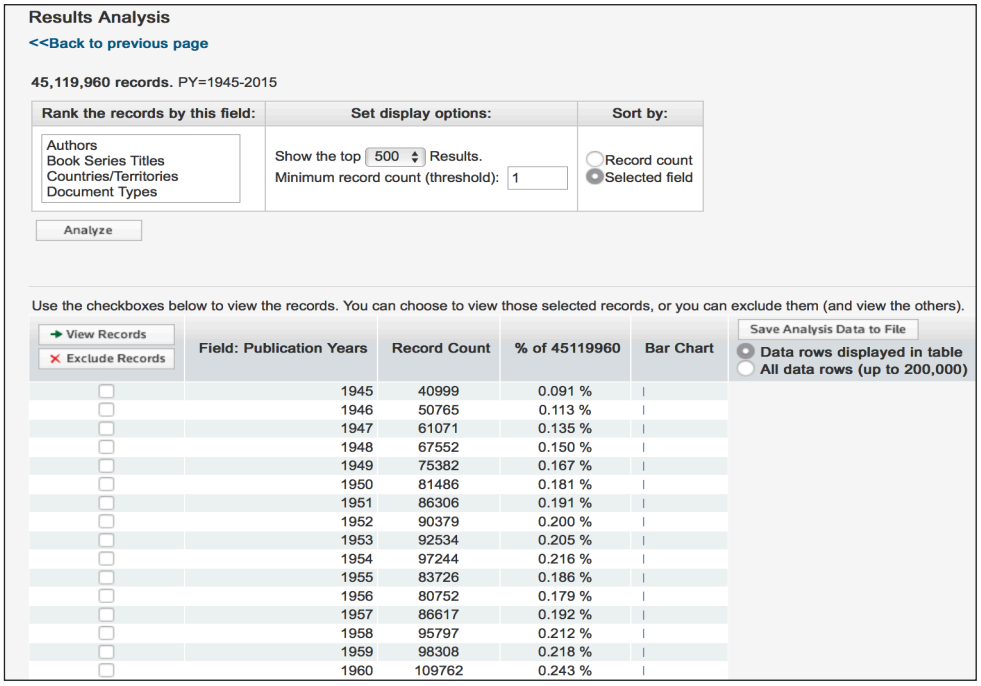
**[8]** We use the term „evolution" as example because we build on exploratory research and questions raised by Chumtong (2015).

**[9]** All data presented in this section were retrieved June 29, 2017. The database is regularly updated with new entries, therefore, the same search at a later point may increase the total corpus.


**Screenshot 2**

The function *Analyze Results* enables to sort the roughly 45 million publications by several categories. Two points are important for our purpose. First, the *Analyze Results* function is extremely valuable as its results are downloadable for further processing. Second, even though the category selection equals the one before, the *Analyze Results* function displays the search result differently from the standard display of results. Instead of showing single publications tagged with the selected category, it unfolds a full list of the record count regarding the selected category. Sorting the search results by publication years will show the record count for every single year, displaying how many publications were published for every particular year between 1945 and 2015 **[Screenshot 3]**. Downloading this data generates a text file with all the numbers we need to determine how the total number of publications are distributed over time.






**Screenshot 3**

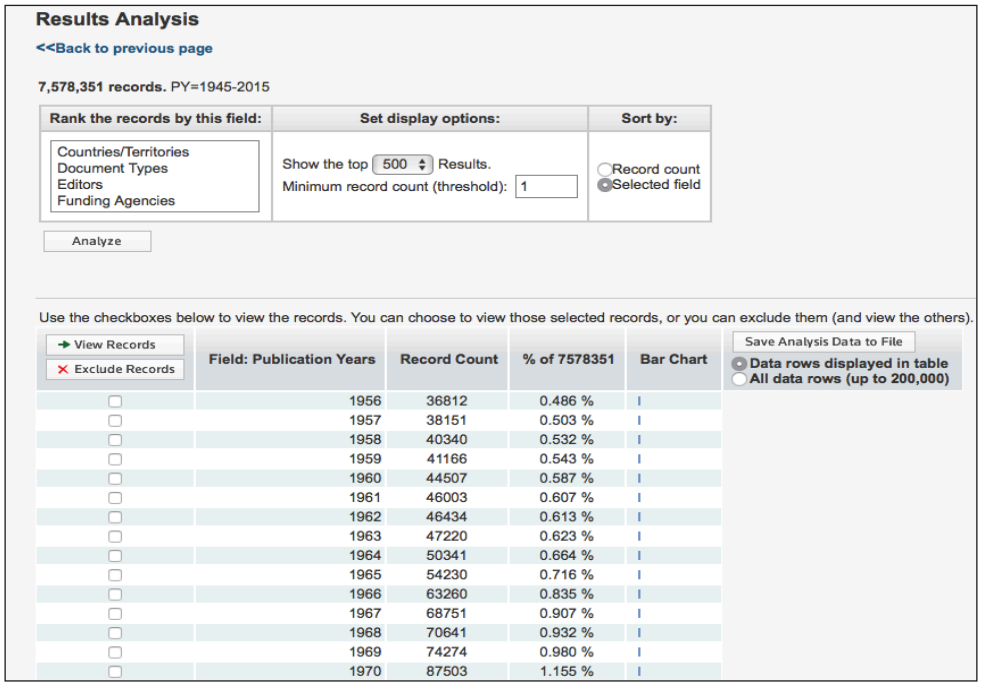
After having defined this corpus, we can extract the numbers of a specific search term, in our case, the term "evolution". The search process will follow the same steps, only this time adding the search term next to the publication year and chose if abstracts should be included. In our example we would like to have an overview of the development of the term during the complete timespan of the database. Therefore typing in `PY=1945-2015 AND TI="evolution"` in the Advanced Search mask will only analyze the semantic content of the title, irrespective of whether they also contain abstracts and keywords. Alternatively, it would be possible to search for `TS="evolution"`, which would include abstracts and keywords, but then we should restrict the time span to `PY=1991-2015`. Otherwise the relative frequencies for the whole period would be screwed. The search results shows 190.157 documents in the SCI containing "evolution" in the title, with a publication year between 1945 and 2015 **[Screenshot 4]**.


**Screenshot 4**

If we then again analyze the results, a downloadable file can be created on the basis of the search, which then can be opened in a spreadsheet application and be normalized against the total SCI corpus we extracted before. Doing so enables us to trace the appearance of the term "evolution" as a part of a title in scientific articles throughout the second half of the 20th century.

As a matter of course it is necessary to repeat the above explained process with the corresponding selection of the scientific field for social sciences to complete the data extraction. To do so, we return to the *Advanced Search* field and start building our corpus for the social sciences. Instead of typing in the necessary information for our search inquiry again, we use the *Edit option* right next to the result in the *Search History* table. The search field will refresh with a highlighted blue background enabling us to overwrite or create a new search set based on the adjustment we used before. For a better overview of the various search inquiries we recommend to create a new set instead of overwriting the old one. The only adjustment we now have to make is to select the SSCI instead of the SCI.

This new search command leads to the result of 7.578.351 publications representing the total number of database entries associated with the social sciences. Via the *Analyze Results* function we again sort the result by publication year and than download the resulting numbers per year, which define our total corpus for the social sciences. This time, however, the downloadable list will not start with the year 1945, but with the year 1956, since this was the year the SSCI was introduced to the *Core Collection* of *Web of Science* **[Screenshot 5]**. Whenever the individual indexes are used in direct comparison one has to keep in mind that the search result will always just present data starting from the year the respective index was introduced.




**Screenshot 5**

After having successfully defined the corpus for the social sciences we can start extracting the data for the term “evolution” inside the SSCI by repeating the steps explained above. We return to the *Advanced Search* field and chose to edit search set #2, which contains all the adjustment settings we used to extract data concerning the term “evolution” inside the SCI. Running our fourth and final search will result in 24.826 **[Screenshot 6]** documents representing the number of publications inside the SSCI containing the term “evolution” in their title. Again we can download this data after ranking the numbers of the result by publication year.


**Screenshot 6**

In the end we have extracted four different sets of data: Two corpora that represent the total number of publications inside the respective research areas (natural sciences set #1 and social sciences set #3) and two data sets containing the total number of publications inside these two corpora that contain the term “evolution” in their title (set #2 and set #4). This data can now be used to produce the diagrams shown in **Figure 3**. To do so one only has to copy the numbers into a spreadsheet application and normalize the results with regard to the respective corpus.


Figure 3: Relative number of publications in the *Web of Science* Science Citation Index (SCI) and Social Science Citation Index (SSCI) that contain the respective term in the title, 1956-2015, smoothing=3, data retrieval September 20, 2016.



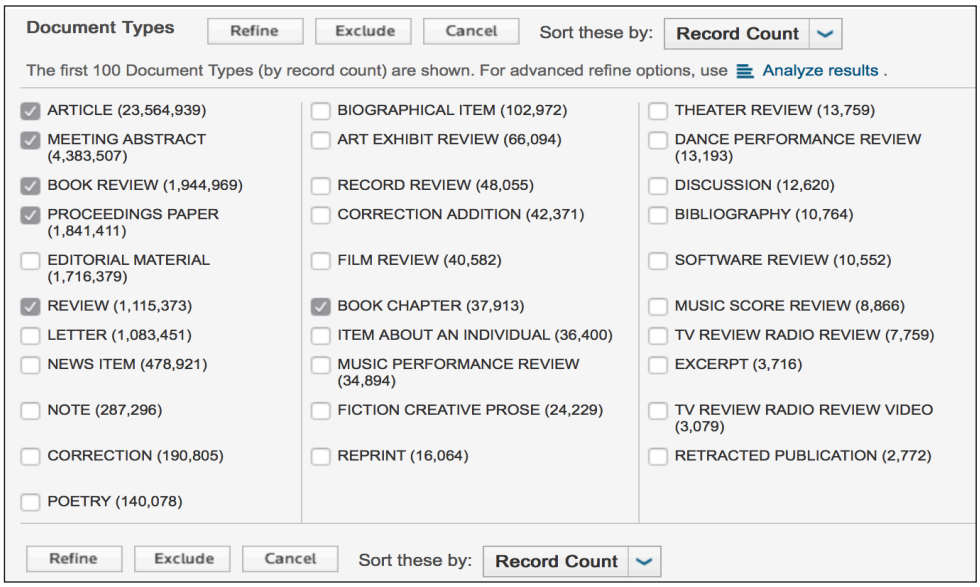
## 2.3. Extracting data: Document types as distinguishing variables

In the last section, we demonstrated how to trace the appearance of a scientific term (“evolution”) in two different corpora. In the following, we present another example, which focuses on what we call a meta-scientific concept (“grand challenges”) **[10]**. Such meta-scientific concepts are more important in the field of science policy than in the technical languages of scientific disciplines; they are used, however, in normal scientific publications as well. In this case, we concentrate specifically on the bibliographic matter and use document types as distinguishing variables. We conduct a topic search [TS] to trace the development of the term “grand challenges” in two different corpora: First, we build a corpus (P) that contains only scientific publications in the narrow sense, that is, documents types like *Articles*, *Book Chapters*, *Book Reviews*, *Meeting Abstracts*, *Meeting Summaries*, *Proceedings Papers*, and *Reviews*. Second, we build a corpus (M) with meta-scientific communication formats, which contains *Editorial Material*, *Letters*, *News Items*, and *Notes*. Because a topic search depends on the existence of abstracts, the time span is set from year 1991 until 2015. The example thus illustrates an extraction method that integrates the possibilities of individual category selection for a more complex corpus design.

**[10]** For a more in-depth discussion of the “grand challenges” concept, including quantitative data, see Kaldewey (forthcoming).

Like the data extraction explained in section 2.2, the determination of the corpus starts with narrowing down all documents within the selected time span by a simple search request for the publication year. In this example, we include the three indexes SCI, SSCI and A&HCI under the *More Settings* drop down menu **[11]** .Typing `PY=1991-2015` in the *Advanced Search* field leads to a first result of 35.403.310 database entries, representing all documents inside the three indexes published during this period. The result can now be refined for the first corpus by opening up the corresponding window on the search mask. To define our corpus (P) we select *Document Types* and start compiling it by clicking on the relevant boxes for the respective document types, namely *Articles*, *Book Chapters*, *Book Reviews*, *Meeting Abstracts*, *Meeting Summaries*, *Proceedings Papers* and *Reviews*. Once the respective types are selected clicking on *Refine* will exclude all the other document types from the first result of over 30 million and only depicts those tagged with one of these document types **[Screenshot 7]**. We now have a result of 31.007.978 database entries. It would be possible to modify the result further by applying more filters, but the corpus determined so far represents all relevant variables in regard to the underlying research interest. To now extract the data we have to follow the same steps already explained. Again, the *Analyze Result* function enables to rank the entries by publication year.






**Screenshot 7**

Now that we successfully defined our first corpus for investigating the development of the "grand challenges" concept in between different document types, the numbers of the phrase itself need to be gathered. As done before, the search entry must be equal to the settings used for corpus (P) and combined with the search term: `PY=1991-2015 AND TS=\"grand challenge*"`. The *-icon at the end of the term serves as a wildcard to include variations such as the plural form ("challenges"). The search inquiry now shows a result of 1.383 entries. This result, however, is not yet what we are aiming for, since it contains all publications and not only those in our corpus (P). Therefore, this result needs to be refined further through the respective categories. As before, opening the corresponding window enables the selection of the document types for a

**[11]** The Emerging Sources Citation Index covers only a small time period (2015-present) which has too little impact on the research of this paper to be included.

refined search, where we can check the boxes for *Article*, *Book Chapter*, *Book Reviews* and so on. The refined search will reduce the number to a new result of 1.117 database entries. This number now includes all necessary information to fit our corpus (P). After this step we have to again rank the upcoming result by publication year via the *Analyze Result* function before downloading it.

Having fulfilled these steps the results give us two data files. The first contains the total number of documents inside the *Web of Science Core Collection* categorized by the document types we considered scientific publications (P) in the narrow sense. Corresponding to this, the second file contains the total number of documents which are adjusted by the same settings, in addition to the appearance of the phrase "grand challenge*" in their title, abstract or as a keyword.

The whole analysis now has to be repeated in regard to the second corpus before proceeding with a qualitative comparison. Therefore we return to the Advanced Search field and start defining our corpus (M): First we narrow down all entries with a publication date in between the years 1991 until 2015 by typing `PY=1991–2015` in the search field. After that we start refining the result via the *Document Types* menu to build the corpus for meta-scientific communication formats. Running the search with the document types *Editorial Material*, *Letter*, *News Item* and *Note* leads to a new result of 3.566.084 entries. This number represents the total amount of database entries inside the three selected indexes of the *Core Collection* we consider meta-scientific communication formats.



Now that we have extracted corpus (M) as the equivalent to corpus (P) our last search inquiry will trace the appearance of "grand challenges" inside corpus (M). As before we narrow down all publications from 1991 until 2015 linked to "grand challenges" by typing `PY=1991–2015 AND TS="grand challenge*"` in the *Advanced Search* field. The result shows 1.383 database entries, which is of course the same amount we got in our search before, since we have not adjusted any settings differently. In the next step we therefore have to again open up the *Document Type* menu and select the respective types we used to design corpus (M). This time, however, the available selection is reduced and even one of the types we used for corpus (M) – the document type *Note* – is missing. This is due to a very low number of publications corresponding to our search criteria. Especially when searching for several terms inside the same corpus, it is important to pay attention to the fact that some categories may not be selectable for all terms of interest. In our exemplary case this causes no problems because we can well do with the other three document types: *Editorial Material*, *Letter* and *News Item*. The new number of 264 database entries is now the final result indicating the appearances of "grand challenges" as a topic inside our individually designed corpus (M). Having all datasets processed through a spreadsheet application, we can eventually produce a diagram comparing the different trajectories **[Figure 4]**.

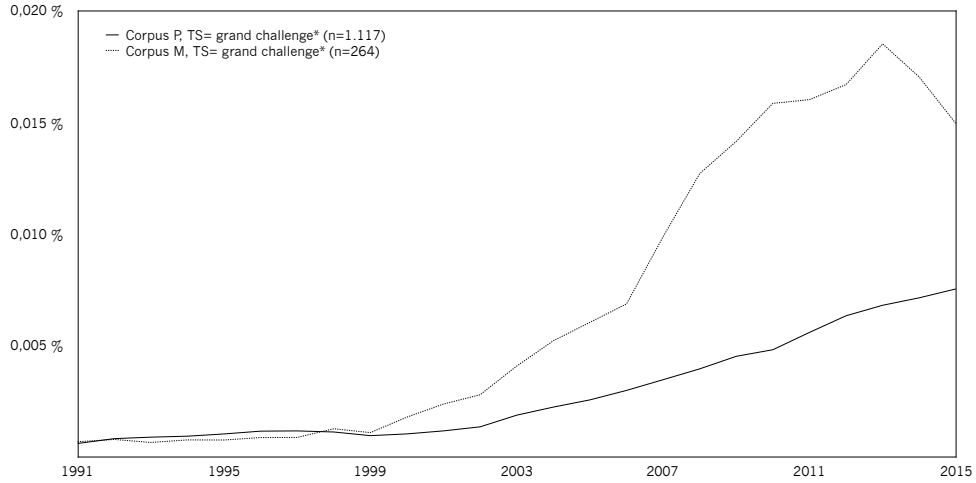



**Figure 4:** Relative numbers of publications in Web of Science core collection, splitted in two individualized corpora (P and M), that contain the respective term, in title, abstract or keywords, 1991-2015, smoothing=3, data retrieval July 28, 2016.



## 2.4 Handling guidelines

Intentionally developed as a citation index, *Web of Science's* core feature is the interlinkage of all publications via citations. In our analysis, we actually made no use of this function – rather, we proposed an "off label" use of the database. Therefore, a modified use of the content comes with certain issues which have to be kept in mind when designing own corpora in the way explained above. This section discusses some disadvantages and advantages of working with the *Web of Science* database in regard to quantitative analyses of scientific and meta-scientific concepts.

One disadvantage is the limited time period covered by the different indexes of the *Core Collection*. The more complex the design of a desired corpus gets, the shorter the time span of the analyzable data will be. In the example from section 2.2 we examined the appearance of the term "evolution" in two research areas, represented by the SCI and the SSCI. Even though the SCI includes publications from 1945 onward, the comparison of the data in Figure 3 starts in 1956, because analyzing the SSCI before that data makes no sense. Hence it is crucial to keep the historical development of the different indexes in mind (see Figure 1 in section 2.1).

A similar problem is given with the difference between *title* [TI] and *topic* [TS] search. The advantage of a topic search relies on the broad semantic data the search will focus on, yet data which results from a topic search is limited to the time period after 1991. Every search inquiry before 1990 must be restricted to a *title* [TI] search, which limits the depth of the data. This problem points to the more general disadvantage, namely the lack of a full-text database. As the archive of *Web of Science* is rather a collection of information about scientific publications than a corpus of publications itself, the interpretation level of the result is bound to this structure of the archive.

On the other hand, *Web of Science* is specifically built for scientific purposes and only handles publications of the scientific sphere, which is why data extracted from this source is a valuable indicator of how scientific language (in contrast to everyday language) developed over time.

In our opinion the *Web of Science* database is well suited for short term quantitative analyses concentrating on the second half of the 20th century, and, particularly, for more recent developments after 1991. Due to the huge amount of categorized publications, as well as its high variety of search adjustments in combination with its accuracy of processing complex search inquires it serves as a reliable source. In comparison to the *Google Ngram Viewer*, users of the *Web of Science* database have a clear overview of the presented data and direct access to the information of every single item uploaded to the archive. The Advanced Search interface enables a clean organization of the search history and offers productive possibilities of editing and tweaking already conducted search inquiries. Particularly when extracting data of several terms inside the same corpus, these options save time during the whole work process.



# 3. SCIENCE MAGAZINE

After discussing the "off label" use of the *Web of Science* database for quantitative semantic analysis, the following section is dedicated to journal archives as another source for quantitative analyses of the development of scientific and meta-scientific concepts. Even though in principle every journal that provides an online archive with a full-text search option can be used for the method presented here, we chose the *Science* Magazine as an exemplary source for the following reasons. *Science* is an internationally known and high valued peer-reviewed journal, which, due its salient position in the academic world, does have significant impact on scholars and policy makers around the world. Owned by the *American Association for the Advancement of Science* (AAAS) the journal publishes its print version weekly to a subscriber base of over one hundred thousand. In this function *Science* can be regarded as representative for the current situation of science and science policy, particularly in the United States. The readership reached by the online services is estimated by half a million. Furthermore, *Science* is decidedly multidisciplinary in character. The editors state the general interest for manuscripts from all areas of scientific research, though the journal's high reputation builds essentially upon contributions from the natural sciences. *Science* today is more than a "normal" scientific journal, it is a broad platform offering publishing space for editorials, news, reports and more. As a multidisciplinary journal containing this range of meta-scientific documents it is a perfect source for extensive analyses of conceptual developments in scientific communication. Finally, the oldest uploaded documents reach back to year 1880, enabling search inquires that cover much longer time periods than, for example, the *Web of Science* database.

The online archive of *Science* allows a full-text search via a regular search field as well as via an Advanced Search option, granting a more complex search mask. The user can adjust the search criteria by author and keywords, citation information, publishing year and the different journals *Science* has integrated over the years. For our purposes, that is, to use a clearly defined corpus, we recommend to chose the *Science* journal only and thus to exclude the other journals (such as *Science Signaling*, *Science Translational Medicine*, *Science Advances*, *Science Immunology*, and *Science Robotics*). A search request that chooses *Science* only and at the same time leaves all other fields empty reveals a base of 261.835 documents stored in archive since 1880 **[12]**. Whatever the user searches, the result consists of a list of all found documents connected to the query. The items found are sorted either chronologically or by best match.

For example, searching for the meta-scientific concept "basic research" will list all documents that contain this phrase somewhere in the full text, which are 9.017 items. However, in contrast to what we know from *Web of Science*, the results cannot be refined by document type, only by year, and there is no function to export data files containing distribution of the items per year. Nevertheless, the archive can be a useful source. In the following section we propose a pragmatic way to extract data from the *Science* archive in order to produce figures similar to those we extracted from the *Web of Science* database.



## 3.1 Working with the Advanced search function

The *Science* web page is first and foremost the online platform of the journal, focusing on the latest articles and breakthroughs in science. Therefore, the mentioned *Advanced journal search* mode is not directly accessible. To navigate there one first has to start a regular search using the search field of the web page, or simply click on the magnifying glass icon. A new window will open showing a side bar with new search options that can be selected. Here the *Advanced journal search* mode will be displayed as one of them. Again, clicking on the link will bring up a new window, now displaying the search mask we can use to generate the data for this paper. Next to the search term field, four main categories can be seen; one field to adjust the format of the results, one field for entering keywords or authors of interest, one field for a citation-specific search and one to limit the results by year, journal and collection type. Even though it is possible to format the result of a search chronologically, there is no setting to list the result by year, like seen in *Web of Science*. Hence, the data for every single year has to be extracted in separate working steps, making the "off label"-use of the *Science* database much more lavish than the one with *Web of Science*.

## 3.2 Extracting data

In this example, our research interest focuses on the term "basic research" and how it appeared inside the main journal of Science since its first publication in the 19th century **[13]**. Similarly to how we proceeded with the *Web of Science* database it is again necessary to first extract the total corpus, that is the number of all items that

**[12]** All data presented in this section were retrieved May 27, 2017. Due to regularly updates of the database the number for the total corpus may have changed.

**[13]** We do this kind of analysis against the background of recent historical research on the historical semantics of "basic research" (Pielke 2012; Schauz 2014; Kaldewey and Schauz forthcoming).

are accessible in the Science archive and their distribution over time. To get the number of all uploaded items, the field for the search term has to stay clear, enabling the search engine to list every document inside the archive only linked to the selected categories. Therefore we select the year, here to find under the category field *Citation* and then choose between the different journals under the category field *Limit Results*. For the latter the main journal *Science* is selected and the year is set to the earliest date (1880).

The resulting list will show every article (or rather, every item marked as a separate document) published in the respective year. For 1880, the total number is 342. This number must be copied into a spreadsheet application, before resetting the search and start with the next publication year of 1881 etc. Depending on the timespan of interest the repetitive extraction of the data can take up to 136 working steps (1880 to 2015). For this example we continued the work process and completed a list in our spreadsheet application, indicating how the total 242.774 items are distributed over time. This list represents the amount of searchable items inside the online *Science* journal archive, which were published from 1880 until 2015. **Figure 5** illustrates how this corpus has developed over time.



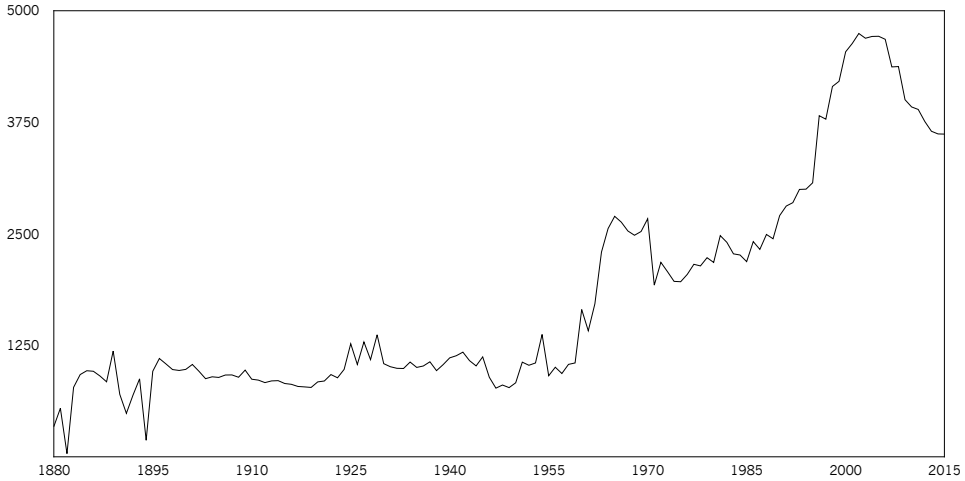



**Figure 5:** Numbers of articles per year in the Science journal archive, 1880-2015, total n=242.774, data retrieval November 8, 2016.

Since the manual completion of such a list has a higher chance for careless mistakes we recommend to use the *Publication date* menu on the right side bar for a better control of your data extraction. Here you find a compilation of several publication years displaying in brackets the amount of uploaded items during this specific time period. Though it is not possible to subdivide the numbers further into single publication years, this information can be helpful in case your data does not match with the overall search result in the *Science* archive.

After finishing the data extraction of the total corpus, the same working steps must be done in regard to the specific term or phrase that is of interest. As before, all settings must be equal to the search inquiry used to design the corpus. It is important

to enclose the phrase with quotation marks if it includes more than one word, like in our example (“basic research”), because otherwise the search engine looks for all documents that contain either one of the two words (“basic” and/or “research”). Running the search for “basic research” for the publication year 1880 shows a result of 0; thus indicating that the term was not used in the 342 publications of this year. Again, this number must be copied into the spreadsheet application and then the search procedure must be repeated for the next year. After doing this inquiry for all remaining years up to 2015 we end up with a new list in our spreadsheet application.

In total our search inquiry identified 8.900 documents that contain the phrase “basic research” in relation to the main corpus we extracted from the Science archive. Now having extracted two data sets, one representing the total corpus and the other representing the number of publications that contain our search term, we can normalize the data and produce a graph that shows the frequency of publications using the term in relation to the total publications in that given year **[Figure 6]**.

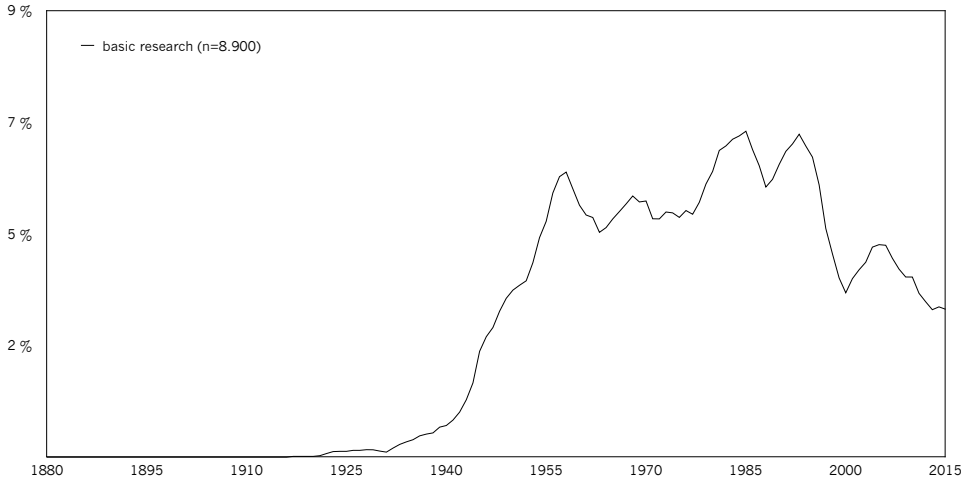



**Figure 6**: Relative numbers of publications in the Science journal archive that contain the respective term, 1880-2015, total n=242.774, smoothing=3, data retrieval December 20, 2016.



## 3.3 Handling guidelines

Using the *Science* archive as a tool for quantitative semantic analysis is not possible without compromises. Such compromises, however, basically will have to be made whenever we extract data from any kind of other journal archive, too. In this section, however, we concentrate on the advantages and disadvantages of the *Science* archive. Furthermore, as this would go beyond the scope of this paper, we do not discuss technical options of programming software that would automatize the kind of queries we have conducted manually.

One important issue in regard to the data of the *Science* archive is the origin of the journal. Seen historically, *Science* has to be regarded as a US American journal, making it a suitable indicator for the situation and developments of American Science. Over time, the journal has become more and more globally oriented, but it remains an open question of how "American" the journal's language is. In contrast to the *Web of Science* database or the *Google Ngram Viewer*, data from the *Science* archive has to be regarded as more limitied in scope, both in regard to geography and clientele, as well as in regard the the range of research areas. This makes it more difficult to express concluding statements regarding the general development of certain scientific and meta-scientific concepts. However, as the journal is definitively of global relevance and multidisciplinary in character, this issue may arguably not interfere as crucial as compared to other more national and more disciplinary journals. The limits still need to be under consideration whenever quantitative data are interpreted qualitatively.

Another problematic issue is the missing option to sort results by document type, since the *Science* archive not only contains journal articles, but also editorials, news items, letters, etc., in which the search term can appear. Furthermore, particularly for the older publications, the quality of the data is not the same as for the newer ones. In some instances it can happen that the search result shows documents in which the search term is not employed **[14]**. Although, we have not witnessed many of these cases in our previous work with the online *Science* journal archive, some errors remain in the search results.



All the same, as an example for how to do quantitative semantic analyses with full text journal archives, the *Science* magazine is valuable particularly when used in addition to analyses made with more quick-and-dirty tools such as the *Google Ngram Viewer*. Due to the covered time period (from 1880 until today) and the huge amount of documents (roughly 250.000), the Science archive allows for various queries that are not possible with other databases.

# APPENDIX

In this Appendix we present three examples of how data extracted from the three sources presented above can be combined. Since a full analysis of the respective examples would demand several new and separate papers, we present the figures with minimal explanation and interpretation only. Our point here is simply to illustrate the possible combinations and comparisons of corpora that come with the off-label use of bibliographic databases and journal archives.

**[14]** For example, if several document types are printed in the same PDF-file (such as various letters and items on the same page), the search function attributes a search term in the PDF to all articles that are, in part, on this page.

## A. Theory of Evolution vs. Evolutionary Theory

Figure A1 to A4 concentrate on the comparison between the terms “evolutionary theory” and “theory of evolution”. Here we want to find out how these related terms were used in the scientific literature, how their appearances over the time relate to each other and if significant changes can be pinned down to specific time periods. The figures show that the term “theory of evolution”, which originates in the 19th century, gradually made room for the concept of “evolutionary theory” in the 20th century. For a further discussion of how such terminological transitions relate to actual developments in the history of evolutionary theory, see Chumtong (2015).

## B. Social Evolution vs. Cultural Evolution

Figure B1 to B4 focus on the terms “social evolution” and “cultural evolution”. Comparable to example A, the data from the *Google Ngram Viewer* (B1) and the *Science* archive (B2) can serve as a starting point for further explorations into how both terms emerged and developed over time. Advanced search features in *Web of Science* enabled us to use research areas as distinguishing variables (B3, B4). Because there is not enough data if one restricts this search to a *title* [TI] search, we conducted a *topic* [TS] search from 1991 to 2015. The resulting figures are helpful to understand conceptual developments in the social sciences: References to “social evolution” date back into the 19th century. In the course of 20th century, particularly in the last decades, talking about “cultural evolution” has become more common, and we may speculate whether today the idea of “cultural evolution” has somehow replaced older notions of “social evolution”.



## C. The Linear Model of Innovation

The examples from Appendix A and B referred to scientific concepts, that is, to technical terms from disciplinary communication contexts. In contrast, Figure C1 to C4 illustrate that similar analyses can be made with what in this paper has been labelled meta-scientifc concepts. We chose the three terms “basic research”, “applied research” and “technological innovation”, because they together constitute what has become known as the “linear model of innovation”. Many scholars assume that this model was a kind of master narrative for 20th century science policy. At the same time, starting in the 1980s, scholars began to declare that the model is “dead” (e.g., Rosenberg 1991: 335). Our figure puts this diagnosis to a test. The result is surprising: Figure C1 to C3 confirm the hypothesis, that the linear model lost its persuasiveness in the 1990s. However, Figure C4, based on a *Web of Science* topic search, draws a different picture: All terms are increasing in frequency over the last 25 years. For further discussions of these developments see Kaldewey and Schauz (forthcoming).

# APPENDIX A: FIGURES

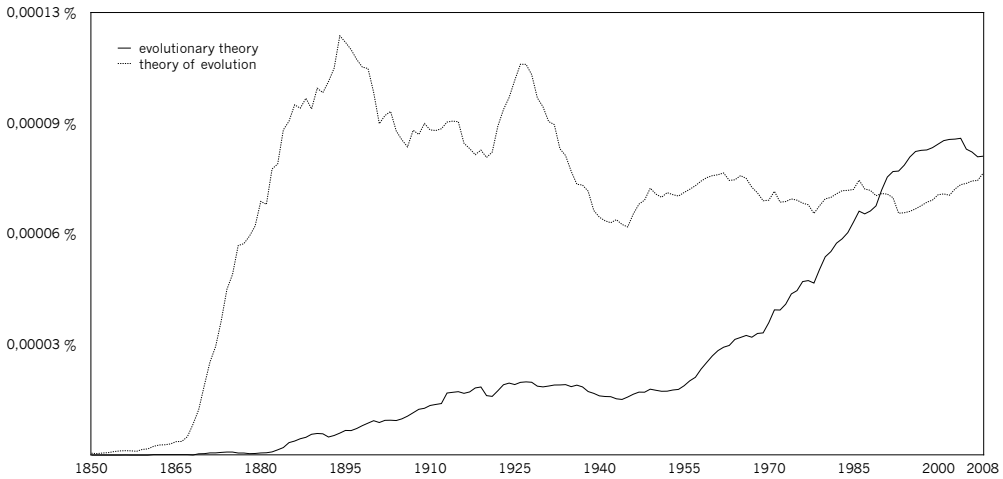



Figure A1. Relative frequencies, extracted from Google Books Ngram Viewer, 1850-2008, English corpus, case-insensitive, smoothing=3.

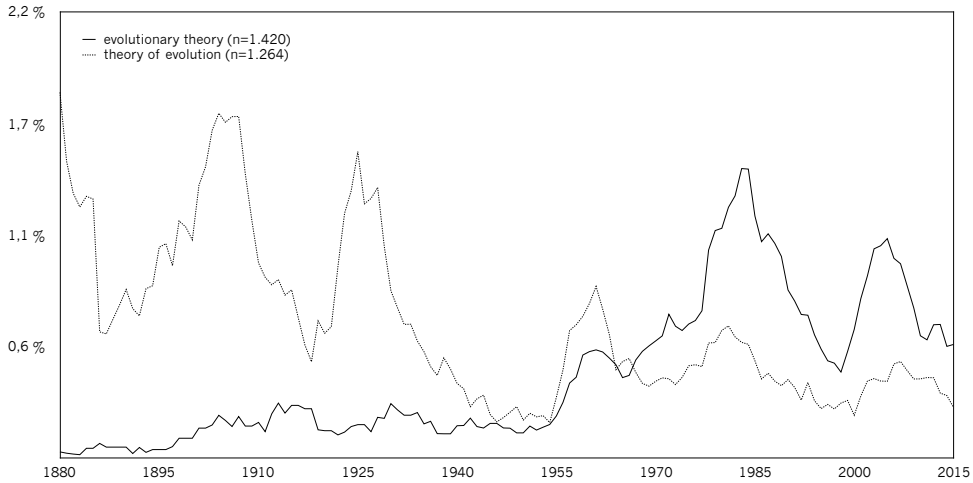



Figure A2. Relative numbers of publications in the Science journal archive that contain the respective terms, 1880-2015, total n=242.774, smoothing=3, data retrieval November 8, 2016.

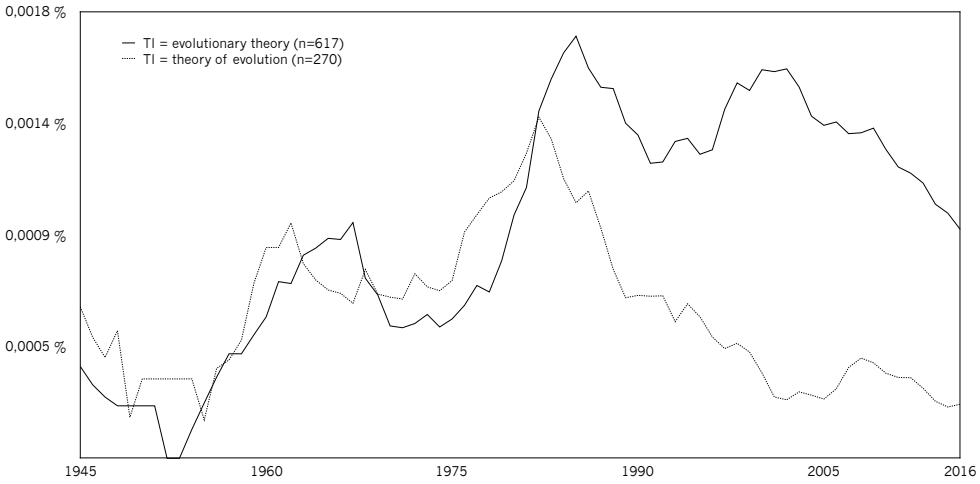



Figure A3. Relative numbers of publications in the Web of Science Science Citation Index (SCI) and Social Science Citation Index (SSCI) that contain the respective terms in the title, 1945-2016, total n=52.355.889, smoothing=3, data retrieval September 20, 2016.

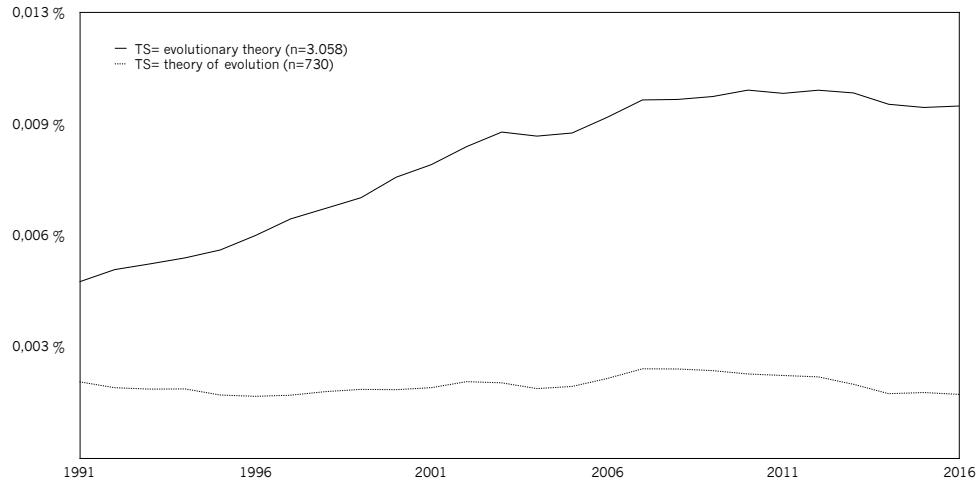



Figure A4. Relative numbers of publications in the Web of Science Science Citation Index (SCI) and Social Science Citation Index (SSCI) that contain the respective terms in title, abstract or keywords, 1991-2016, total n=34.833.394, smoothing=3, data retrieval September 20, 2016.

# APPENDIX B: FIGURES

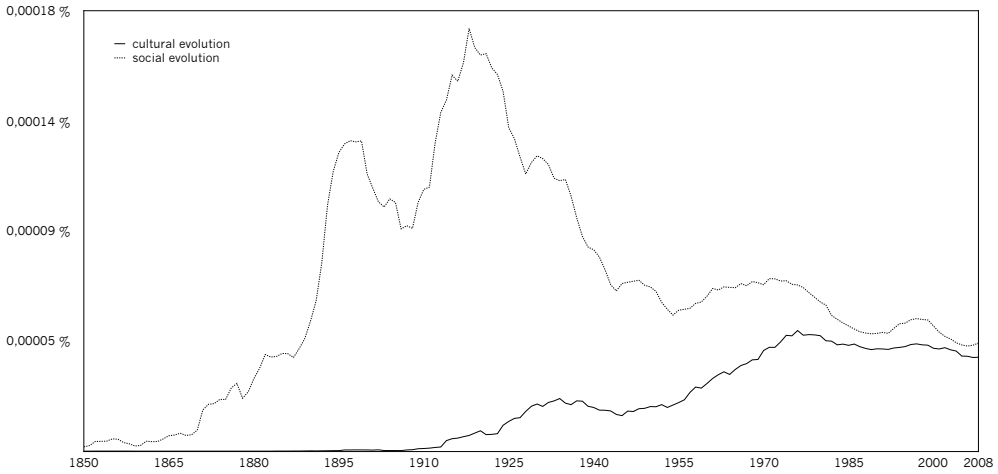



Figure B1. Relative frequencies, extracted from Google Books Ngram Viewer, 1850-2008, English corpus, case-insensitive, smoothing=3.

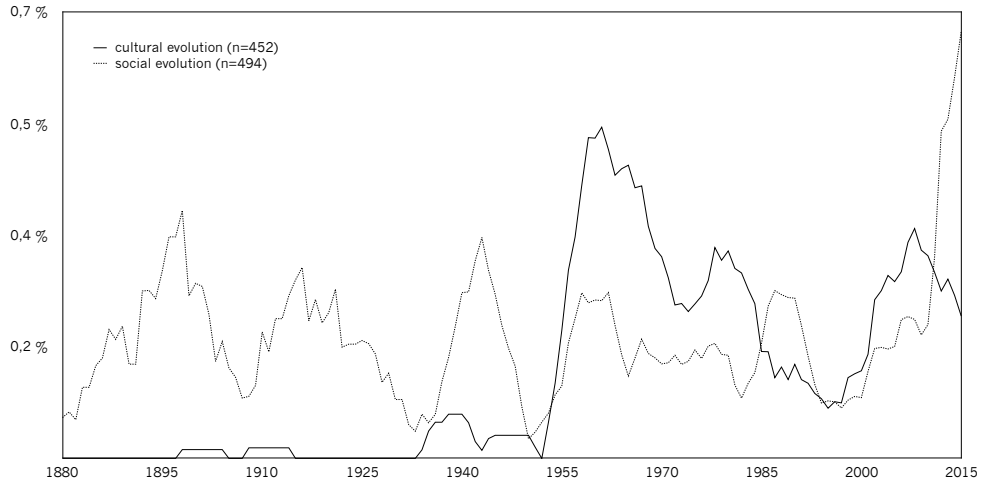



Figure B2. Relative numbers of publications in the Science journal archive that contain the respective terms, 1880-2015, total n=242.774, smoothing=3, data retrieval November 8, 2016.

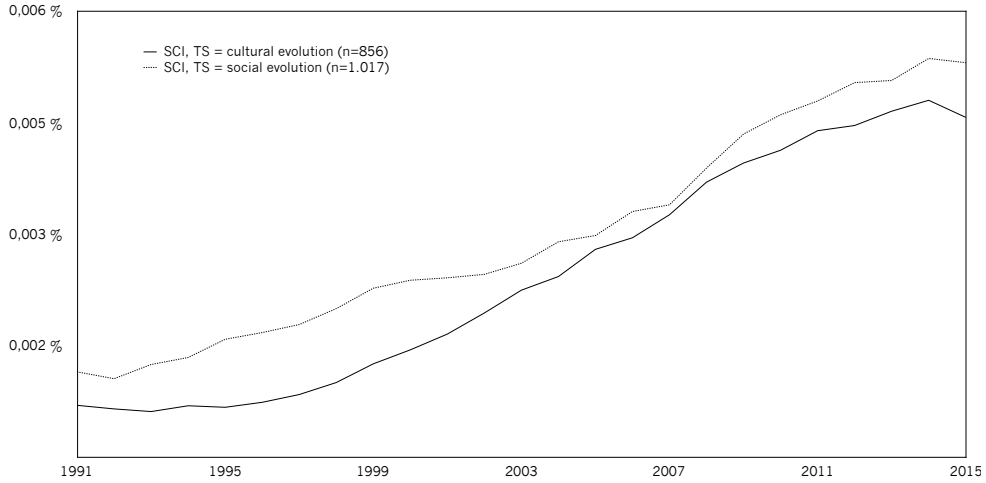



Figure B3. Relative numbers of publications in the Web of Science Science Citation Index (SCI) that contain the respective terms in title, abstract or keywords, 1991-2015, total n=29.846.675, smoothing=3, data retrieval September 20, 2016.

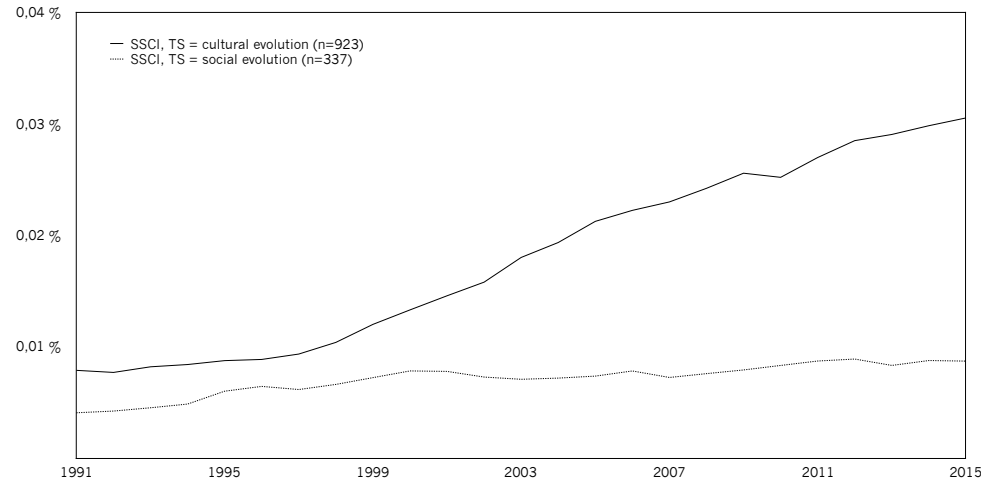



Figure B4. Relative numbers of publications in the Web of Science Social Science Citation Index (SSCI) that contain the respective terms in title, abstract or keywords, 1991-2016, total n=4.520.047, smoothing=3, data retrieval September 20, 2016.

# APPENDIX C: FIGURES

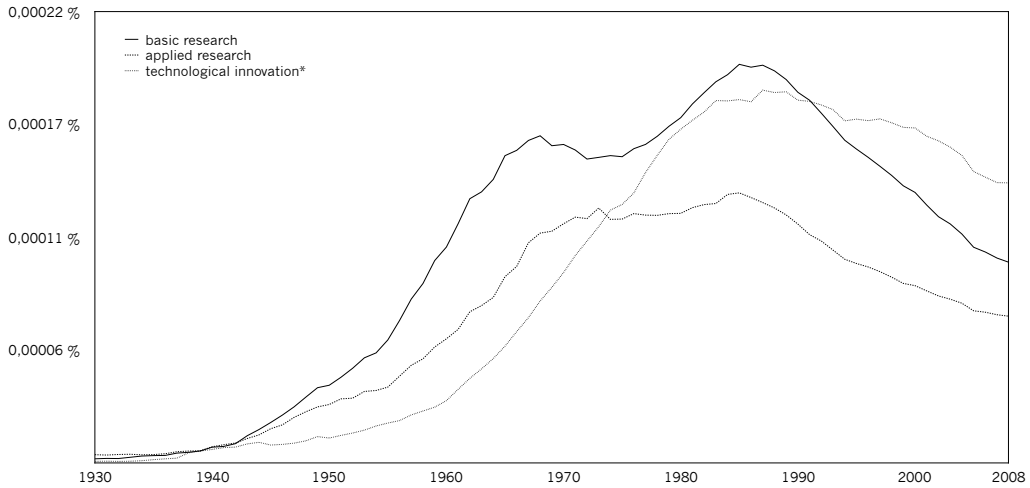



Figure C1. Relative frequencies, extracted from Google Books Ngram Viewer, 1930-2008, English corpus, case-insensitive, smoothing=3.

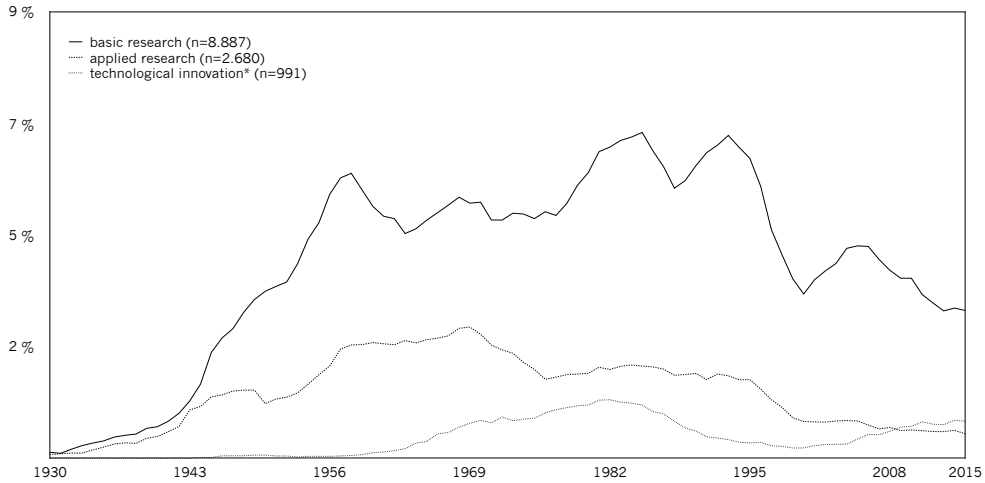



Figure C2. Relative numbers of publications in the Science journal archive that contain the respective terms, 1930-2015, total n=242.774, smoothing=3, data retrieval November 8, 2016.

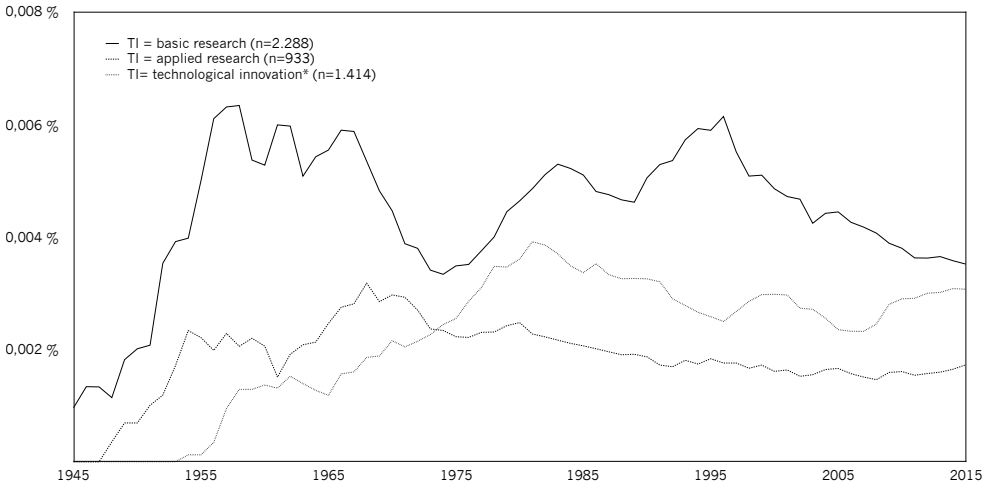



Figure C3. Relative numbers of publications in the Web of Science core collection that contain the respective terms in the title, 1945-2015, total n=54.316.656, smoothing=3, data retrieval August 11, 2016.

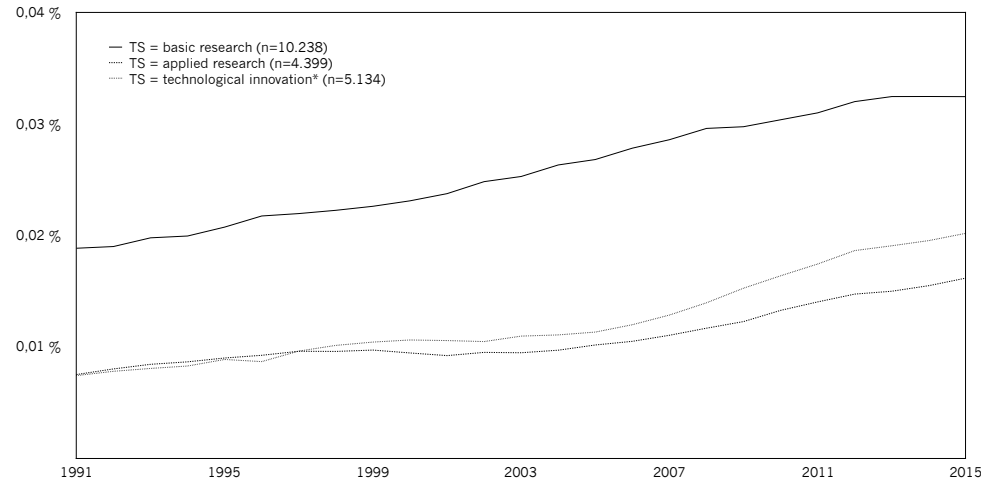



Figure C4. Relative numbers of publications in the Web of Science core collection that contain the respective terms in title, abstract or keywords, 1991-2015, total n=35.326.253, smoothing=3, data retrieval August 11, 2016.

# REFERENCES


» Chumtong, Jason (2015): Die Evolution der Evolutionstheorie. Eine wissenschaftssoziologische Reflexion zur Semantik des Evolutionsbegriffes. Bachelor-Arbeit im Studiengang Politik und Gesellschaft, Philosophische Fakultät, Rheinische Friedrich-Wilhelms-Universität Bonn.

» Kaldewey, David: "The Grand Challenges Discourse. Transforming Identity Work in Science and Science Policy." *Minerva* (forthcoming).

» Kaldewey, David and Désirée Schauz, eds.: *Basic and Applied Research. Language and the Politics of Science in the Twentieth Century*. New York: Berghahn Books (forthcoming).

» Klein, Daniel B. (2013): "Ngrams of the Great Transformations." George Mason University, Department of Economics, Working Paper No. 13-10.

» Lin, Yuri et al. (2012): "Syntactic Annotations for the Google Books Ngram Corpus". *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju, Republic of Korea, pp. 169–174.

» Michel, Jean-Baptiste et al. (2011): "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, pp. 176–182.

» Ophir, Shai (2016): "Big Data for the Humanities Using Google Ngrams. Discovering Hidden Patterns of Conceptual Trends." *First Monday* 21(7).

» Pechenick, Eitan Adam, Christopher M. Danforth and Peter Sheridan Dodds (2015): „Characterizing the Google Books Corpus. Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution." *PLOS One*.

» Pettit, Michael (2016): "Historical Time in the Age of Big Data. Cultural Psychology, Historical Change, and the Google Books Ngram Viewer." *History of Psychology* 19(2), pp. 141–153.

» Pielke, Roger A. (2012): "'Basic Research' as a Political Symbol." *Minerva* 50(3), pp. 339–361.

» Rosenberg, Nathan (1991): "Critical Issues in Science Policy Research." *Science and Public Policy* 18(6), pp. 335–346.

» Roth, Steffen, Carlton Clark and Jan Berkel (2017): "The Fashionable Functions Reloaded. An Updated Google Ngram View of Trends in Functional Differentiation (1800-2000)". In: *Research Paradigms and Contemporary Perspectives on Human-Technology Interaction*. Ed. by Anabela Mesquita. Hershey: IGI Global, pp. 236–265.

» Schauz, Désirée (2014): "What is Basic Research? Insights from Historical Semantics." *Minerva* 52(3), pp. 273–328.

# ABOUT THE AUTHORS

Jason Chumtong holds a BA degree in political science and sociology from the Rheinische Friedrich-Wilhelms-Universität Bonn. In his bachelor thesis (2015) he discussed the evolution of evolutionary theory in the natural and social sciences. Between 2013 and 2017 he worked as a research assistant at the Forum Internationale Wissenschaft (FIW). In October 2015 he was visitng researcher at the University of Oslo (IKOS). Since September 2017 he is inrolled in the postgraduate programme "Science and Technology in Society" at the University of Edinburgh.

David Kaldewey is Junior Professor for Science Studies and Sociological Theory at the University of Bonn and leader of the research group "Discovering, Exploring, and Addressing Grand Societal Challenges," funded by the Mercator Foundation. He holds a PhD in sociology from Bielefeld University. In his book *Wahrheit und Nützlichkeit* (2013) he explored discourses on the goals and values of science in a long-durée perspective. Several forthcoming publications deal with the changing relationship of science and politics, particularly with the contemporary pluralization of science policy discourses and how they transform the identity work of scientists and policy makers.

# FIW WORKING PAPER

**Titel:** Zum Forschungsprogramm des Forum Internationale Wissenschaft der Universität Bonn
**Autor:** Rudolf Stichweh
**Datum:** September 2015
**ISBN:** 978-3-946306-00-9



**Titel:** Following the Problems. Das Programm der Nachwuchsforschergruppe „Entdeckung, Erforschung und Bearbeitung gesellschaftlicher Großprobleme“
**Autoren:** David Kaldewey, Daniela Russ und Julia Schubert
**Datum:** September 2015
**ISBN:** 978-3-946306-01-6

**Titel:** Politische Demokratie und die funktionale Differenzierung der Gesellschaft. Zur Logik der Moderne
**Autor:** Rudolf Stichweh
**Datum:** April 2016
**ISBN:** 978-3-946306-02-3

**Titel:** The Soviet Organisational Society. Political Control, the Soviet Village, and World Society
**Autor:** Evelyn Moser
**Datum:** Mai 2016
**ISBN:** 978-3-946306-03-0


**Titel:** Markets, Order and Noise: Two Contributions to a Comprehensive Understanding of Modern Markets
**Autor:** Pascal Goeke, Evelyn Moser
**Datum:** September 2016
**ISBN:** 978-3-946306-04-7


**Titel:** CSR public policies in India's democracy: Ambiguities in the political regulation of corporate conduct
**Autor:** Damien Krichewsky
**Datum:** März 2017
**ISBN:** 978-3-946306-05-4




**Titel:** The Making of the '*Geisteswissenschaften*' A Case of Boundary Work?
**Autor:** Julian Hamann
**Datum:** Juni 2017
**ISBN:** 978-3-946306-06-1


**Titel:** Beyond the Google Ngram Viewer: Bibliographic Databases and Journal Archives as Tools for the Quantitative Analysis of Scientific and Meta-Scientific Concepts
**Autor:** Jason Chumtong, David Kaldewey
**Datum:** August 2017
**ISBN:** 978-3-946306-07-8