# Researchers and their data. A study based on the use of the word data in scholarly articles

Frédérique Bordignon, Marion Maisonobe

RESEARCH ARTICLE

# Researchers and their data: A study based on the use of the word *data* in scholarly articles

**Frédérique Bordignon**[1,2] and **Marion Maisonobe**[3]

[1]Ecole des Ponts, Marne-la-Vallée, France
[2]LISIS, INRAE, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France
[3]Laboratoire Géographie-cités, CNRS, Université Paris 1, Université Paris Cité, EHESS, Aubervilliers, France

**Keywords:** corpus-based study, data, research article, science studies, scientometrics, textometrics

## ABSTRACT

*Data* is one of the most used terms in scientific vocabulary. This article focuses on the relationship between data and research by analyzing the contexts of occurrence of the word *data* in a corpus of 72,471 research articles (1980–2012) from two distinct fields (Social sciences, Physical sciences). The aim is to shed light on the issues raised by research on data, namely the difficulty of defining what is considered as data, the transformations that data undergo during the research process, and how they gain value for researchers who hold them. Relying on the distribution of occurrences throughout the texts and over time, it demonstrates that the word *data* mostly occurs at the beginning and end of research articles. Adjectives and verbs accompanying the noun *data* turn out to be even more important than *data* itself in specifying data. The increase in the use of possessive pronouns at the end of the articles reveals that authors tend to claim ownership of their data at the very end of the research process. Our research demonstrates that even if data-handling operations are increasingly frequent, they are still described with imprecise verbs that do not reflect the complexity of these transformations.

## 1. INTRODUCTION

While data are an essential component of the scientific method, it is commonly accepted that they play a growing role in our contemporary societies. This evolution is related to a change of scale in our access and means to process data (Chen & Zhang, 2014). As this change of scale also concerns research data, one may wonder if it has an effect on the way researchers use and refer to data. Given that some consider the changes under way as capable of transforming how research is done across disciplines, the investigation we provide is timely. The aim of this paper is indeed to contribute to the science studies that focus on what Borgman (2015) calls "data scholarship," that is, the complex arrangement of the relationship between data and research. To this end, we propose to explore the traces of this relationship in scientific articles.

Most scientometrics studies dealing with the advent of the "Big Data era" document the development of research areas specifically related to data handling (Huang, Schuehle et al., 2015; Maisonobe, 2022), but no comprehensive, longitudinal and transversal study has been performed on the content of scholarly articles' full text to monitor the way academic scholars refer to data. However, Coxhead (2000) has evidenced that the word *data* is among the most used terms of academic vocabulary, as did Hyland and Tse (2007) who also insisted on field-to-field variations.

an open access journal

Citation: Bordignon, F., & Maisonobe, M. (2022). Researchers and their data: A study based on the use of the word *data* in scholarly articles. *Quantitative Science Studies*, *3*(4), 1156–1178. https://doi.org/10.1162/qss_a_00220

DOI: https://doi.org/10.1162/qss_a_00220

Peer Review: https://publons.com/publon/10.1162/qss_a_00220

Received: 10 May 2022
Accepted: 2 October 2022

Corresponding Author: Frédérique Bordignon
frederique.bordignon@enpc.fr

Handling Editor: Ludo Waltman

According to Rosenberg (2013, 2018), the word *data* progressively gained popularity among scholars during the 19th century and throughout the 20th century, when it started to be more often used with the meaning it is attributed today. A quick glance at the Web of Science Core Collection suggests the continuation of this positive dynamic. Between 1991 and 2019 the share of article abstracts containing the word *data* increased from 13.6% to 18.6%. In 2020, this share exceeds 20%, which means that the word *data* is now present in the abstracts of one out of five academic papers[1].

The widespread use of the word and its frequent appearance in the abstracts of scientific articles, however, give us only limited insight into its use by researchers within their articles and over time. Moreover, as many analysts have pointed out in reaction to the sometimes excessive enthusiasm generated by the Big Data movement (Boyd & Crawford, 2012; Tenopir, 2016), the term *data* is so polysemous that it seems difficult to understand it independently of its context. In particular, most specialists share the idea that *data* does not have a stable meaning, its significance and utility varying throughout the research process (Hilgartner & Brandt-Rauf, 1994; Latour & Woolgar, 1986; Leonelli, 2020; Longino, 2020).

To account for this diversity of use and meaning, and to better understand how data are mobilized by researchers, we offer to investigate how the rhetorical function of *data* emerges in academic writing by analyzing in which contexts *data* occurs in scientific articles, which are among the written outputs that researchers use to formulate their argument and reasoning. To carry out this investigation, we use a computational method of discourse analysis, starting from the pivot term *data* and looking for different lexicogrammatical realizations in our corpus of full texts. We undertake the textometric analysis of a corpus of 72,471 scientific articles with 896,159 occurrences of the word *data* retrieved from scholarly articles available in the ISTEX database (Excellence Initiative for Scientific and Technical Information; see below for details). We consider the evolution of the use of the term and its context of use (its neighborhood) over the course of the articles and over a 30-year span (1980s to 2010s). To echo the work revealing the differences in practices between disciplinary fields, we provide a comparative study between physical sciences and social sciences.

Our main research question is therefore the following: What does the linguistic environment of the word *data* tell us about the very nature of data in two different disciplinary fields, and over a 30-year period? This study is indeed a corpus-based lexical study that aims to shed light on the different issues raised by research on data, namely the difficulty of defining what is considered as data, the transformations that data undergo during the research process, and how they gain value for the researchers who hold them. After the state of research, the article presents the data and methods used to carry out this study and the results obtained by focusing on the occurrence of *data* throughout the article, and in different contexts of use, including the combination with adjectives, verbs, and possessive pronouns. In addition to a tailored categorization of the types of adjectives and verbs pertaining to the word *data*, our research brings out interesting stylized facts about the evolution of the use of the term across articles and interesting distinctions between scientific domains.

## 2. BACKGROUND

### 2.1. Drawing on Scientific Articles

Each research field develops its own inscriptions (Latour, 1999) to record, describe, and represent what it considers to be data. Some of those traces can be found in scientific

[1] Analysis performed on January 5, 2022, on the online version of the Web of Science Core Collection, with a restriction to articles containing an abstract written in English (all the abstracts with at least one occurrence of the word *the* are considered in this analysis).

publications, and in particular in the articles that we intend to analyze here. While the articles are of course not a faithful account of what happened in the laboratory, according to Knorr and Knorr (1978), they provide the selected measurement inscriptions of the laboratory with "contexture."

In discourse epistemetrics, a specialty at the interface between quantitative science studies and applied linguistics (Demarest & Sugimoto, 2015), it is agreed that the style, composition units, rhetorical moves, and registers found in peer-reviewed research articles demonstrate regularities and specificities allowing us to distinguish them from other kinds of documents and to submit them to specific bibliometric and textometric analyses. Also, from a diachronic linguistic observation, it is possible to derive a sociohistorical interpretation, in line with those who consider that the linguistic study of scientific discourses can inform the research process itself (Bazerman, 1981; Mullins, Snizek, & Oehler, 1988). We follow the footsteps of works that take the scholarly article as an object and as a reflection of research practices, social context, and "discourse community" (Swales, 1988), in which researchers are involved.

By studying the use of the word *data* in research articles, we assume that it can inform us about the way researchers use and value data.

### 2.2. The Polysemy of the Term *Data*

The heterogeneity of the objects falling under the term *data* has been highlighted by many studies on the topic, which then hardly succeed in providing a consensual and accurate definition of what this term covers (Borgman, 2015; Zins, 2007). Others have focused on the word *data* itself. Rosenberg (2013) relies on *datum*, the Latin etymology of the word *data*, which refers to something given in an argument, something taken for granted. According to him the term *data* is thus used to refer to claims accepted for the sake of argument and has no intrinsic formal characteristics. Rosenberg concludes that the use of the word *data* provides a rhetorical basis.

In sociology of science, the use of the word *data* has been criticized when it refers to objects that are rather constructed (Latour, 1999; Terrier, 2011). Latour (1999) suggests stopping speaking of *data* and using *sublata* instead, meaning achievements, and illustrating the cumulative nature of knowledge (Perret & Le Deuff, 2019). In the same line of thought, but in a reflection on digital humanities, Drucker (2011) also suggests reconsidering *data* and proposes to speak of *capta*: "data are capta, taken not given, constructed as an interpretation of the phenomenal world, not inherent in it." Even if neither *capta* nor *sublata* have passed into common usage, there is a consensus that data are not indeed already there (Borgman, 2015).

Drawing on the fact that adjectives qualify nouns, by their very nature, we propose to explore the meaning of the word *data* by taking advantage of the semantic content conveyed by the adjectives it combines with.

### 2.3. Data Transformation, Value, and Ownership

Data are the outcome of a number of operations, manipulations, and translations so that they can be given meaning and direction for action by those who process them (Hilgartner & Brandt-Rauf, 1994; Martin-Scholz, 2017). Many studies have explored the complexity of these operations, such as Walford (2013), drawing on an ethnographic study of scientific research carried out in the Amazon, Denis and Goëta (2017) and Plantin (2019), who have highlighted several dimensions of the process of cleaning for opening data (and make them suitable for reuse), or Ruppert and Scheel (2021) who study the "data practices" within the field of official population

statistics. They argue that these numerous operations are the central features of the transactional process through which data are both obtained for some and given for others, just as Hilgartner and Brandt-Rauf (1994) propose to use the term *data* in a broad manner that encompasses both inputs and outputs. Accordingly, data should be conceptualized as part of an evolving "data stream" (Hilgartner & Brandt-Rauf, 1994; Smolczewska Tona, 2021) or "data journey" (Leonelli, 2020). This is what leads Gitelman (2013) to consider *raw data* as an oxymoron, data being always "cooked" and never entirely "raw"; there are phenomena of the world filtered by our means of perception (Longino, 2020). Nevertheless, *raw data* is a common term for data that has not yet been cleaned; they are indeed awaiting transformation and do not yet have the status of certified data (Walford, 2013). We will consider that an article constitutes the authors' narrative of a part of the "data journey," the one they wished to share with the reader.

It is in this transformation that data become more valuable and ownership is shaped. The data have become a symbol of "what the researcher seeks and more importantly, needs" (Markham, 2013).

Whether observations, traces or other entities become data depends on who collects them and how, with different practices in different disciplines (Borgman, 2015; Leonelli, 2020; Ruppert & Scheel, 2021; Zins, 2007).

This body of literature prompted us to analyze verbs (and the actions they refer to) and to investigate the use of the possessive pronouns *my* and *our*.

Despite contradictory recommendations and still much controversy about whether to use the plural or the collective singular to refer to data, usage allows both forms (Rosenberg, 2013) and we opt for the plural, which seems more natural to us as native speakers of French.

## 3. DATA AND METHODS

### 3.1. Data

This study is a use case of the ISTEX database, an archive of scientific publications[2]. It offers French researchers online access to a retrospective collection of more than 23 million scientific publications in all disciplines and from many different publishers including Elsevier, Wiley, Springer, Oxford University Press, British Medical Journal, IOP Publishing, Nature, Royal Society of Chemistry, De Gruyter, Ecco Press, Emerald, and Brill. We therefore take advantage of this very extensive database, which offers a legal framework for text mining and saves us the tedious task of (manually) downloading full texts from the various publishers' websites. Finally, ISTEX offers technical services to export plain texts in different formats expected by text-mining tools, and also enriches the metadata provided by publishers with several discipline classifications (Dazy, 2014).

It seems essential to us that researchers of quantitative science studies take advantage of this type of database, partly designed for them[3], as they can access and process them to answer their research questions without the advanced skills of computer scientists.

We used the ISTEX API to build our corpus. We simply searched for the word *data* everywhere in the documents. After trials and visual checks, we chose to filter out documents whose quality index was under 5 (out of 10; the score is calculated by ISTEX according to checks on

[2] https://www.istex.fr.

[3] It should be noted that ISTEX is a public resource supported by the main French research organization: the CNRS (Centre National de la Recherche Scientifique).

OCR quality and the integrity of texts and metadata) to avoid possible textual residues impossible to interpret.

Rather than querying the database as a whole without anticipating possible biases due to volumetric differences between disciplines, whether as a result of random extraction, irregular ISTEX updates (i.e., according to new agreements with publishers), or heterogeneity in the OCR quality across time and editors, we have chosen to create two subcorpora based on two very broad subject areas which will allow us to make comparisons. We chose Social sciences and Physical sciences because they are unlikely to overlap too much and because they comprise distinct discourse communities. To distinguish between publications in Social sciences and Physical sciences, we relied on the Scopus classification scheme, which is available within ISTEX (Elsevier, 2021). We extracted, for each of the two domains under study, a set of 2,500 articles in 2-year slices to ensure that we have a comparable and sufficient number of articles per period, especially because, to make our results clearer in the charts, we will later group them by decade. In practice, we use the following model query with the random export option to query the database:

> data AND host.genre.raw:("journal") AND genre.raw:("research-article") AND publicationDate:[*year* TO *year+1*] AND language.raw:("eng") AND qualityIndicators.score:[5 TO 10] AND categories.scopus.raw:"1 - Social sciences"

This query was repeated for each 2-year set since 1980 to extract the publications in Social sciences and those in Physical sciences (changing: "1 - Social sciences" to "2 - Physical sciences"). Each time, we exported the corpus files in TXT (for the plain text) and JSON formats (for the metadata). We then took 2012 as the upper limit, because from 2013 onwards, the ISTEX database is unbalanced with fewer and fewer articles available in Social sciences.

We therefore have a homogeneous corpus whose text files are of good quality and which allows us to carry out a comparative analysis over time and within two major scientific fields.

As these files are copyright protected, we cannot make them available, but this methodology allows authorized users to access the database and reproduce the operations. In any case, the data we have extracted from texts are available for download (Bordignon & Maisonobe, 2022).

### 3.2. Tools

We used the software TXM (Heiden, Magué, & Pincemin, 2010) to index this corpus and, in particular, to annotate it with TreeTagger. TreeTagger is a tool that processes words from a text and labels them with a part-of-speech tag. Part-of-speech tagging is commonly used in corpus linguistics to identify word categories (e.g., verb, noun, adjective, ...) and also grammatical features (verb tense, plural/singular, ...). For example, it helps in distinguishing the verb *leaves* from the noun *leaves* thanks to the context of the sequence of words they occur in.

Then TXM is used to query the corpus and retrieve the contexts where *data* occurs and its position in the text (i.e., the rank of the sentence in the whole document). The resulting data can then be exported in CSV and processed in Tableau Software.

### 3.3. Queries

Queries for TXM must be constructed according to the Corpus Query Language (CQL) based on the combination of regular expressions and the parts of speech previously tagged. These

queries were elaborated through a long iterative process, including quality controls and numerous tests on our corpus.

For the purposes of this study, we develop queries able to extract adjectives combined with *data*, to extract the verbs and past participles that authors use to express their actions on data to assess, and the use of the possessive pronouns *my* and *our* combined with *data*.

Given this example:

*(...) in Fig. 8, we compare the uncorrected experimental data with the LA 150 data (...)*

and with the following query in TXM

*[word="I|we" & enpos="PP"] []{0,3} [!word="not|.*n't|cannot"] [word="data"%c] | [word= "data"%c][word ="that"]* [word="I|we" & enpos="PP"] []{0,3} [!word="not|.*n't|cannot"] [enpos="V.*"]*

we can retrieve the following result

*we_PP compare_VVP the_DT uncorrected_JJ experimental_JJ data_NNS*

where the verb *to compare* preceded by the pronoun *we* can be identified easily thanks to part-of-speech tagging (where the tag "VVP" stands for "verb, present non-3rd-person" and thus enables the identification of the verb).

For each occurrence retrieved by these queries, we also recorded the position in the text with the sentence reference (i.e., rank) provided by TXM; this means that each occurrence position in the text is between 1 and 100%, the value 1% representing the beginning of the text and the value 100% corresponding to the very end of the article. Our aim here was to resonate with existing studies that consider *data* as mobile entities and to check whether this is apparent in the course of the text.

We also built a special query not intended to identify the word *data* in any context, but to identify the position of the bibliography section in each document (i.e., the rank of the section title). This query served to populate a variable that we used to exclude all occurrences of *data* that are beyond this limit (i.e., occurring in the titles of publications or sources present in the articles' reference section). Our objective here is to prevent the results from being biased by occurrences of the word *data* in the full reference of a cited document, with, for example, *data* occurring in the title of a cited document.

By removing all publications where the term *data* appears only in the reference section, we finally obtained a corpus of 896,159 occurrences of *data* in the main text, provided by 72,471 research articles (Table 1).

**Table 1.** Corpus overview. Distribution of articles and occurrences of the term *data* by year and subject area

| | Number of articles | | | | Number of occurrences of *data* | | | |
|---|---|---|---|---|---|---|---|---|
| **Subject area** | 1980–1990 | 1991–2001 | 2002–2012 | All | 1980–1990 | 1991–2001 | 2002–2012 | All |
| **Physical sciences** | 11,945 | 12,277 | 11,889 | 36,111 | 158,678 | 195,177 | 137,246 | 491,101 |
| **Social sciences** | 12,022 | 12,700 | 12,090 | 36,812 | 114,451 | 153,878 | 143,995 | 412 324 |
| **All** | **23,859** | **24,874** | **23,738** | **72,471** | **271,377** | **346,812** | **277,970** | **896,159** |

## 4. RESULTS

### 4.1. Data Occurrences Throughout the Article

The following figures present the distribution of occurrences of the word *data* throughout the text, following the examples of Bertin, Atanassova et al. (2016) and Hsiao and Schneider (2021) while they study the location of in-text citations. Thus, in Figure 1, the *x*-axis refers to the progression of the text. The *y*-axis indicates the share of occurrences related to each 10% interval of the text. The figure shows that it is at the beginning and the end of the article that the authors use the word *data* the most, with slightly different curves for Social sciences and Physical sciences. However, we cannot conclude that less data is mobilized in between. On the contrary, we think that this is indicative of the authors specifying the nature of the data throughout the text by naming them differently (i.e., without using the word *data*, but by naming what can instantiate the word *data*, such as surveys, measurements, subjects, materials, numbers, and photos).

The fact that the word *data* occurs more often at the beginning and end of articles might come from the fact that the introduction and conclusion are the sections that contain more

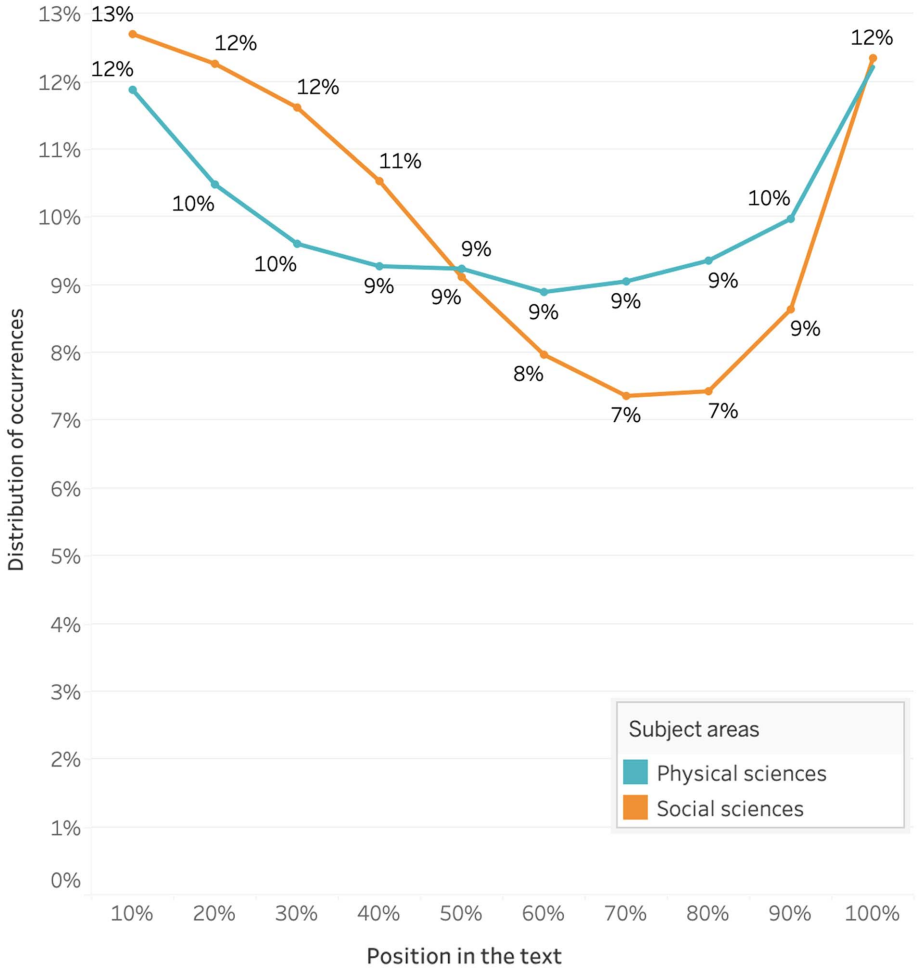


**Figure 1.** *Data* occurrences throughout the article.

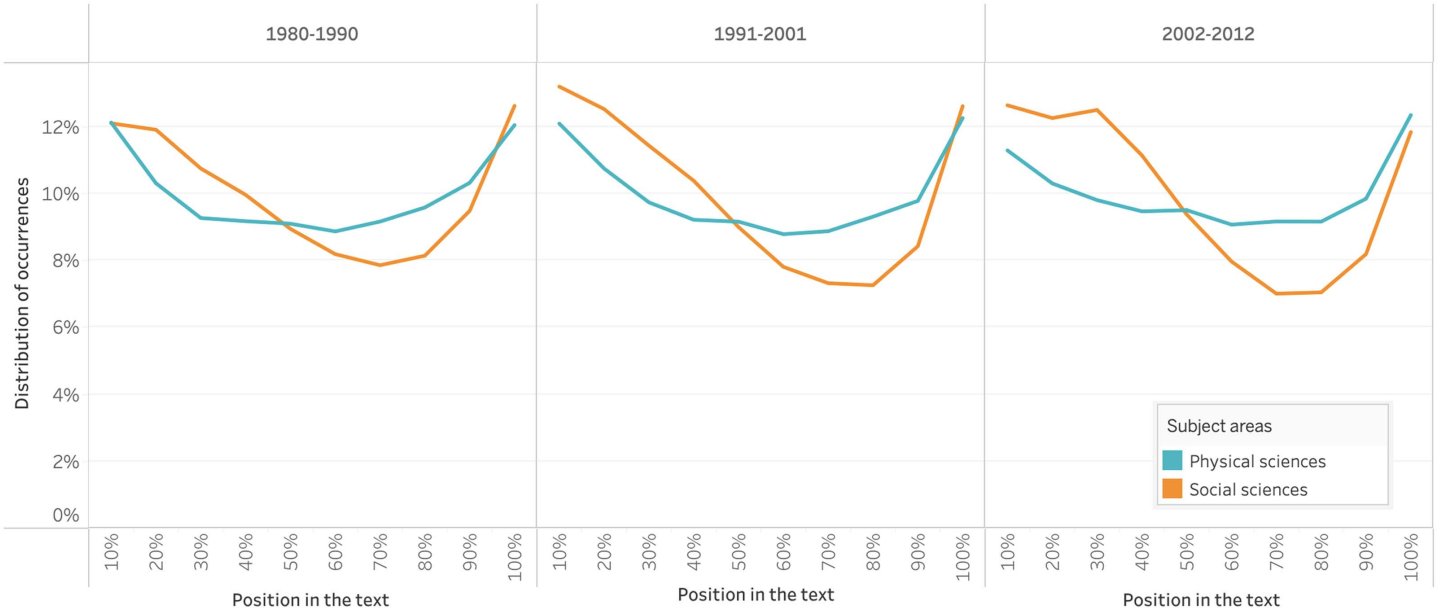


**Figure 2.** Evolution of the distribution of *data* occurrences throughout the article.

general statements. The end of articles is also often dedicated to discussion. It gives an opportunity to authors to step back and theorize the results obtained.

Figure 2 shows that this distribution has not changed much over time, reflecting the regularity of the structure of research articles over the period under study.

### 4.2. On the Contribution of Adjectives to the Definition of Data

When we see in the literature how difficult it is to define data, and that there is a consensus on their evolving nature, it becomes clear that it is relevant to study adjectives combined with *data* insofar as the role of the adjective is to modify the noun, by bringing further details. This is precisely what the authors need to specify the nature of the data they are talking about.

The word *data* can also be preceded by a noun (adjunct/attributive noun) which specifies its meaning in the manner of an adjective. But the numerous tests on our corpus lead us to prefer the analysis of adjectives and past participles. Indeed, the combination of these attributive nouns and *data* generates mainly types of data (e.g., *population data*, *morbidity data*, *pollen data*, *citation data*), leaving aside some of the nuances that the authors wish to specify concerning their data and that only adjectives can convey.

Without engaging in an in-depth linguistic study of the adjective in the English language, we propose a simple and intuitive classification, with seven semantic classes for adjectives combined with *data*. This proposal is the result of our reflection after a thorough examination of all our extractions and manual tagging of all lexical combinations. It does not derive from any other classification and is not intended to be applicable beyond co-occurrences with the word *data*.

- **Kinds of data.** We borrow the term *kind* from Morgan (2020), who uses it to refer to what is produced by different kinds of instruments, to be taken in a broad sense as it can also refer to strategies developed in the Social sciences. These kinds of data are revealed by

the properties conveyed by the adjectives. These properties might be general (*simulated, statistical, isotopic, global, random, theoretical, exploratory, archival, relational, raw, digital, analytical, synthetic*), or specific to science and technique generally speaking (*experimental, observational, empirical, structural, numerical, electronic*), or specific to a particular field (*thermodynamic, kinetic, nuclear, molecular, crystallographic, astrometric, photometric, seismic, atomic, meteorological, optical*).

- **Dimension.** These adjectives can be used to define the extent of the data ( *few, extensive, small, full, comprehensive*), but in a large majority of cases (*big, large, entire*), this dimension applies to a set designated by the names *bank, base, set, package*, for instance.
- **Spatiotemporal properties.** These adjectives provide a temporal information about the data (*new, actual, preliminary, initial, recent, current, present, existing, first, final, primary, secondary, historical*) or specify their spatial extent (*local, national, international, Brazilian, geocoded, western*).
- **Value.** The value of data might be subjective or at least results from the researcher's analysis. We only include in this category adjectives that convey a value (or lack of value) inherently, independently of any context (*accurate, correct, false, falsified, inaccurate, reliable, sufficient, limited, original, basic, relevant, high-quality, noisy, useful*).
- **Availability.** These adjectives provide information on the availability of data (*available, observed, additional, measured, corresponding, internal, published*), or on the contrary on the fact that they are missing for the researcher (*lossy, incomplete, inaccessible, insufficient, complete, missing, unpublished*).
- **Data to data comparison.** These adjectives qualify the data by comparing them to each other (*alternative, other* (it might be considered as a determiner as well), *same, conflicting, different*) or by placing them in relation to other data of the same group (*joint, various, mixed, homogeneous, combined, balanced*). In both cases, these same adjectives can be used for a comparison with a completely different data set.
- **Alpha-numerical adjectives.** We also make a special class of adjectives that have an alpha-numeric form. We are unable to qualify them with certainty because it is necessary to have both context and more detailed knowledge of the scientific field in question (e.g.: *1o-2 data, 1s data, 4dvar data, 18-winter data*)

We manually tagged into one of these seven classes all the adjectives for which we identified at least 20 occurrences in the whole corpus (i.e., 3,556 adjectives for 228,988 occurrences involving the word *data*) occurring before the bibliography section. The distribution is detailed in Table 2 and the most frequent adjectives by class and domain are presented in Figure 3.

The identification of adjectives and the fact that 3,556 different adjectives are used more than 20 times in the corpus show that it is by combining *data* with an adjective that the word *data* takes on its meaning. Adjectives provide the semantic precision that the word *data* alone lacks, giving rise to the abundant considerations in the literature in an attempt to provide a definition. To confirm this, we can see that adjectives expressing a kind of data are the most frequent with more than half of the total number of uses. This proportion is even more important in Physical sciences (59.2%, against 44.5% in Social sciences), but this difference is compensated by a more important proportion of adjectives expressing a spatiotemporal property in Social sciences (20.9% against 11.3% in Physical sciences). Secondly, adjectives related to the availability of data are used in more than 11% of the occurrences of *data*. The access to the data and their intrinsic characteristics thus seems to be more discussed than their size. Interestingly, while the issue of "big data" and its use in science is very important, this aspect does

**Table 2.** Distribution of *data* across the classes of adjectives it combines with

| | # occurrences | | % occurrences | | # occurrences | % occurrences |
|---|---|---|---|---|---|---|
| | Physical sciences | Social sciences | Physical sciences | Social sciences | All | All |
| **Kinds of data** | 78,745 | 43,570 | 59.2 | 44.5 | 121,358 | 53.0 |
| **Spatiotemporal properties** | 15,101 | 20,504 | 11.3 | 20.9 | 35,292 | 15.4 |
| **Availability** | 14,493 | 12,421 | 10.9 | 12.7 | 26,659 | 11.6 |
| **Value** | 10,351 | 9,944 | 7.8 | 10.2 | 20,085 | 8.8 |
| **Data to data** | 7,715 | 7,213 | 5.8 | 7.4 | 14,745 | 6.4 |
| **Dimension** | 4,334 | 3,568 | 3.3 | 3.6 | 7,822 | 3.4 |
| **Alpha-numerical** | 2,324 | 714 | 1.7 | 0.7 | 3,027 | 1.3 |
| **All** | **133,063** | **97,934** | **100.0** | **100.0** | **228,988** | **100.0** |

not emerge clearly in our corpus. Adjectives expressing the dimension of data sets constitute barely more than 3% of occurrences. Moreover, the phrase *big data* is almost absent from our corpus (around 0.001% of occurrences).

Figure 3 shows the 20 most frequent adjectives of each class by subject area; the percentage is calculated on the whole field.

The most frequent adjective in Physical sciences, is, by far, *experimental*, with 20.9% of all occurrences of adjectives in the domain (vs. 2% in Social sciences). The second most frequent adjective in Physical sciences, which is also the most frequent in Social sciences, is *available*, with around 4% of the occurrences in both disciplines.

The distribution of adjectives qualifying the word *data* throughout the text generally follows that of the word *data* itself (Figure 4). Nevertheless, it is interesting to note that the adjectives expressing relations between data have a flatter curve in Physical sciences, denoting the fact that the authors continuously assemble their data or balance them with other teams' data to progress in their demonstration, up to the results and conclusions. In contrast, in Social sciences, it is at the beginning and especially at the very end that these relationships are rather settled. This can be put in relation to the adjectives of availability, the use of which drops off rapidly at the beginning of the articles in Physical sciences, whereas it rebounds and lasts for at least a third of the text in Social sciences. Finally, it should be noted that the kinds of data are also expressed during the first third of the text in Social sciences, whereas these details seem to be dealt with more quickly in Physical sciences.

Our corpus allows us to generate Figure 5, which shows the evolution of the distribution of each type of adjective over time in each domain.

Even if we can see some oscillations (which may be due to the very nature of the ISTEX corpus), we cannot say that there is a clear shift over this 30-year period. However, if we look very closely, we can see that the curve for data to data adjectives in Social sciences also tends to flatten out, with a less sharp decrease in data to data adjectives over the first two-thirds of the articles. This trend is very subtle indeed and requires further investigation; it may be indicative of a more intensive use of quantitative approaches in Social sciences.

| Kinds of data | | | | Dimension | | | | Spatio-temporal properties | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Physical sciences** | | **Social sciences** | | **Physical sciences** | | **Social sciences** | | **Physical sciences** | | **Social sciences** | |
| experimental | 20.90% | empirical | 2.54% | large | 0.48% | large | 0.51% | initial | 2.11% | present | 1.87% |
| raw | 1.50% | experimental | 2.00% | few | 0.35% | entire | 0.29% | present | 1.37% | new | 1.23% |
| nuclear | 1.21% | qualitative | 1.68% | entire | 0.24% | full | 0.25% | new | 1.17% | longitudinal | 1.19% |
| observed | 1.07% | raw | 1.63% | full | 0.20% | few | 0.25% | actual | 0.46% | historical | 1.16% |
| spectral | 1.04% | quantitative | 1.24% | whole | 0.19% | extensive | 0.22% | historical | 0.46% | monthly | 0.87% |
| real | 0.87% | demographic | 1.00% | extensive | 0.19% | little | 0.22% | current | 0.38% | actual | 0.84% |
| simulated | 0.84% | aggregate | 0.86% | many | 0.18% | comprehensive | 0.18% | recent | 0.37% | annual | 0.84% |
| seismic | 0.82% | cross-sectional | 0.85% | total | 0.16% | whole | 0.18% | previous | 0.34% | quarterly | 0.82% |
| isotopic | 0.66% | observed | 0.85% | small | 0.16% | several | 0.16% | spatial | 0.31% | recent | 0.69% |
| empirical | 0.65% | statistical | 0.78% | several | 0.15% | small | 0.15% | preliminary | 0.28% | primary | 0.66% |
| structural | 0.65% | normative | 0.76% | little | 0.11% | total | 0.15% | existing | 0.24% | current | 0.62% |
| thermodynamic | 0.64% | numerical | 0.68% | comprehensive | 0.10% | many | 0.14% | local | 0.22% | secondary | 0.54% |
| spectroscopic | 0.61% | real | 0.67% | much | 0.07% | overall | 0.13% | primary | 0.20% | spatial | 0.51% |
| numerical | 0.61% | financial | 0.64% | overall | 0.07% | much | 0.12% | first | 0.17% | initial | 0.49% |
| kinetic | 0.61% | behavioral | 0.56% | long | 0.05% | large-scale | 0.06% | longitudinal | 0.17% | preliminary | 0.45% |
| quantitative | 0.56% | descriptive | 0.54% | numerous | 0.05% | considerable | 0.06% | second | 0.12% | daily | 0.40% |
| meteorological | 0.52% | detailed | 0.52% | extended | 0.04% | numerous | 0.03% | final | 0.12% | national | 0.36% |
| synthetic | 0.45% | individual | 0.50% | large-scale | 0.04% | cumulative | 0.03% | temporal | 0.11% | existing | 0.35% |
| observational | 0.45% | observational | 0.45% | maximum | 0.04% | short | 0.03% | secondary | 0.11% | previous | 0.26% |
| analytical | 0.45% | simulated | 0.45% | short | 0.04% | minimal | 0.03% | real-time | 0.11% | first | 0.25% |
| ... n=947 | | ... n=832 | | ... n=65 | | ... n=63 | | ... n=311 | | ... n=430 | |

| Value | | | | Availability | | | | Data to data | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Physical sciences** | | **Social sciences** | | **Physical sciences** | | **Social sciences** | | **Physical sciences** | | **Social sciences** | |
| original | 1.05% | relevant | 1.10% | available | 3.49% | available | 4.05% | other | 1.09% | other | 1.39% |
| relevant | 0.46% | original | 1.08% | unpublished | 1.09% | missing | 1.66% | same | 1.07% | same | 1.33% |
| noisy | 0.45% | basic | 0.49% | additional | 0.62% | complete | 0.89% | different | 0.54% | different | 0.68% |
| reliable | 0.38% | reliable | 0.49% | corresponding | 0.53% | unpublished | 0.79% | similar | 0.38% | comparable | 0.44% |
| basic | 0.33% | necessary | 0.42% | missing | 0.49% | additional | 0.79% | independent | 0.25% | similar | 0.39% |
| necessary | 0.32% | appropriate | 0.32% | measured | 0.46% | limited | 0.40% | various | 0.21% | comparative | 0.36% |
| accurate | 0.25% | useful | 0.22% | limited | 0.42% | incomplete | 0.39% | combined | 0.20% | relational | 0.24% |
| appropriate | 0.21% | objective | 0.22% | supplementary | 0.41% | sufficient | 0.35% | relational | 0.15% | various | 0.21% |
| useful | 0.17% | accurate | 0.21% | complete | 0.40% | own | 0.33% | parallel | 0.14% | combined | 0.21% |
| exact | 0.14% | consistent | 0.20% | sufficient | 0.24% | corresponding | 0.23% | relative | 0.12% | independent | 0.18% |
| precise | 0.13% | official | 0.19% | published | 0.19% | insufficient | 0.22% | comparable | 0.11% | correlational | 0.16% |
| good | 0.11% | rich | 0.18% | insufficient | 0.18% | enough | 0.19% | related | 0.10% | separate | 0.14% |
| common | 0.11% | valid | 0.18% | enough | 0.17% | external | 0.13% | comparative | 0.10% | related | 0.12% |
| suitable | 0.11% | important | 0.17% | incomplete | 0.16% | sparse | 0.12% | separate | 0.07% | diverse | 0.09% |
| consistent | 0.10% | good | 0.16% | own | 0.14% | internal | 0.12% | scattered | 0.07% | relative | 0.09% |
| selected | 0.10% | adequate | 0.15% | recorded | 0.14% | supplementary | 0.12% | binding | 0.06% | twin | 0.07% |
| unique | 0.10% | noisy | 0.13% | due | 0.14% | online | 0.12% | ancillary | 0.06% | dichotomous | 0.06% |
| important | 0.09% | subjective | 0.13% | surrogate | 0.11% | due | 0.11% | composite | 0.05% | alternative | 0.06% |
| simple | 0.09% | pertinent | 0.11% | sparse | 0.10% | possible | 0.09% | auxiliary | 0.05% | conflicting | 0.06% |
| critical | 0.09% | common | 0.11% | internal | 0.09% | usable | 0.08% | hierarchical | 0.04% | differenced | 0.05% |
| ... n=201 | | ... n=201 | | ... n=75 | | ... n=74 | | ... n=84 | | ... n=87 | |

**Figure 3.** Percentage of occurrences of the 20 most frequent adjectives combined with *data*, by adjective class and subject area (see full list and reusable version in the available data set (Bordignon & Maisonobe, 2022)).

### 4.3. Data Transformation Through Verb Usage

As many studies have shown that data undergo transformations, we assume that it is by identifying verbs whose subjects are the authors of the articles (using *we* or *I*) that we will be able to have a better picture of how data are used and how transformations are performed. We build a query for this purpose, taking care to avoid retrieving negative forms that reflect the opposite of

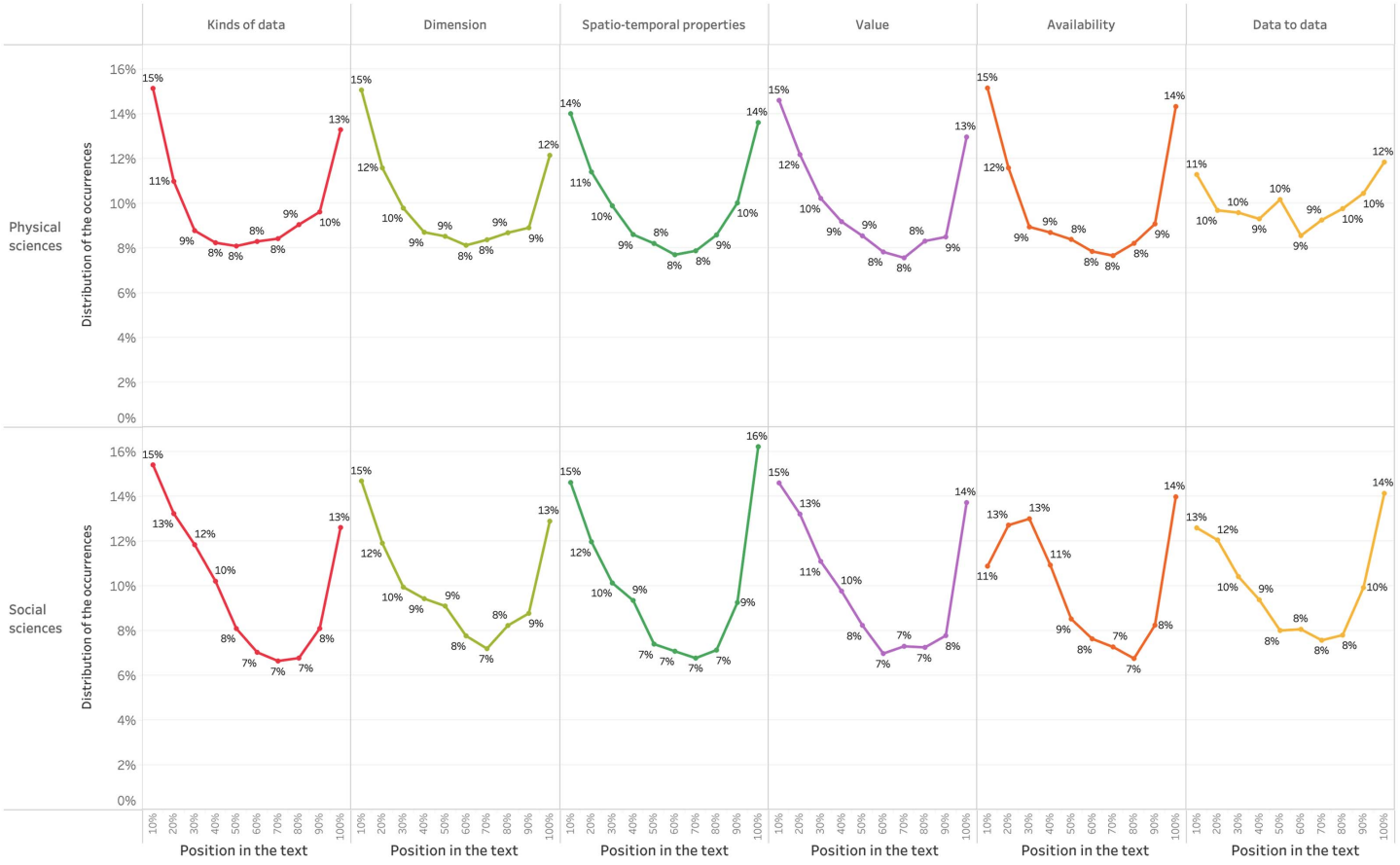


**Figure 4.** Distribution of *data* throughout the text, represented by the adjective classes it combines with and the subject areas.

what we are looking for. Our query also allows us to identify more complex inverted forms such as "*the data we analyzed*." We also search for past participles, which are somehow halfway between the adjectives from which they take on the function of qualification (notably after the verb *to be*), and the verbs from which they are lexically derived. To process the results accurately, we homogenize some equivalent forms but whose spellings may differ (e.g., *analyse* and *analyze*).

As for adjectives qualifying *data*, we propose a classification of the verbs and past participles combined with *data*. We propose three main classes that mirror three phases of the data scholarship process: assemblage, analysis, and inscription.

- **Assemblage.** Data assemblage verbs and participles are those allowing authors to explain how they carried out the delineation and curation of their data. We borrow the term from Law (2004), who presents the concept of method assemblage as "the process of enacting or crafting bundles of ramifying relations that condense presence and (therefore also) generate absence by shaping, mediating and separating these." Data assemblage is displayed through verbs related to data collection (*collect*, *gather*, *accumulate*), to their selection (*filter*, *remove*, *truncate*), and to their arrangement (*merge*, *aggregate*, *associate*). Some rare verbs within this class express the fact of possessing these collected data (*own*, *have*, *hold*). At this stage, data are inputs and are interrelated, echoing the data to data adjectives.

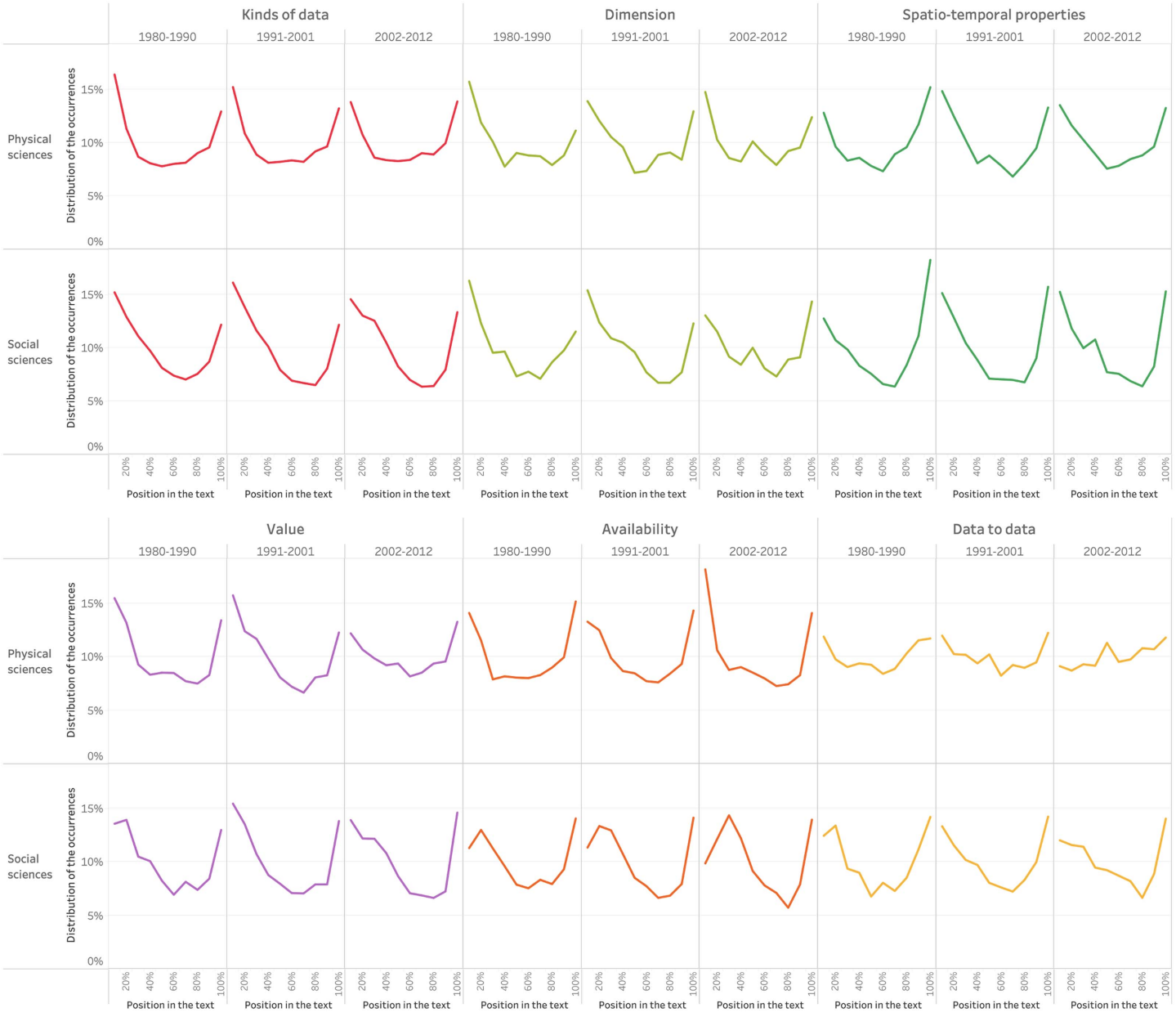


**Figure 5.** Evolution of the distribution of *data* throughout the text, according to the adjective classes it combines with and the subject areas.

- **Analysis.** With data analysis verbs and the use of the passive voice, authors explain the analysis carried out with the data and the associated methodology. The data are seen from afar: They are studied (or *examined, compared, interpreted*) or evaluated (or *assessed, checked, verified*) thus reminding us of value adjectives.
- **Inscription.** In this third class, verbs and participles denote inscriptions, in Latourian terms (Latour, 1999): They refer to "the transformations through which data become materialized into a sign" (*list, normalize, plot, code, compute*). They can also refer to the presentation of these inscriptions ( *present, show, report, publish*). The data are then outputs.

**Table 3.** Distribution of *data* according to verbs and past participles it combines with in the two subject areas

| | | Physical sciences | | | Social sciences | | | |
|---|---|---|---|---|---|---|---|---|
| | | Past participles | Verbs | Subtotal | Past participles | Verbs | Subtotal | All |
| **Assemblage** | **Number of occurrences** | 29,881 | 1,274 | **31,155** | 24,553 | 1,998 | **26,551** | **57,263** |
| | **Percentage by subject area** | 95.9 | 4.1 | **100** | 92.5 | 7.5 | **100** | – |
| **Inscription** | **Number of occurrences** | 32,665 | 1,422 | **34,087** | 18,842 | 1,399 | **20,241** | **54,004** |
| | **Percentage by subject area** | 95.8 | 4.2 | **100** | 93.1 | 6.9 | **100** | – |
| **Analysis** | **Number of occurrences** | 14,886 | 1,953 | **16,839** | 14,915 | 3,073 | **17,988** | **34,561** |
| | **Percentage by subject area** | 88.4 | 11.6 | **100** | 82.9 | 17.1 | **100** | – |
| **Others** | **Number of occurrences** | 433 | 36 | **469** | 268 | 69 | **337** | **799** |
| | **Percentage by subject area** | 92.3 | 7.7 | **100** | 79.5 | 20.5 | **100** | – |
| **All** | **Number of occurrences** | **77,865** | **4,685** | **82,550** | **58,578** | **6,539** | **65,117** | **146,627** |
| | **Percentage by subject area** | **94.3** | **5.7** | **100** | **90.0** | **10.0** | **100** | – |

We have unified verbs and past participles under their corresponding lemma (e.g., *obtain* and *obtained* have *obtain* as lemma). We finally left out the phrases with the term *set*, because the sequence *data set* (supposed to be equivalent to *dataset* in one word) was problematic when tagged in parts-of-speech (*set* being then often wrongly considered as a participle; e.g., "*the results were used as a learning data set for a predictor of peptide detectability*"). And eventually, we excluded the different variants of *do* and *make*, which do not bring any information on the action performed on the data itself.

We classified all verbs and participles with more than 20 occurrences (in the whole corpus) into one of the three categories (i.e., 457 terms for 146,627 occurrences involving the word *data*, occurring before the bibliography section). Past participles account for more than 90% of the occurrences (Table 3), consistent with the fact that the passive voice suggests the researcher's actions that the data undergo. The distribution by subject area across classes is detailed in Table 4 and the most frequent forms associated are presented in Figure 6 by class and domain.

**Table 4.** Distribution by percentage of occurrences of *data* across the classes of verbs/past participles in the two subject areas

| | Physical sciences | Social sciences | All |
|---|---|---|---|
| **Assemblage** | 37.7 | 40.8 | 39.1 |
| **Analysis** | 20.4 | 27.6 | 23.6 |
| **Inscription** | 41.3 | 31.1 | 36.8 |
| **Others** | 0.6 | 0.5 | 0.5 |
| **All** | **100.0** | **100.0** | **100.0** |

| Assemblage | | | | Analysis | | | | Inscription | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Physical sciences | | Social sciences | | Physical sciences | | Social sciences | | Physical sciences | | Social sciences | |
| obtain/ obtained | 10.52% | collect/ collected | 14.36% | use/ used | 6.65% | use/ used | 9.36% | present/ presented | 4.70% | present/ presented | 4.67% |
| collect/ collected | 6.51% | obtain/ obtained | 6.05% | analyse/ analysed | 2.07% | analyse/ analysed | 4.24% | show/ showed | 4.23% | report/ reported | 1.94% |
| take/ taken | 3.36% | gather/ gathered | 2.32% | compare/ compared | 1.21% | base/ based | 1.40% | measure/ measured | 3.51% | generate/ generated | 1.52% |
| give/ given | 3.22% | provide/ provided | 1.85% | base/ based | 0.78% | examine/ examined | 0.83% | report/ reported | 2.00% | show/ showed | 1.40% |
| fit/ fitted | 1.22% | give/ given | 1.54% | find/ found | 0.72% | consider/ considered | 0.73% | publish/ published | 1.75% | derive/ derived | 1.13% |
| acquire/ acquired | 1.21% | take/ taken | 1.39% | consider/ considered | 0.65% | find/ found | 0.64% | record/ recorded | 1.56% | publish/ published | 1.09% |
| require/ required | 1.03% | need/ needed | 1.06% | discuss/ discussed | 0.46% | discuss/ discussed | 0.64% | generate/ generated | 1.42% | draw/ drawn | 0.89% |
| provide/ provided | 0.95% | have/ had | 1.06% | determine/ determined | 0.35% | compare/ compared | 0.60% | plot/ plotted | 1.18% | record/ recorded | 0.88% |
| need/ needed | 0.87% | require/ required | 1.01% | relate/ related | 0.35% | relate/ related | 0.60% | store/ stored | 0.99% | describe/ described | 0.73% |
| gather/ gathered | 0.69% | aggregate/ aggregated | 0.66% | interpret/ interpreted | 0.33% | employ/ employed | 0.53% | derive/ derived | 0.87% | store/ stored | 0.61% |
| sample/ sampled | 0.51% | include/ included | 0.56% | evaluate/ evaluated | 0.32% | subject/ subjected | 0.39% | calculate/ calculated | 0.76% | transform/ transformed | 0.49% |
| include/ included | 0.44% | come | 0.43% | observe/ observed | 0.30% | interpret/ interpreted | 0.34% | correct/ corrected | 0.63% | code/ coded | 0.46% |
| have/ had | 0.41% | combine/ combined | 0.39% | know/ known | 0.29% | know/ known | 0.25% | describe/ described | 0.60% | compile/ compiled | 0.45% |
| transfer/ transferred | 0.34% | group/ grouped | 0.39% | sense/ sensed | 0.29% | see/ seen | 0.24% | process/ processed | 0.57% | express/ expressed | 0.44% |
| combine/ combined | 0.31% | limit/ limited | 0.37% | examine/ examined | 0.24% | utilize/ utilized | 0.24% | list/ listed | 0.52% | summarize/ summarized | 0.42% |
| select/ selected | 0.29% | pool/ pooled | 0.33% | see/ seen | 0.21% | observe/ observed | 0.22% | represent/ represented | 0.51% | average/ averaged | 0.42% |
| limit/ limited | 0.29% | enter/ entered | 0.32% | treat/ treated | 0.21% | evaluate/ evaluated | 0.21% | summarize/ summarized | 0.51% | produce/ produced | 0.40% |
| associate/ associated | 0.26% | contain/ contained | 0.29% | employ/ employed | 0.20% | treat/ treated | 0.21% | normalize/ normalized | 0.48% | measure/ measured | 0.39% |
| receive/ received | 0.24% | acquire/ acquired | 0.28% | assume/ assumed | 0.19% | test/ tested | 0.18% | transform/ transformed | 0.45% | extract/ extracted | 0.38% |
| supply/ supplied | 0.24% | supply/ supplied | 0.28% | read | 0.19% | submit/ submitted | 0.18% | distribute/ ditributed | 0.43% | plot/ plotted | 0.35% |
| … n=102 | | … n=101 | | … n=133 | | … n=133 | | … n=217 | | | … n=214 |

**Figure 6.** Percentage of occurrences of the 20 most frequent verbs/past participles combined with *data*, by class and subject area (see full list and reusable version in the available data set (Bordignon & Maisonobe, 2022)).



Figure 6 displays the 20 most frequent forms of each class by field; the percentage is calculated on the whole field.

The verbs *obtain* and *collect*, expressing the task of data collection, are by far the most used in both subject areas. Next are the verbs *use* and *analyze*, as well as *present* and *show* (especially in Physical sciences). But these frequently used verbs are lacking specificity. To specify their actions, the authors then resort to other verbs that are more precise. But still, the class of assemblage verbs is the smallest, with hardly more than 100 variants. This means that the data assemblage phase is underdescribed. It is in the class of inscriptions that we find the most variants, more than 200, which describe the inscriptions generated.

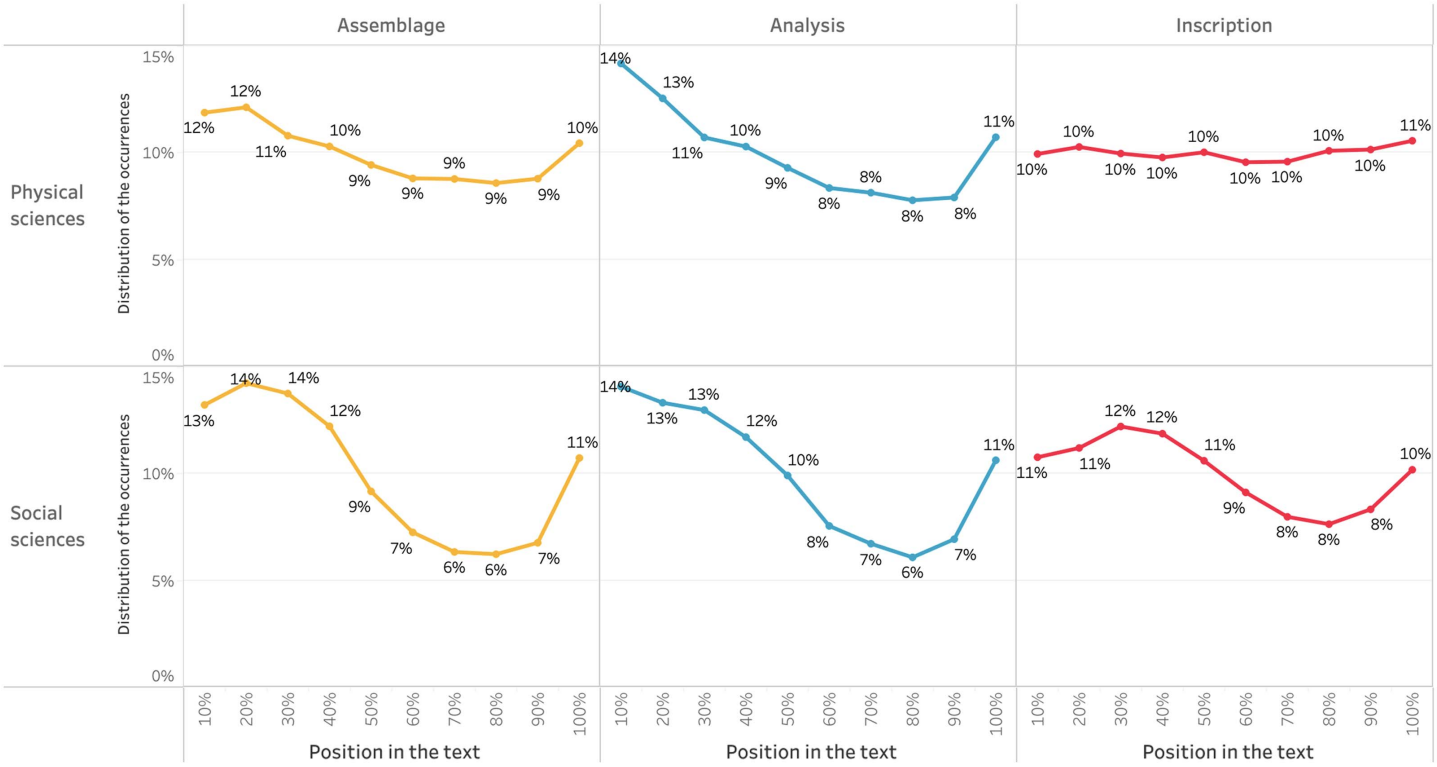


**Figure 7.** Distribution of the occurrences of *data* throughout the text, represented by the verbs/past participles classes it combines with and the subject areas.

Figure 7 reveals an important difference between the two domains: The curves in Social sciences follow those of the distribution of *data* in the text, whereas they are different in Physical sciences. Indeed, we can see that even if the actions of assemblage are more frequent at the beginning and the end of the text, the curve tends to flatten with occurrences that are distributed on the whole article. The homogenization of the distribution is even more striking for the inscriptions, with a nearly flat curve. This reveals that physical scientists have a permanent and continuous use of data, generate new inscriptions, and need to rearrange their data to carry out their research and thus their demonstration in the text.

This result is consistent with the findings in Figure 4 with a similar distribution for the data to data adjectives; this means that assemblage operations and inscriptions are recurrent throughout the text. Figure 8 shows that these practices have not changed much over time.

### 4.4. Data Ownership and Possessive Pronouns

The literature review revealed that the work around data, from collection to interpretation, is important and value-creating. Bearing in mind that researchers gain recognition with their data, we looked for expressions of ownership of data in texts through possessive pronouns by isolating the use of *our/my data*.

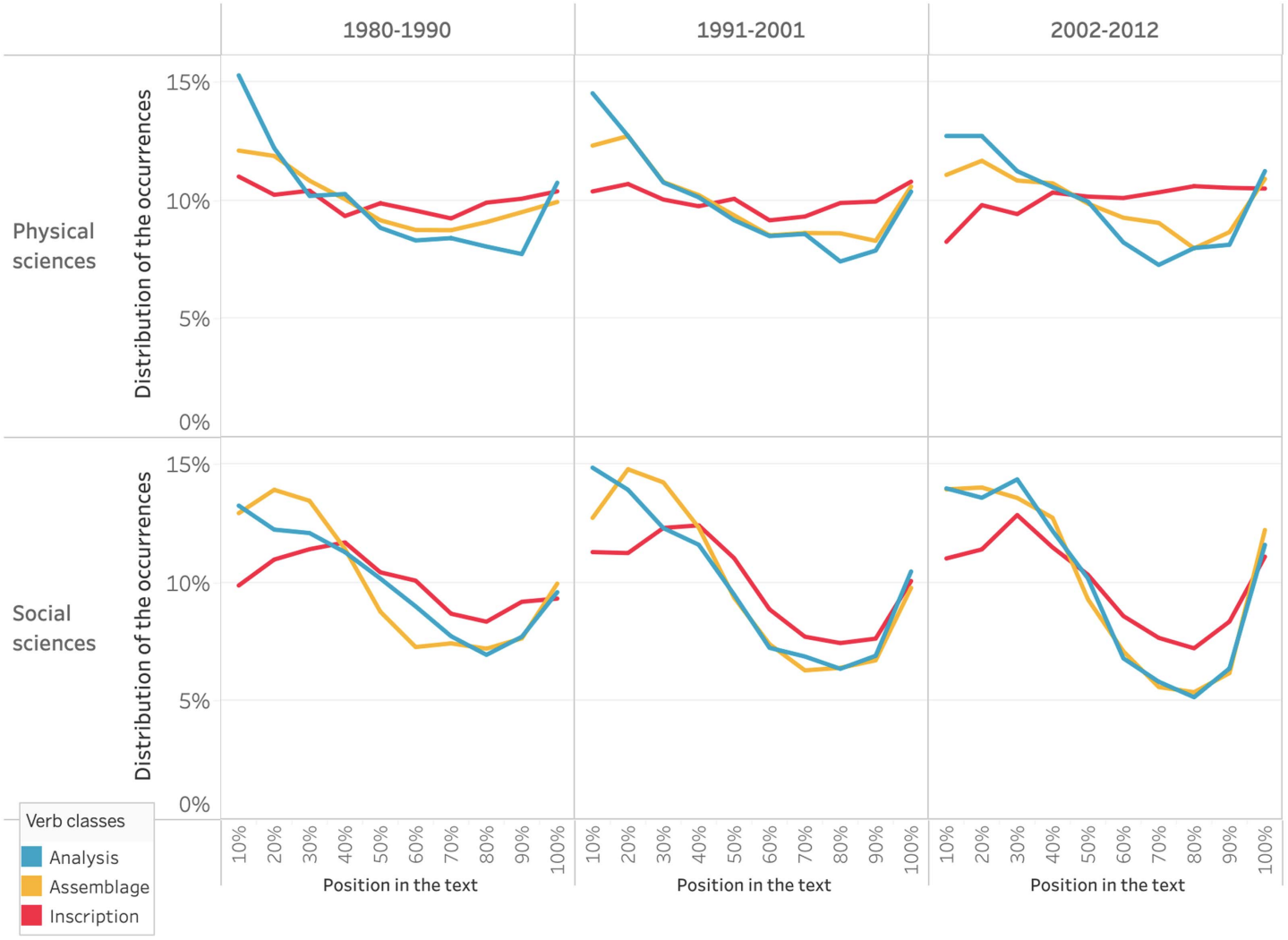


**Figure 8.** Evolution of the distribution of *data* throughout the text, according to the verb/past participles classes it combines with and the subject areas.



We found 23,102 occurrences of possessive pronouns used before the word *data* and recorded before the bibliography. Of these, 15,481 occurrences are related to the pronouns *my* or *our*, or 67%.

Figure 9 shows very clearly that the authors appropriate the data throughout the article, as if after having used them, they finally allow themselves to assert that they are the owners and that it is with this status that they mention them in the conclusion of their article. This is probably also a way of "sharing" them (at least the associated results) within the community by insisting on who owns them, and the value added by the analysis.

Authors anticipate the future credit that readers would have to give them by citing the conclusions of their research, and sometimes the data themselves. But at the same time, if *my/our data* appears as the subject (typically in "*our data suggest* ..." or "*our data show* ..."), it helps

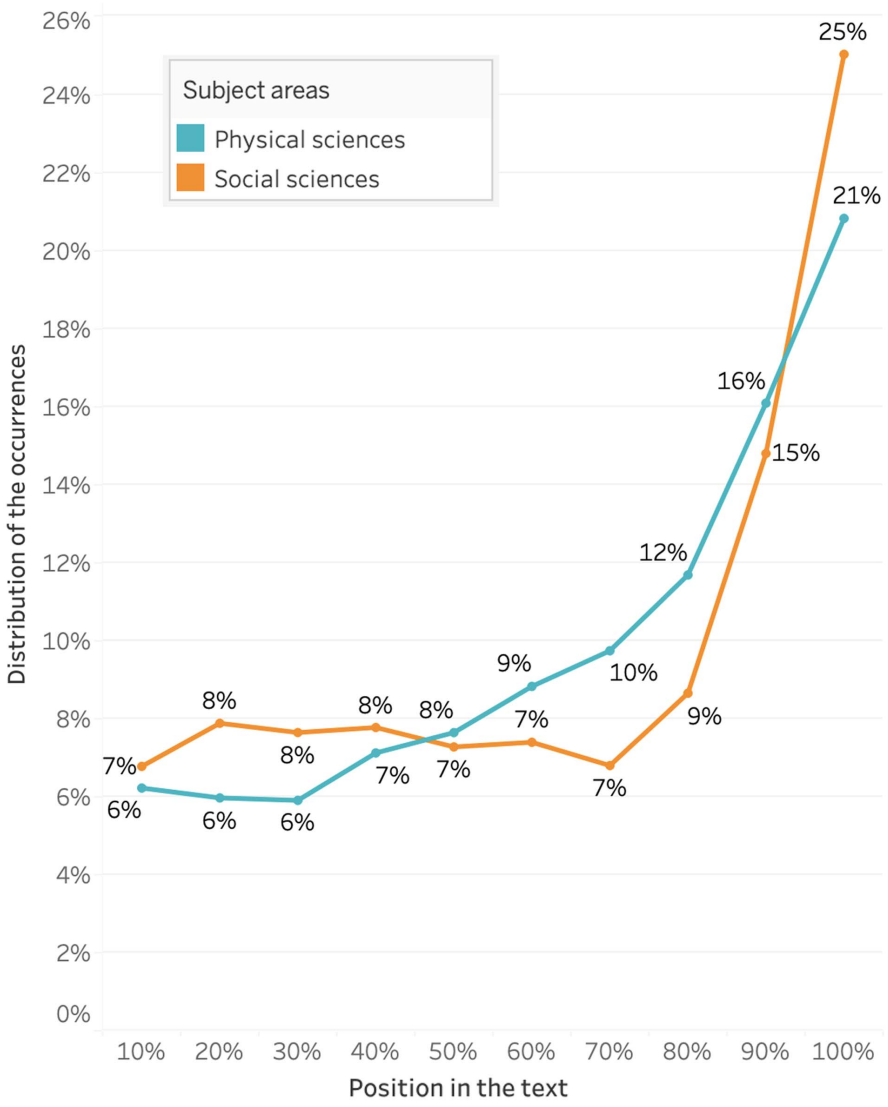


**Figure 9.** Distribution of *my/our data* throughout the text in the two subject areas.



authors distance themselves, shift responsibility to the data (as Salager-Meyer (1999) states it) with a typical turn of phrase where some nonhuman entity such as data actually speaks for itself. Therefore, they anticipate criticism as they progress in presenting their results. As the article reaches its conclusion, authors have an interest in gaining confidence and may seek to emphasize the personal and specific contribution to the literature that the article and its results represent.

Figure 10 shows the evolution over the three decades we are studying. Without showing major changes, it nevertheless reveals that in the most recent period, the occurrences distribution curve in Social sciences increasingly looks like that of Physical sciences with a less marked growth at the end of the text.

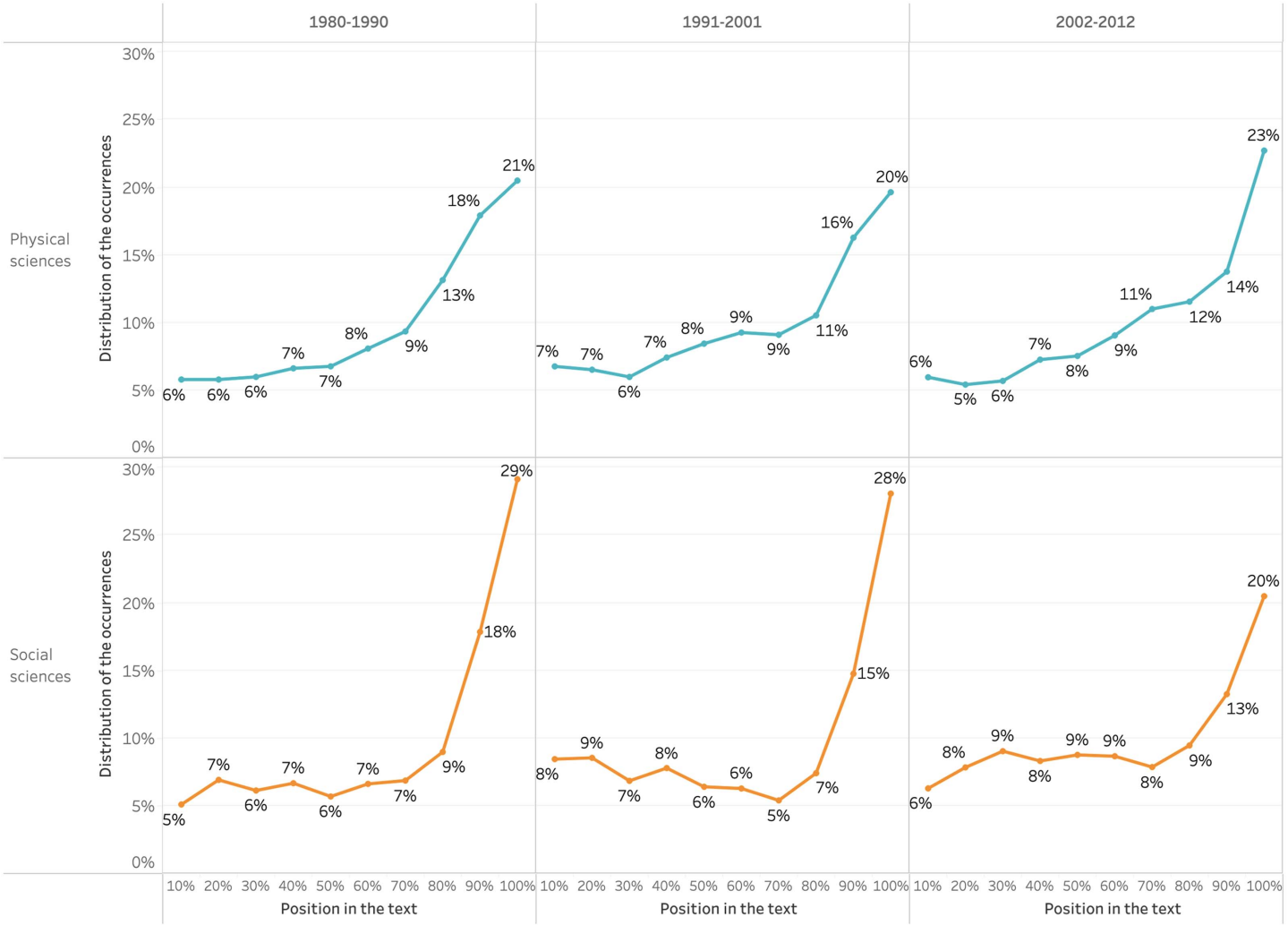


**Figure 10.** Evolution of the distribution of *my/our data* throughout the text in the two subject areas.

## 5. DISCUSSION

As our literature review at the beginning of this article demonstrates, we set our study in a theoretical framework that takes scientific writings as research objects and considers that their linguistic analysis can inform the research process itself, even if only partially. Our approach is intended to complement sociological or ethnographic studies that point out the difficulty of defining what data are and which place them in a transformation process. Therefore, by choosing to work on adjectives and verbs, we are able to provide results that confirm both the polysemous nature of the word *data* and the difficulty of defining it. But above all, our study draws on a large corpus of scientific articles overlapping two disciplinary fields and thus sheds new light on this polysemy by providing a large number of examples of lexical combinations, which we present with typologies of adjectives and verbs/past participles to help us understand the nature of data.

Our findings confirm the transformation process but also that the task of assemblage is poorly rewarded, as shown by previous studies whose scope is wider because they deal with

data curation, mainly in archives, whereas our focus here is rather on what researchers decide to mention in the articles they publish. We believe, however, that the large number of words we have extracted and which highlight these transformation processes can contribute to the works related to the description of data curation and data reuse. Indeed, while many instances of data use that we have identified in the articles may be instances of primary data use (i.e., researchers describe collecting, manipulating, and analyzing their own data), researchers may also reuse data created by others that have been made available through data repositories. The lists of terms we provide help to fuel studies on these topics and to enrich, with a greater level of detail, the Data Practices and Curation Vocabulary (a shared vocabulary (Chao, Cragin, & Palmer, 2015), limited to 187 terms, and intended for use by data curators and other persons involved in the curation process (including preservation, sharing or management)).

In line with the principles of epistemetrics, we have built up a corpus that allows for a comparative analysis between fields, a diachronic analysis and an analysis along the text, which we suggest reflects the "data journey" agreed upon by previous studies. Despite the increasing use of the word *data* in titles and abstracts, our research shows a significant regularity over the period studied in the way researchers refer to data within scholarly articles. In particular, authors prefer to use this word at the beginning and the end of articles. This predominance is consistent with other findings suggesting (Master, 1991) that it is in the opening and concluding sections of research articles that we find phrases involving more abstract and general concepts, as data can be.

We have identified a few differences between the two subject areas, and those differences continue over time:

- data and their characteristics are discussed longer at the beginning of Social sciences articles;
- in Physical sciences only, assemblage operations and inscriptions are recurrent throughout the text; and
- there is more description of the spatiotemporal properties of data and more discussion of their availability in Social sciences and these features continue over time.

In both subject areas, we observe the same tendency to take more ownership of the data throughout the narrative, which is illustrated by the more frequent use of the possessives *my* or *our* at the end of the articles, actually reaching the end of the “data journey.” We have seen that the trend in Social sciences to appropriate data at the very end of the text has changed in the most recent period, with a redistribution of some of the uses of *my/our* earlier in the text. This is a discourse pattern similar to the one we have in Physical sciences. It is difficult to draw conclusions about a possible change in the behavior of social scientists towards *their* data, and it would be worth organizing interviews to clarify this. Our work opens up here an opportunity for further research.

The verb analysis confirms the seminal research of Knorr and Knorr (1978) on the topic: The complex and painful task of data acquisition is somewhat absent from articles, even if assemblage verbs are the most frequent. They refer to assemblage operations but with rather imprecise verbs, such as *obtain* and *collect*. It is in the class of inscriptions that we find the most variants, meaning authors prefer to focus on their results, to produce their own new inscriptions and thus new data, and to claim ownership of them at the end of the process. This is what leads us to think that it would be interesting to do the same analysis on another kind of scientific output, the data paper, to see if other trends emerge.

Regarding the limits of this study, we can say that it would have been interesting to investigate on words belonging to the data family, such as *database* and *dataset*. This is something we are considering for future work, but it made sense to start by investigating the word *data* to grasp its polysemy, before studying derived words such as *database* or *dataset*. Moreover, the work of indexing and tagging terms in such a large corpus is very time-consuming and the many necessary controls are tedious. As for the ISTEX database, it should be noted that it does not cover recent publications and that its coverage depends on agreements with publishers so that the types of analyses that can be derived from it are limited. Given the evolution of the coverage, we also need to be cautious with the interpretations of the longitudinal observations. Nevertheless, by using a random sampling method, we believe that we have somehow managed to address this coverage pitfall. Consequently, we found that ISTEX is a resource that can be used if care is taken to extract a fairly large corpus and to ensure its homogeneity, as we have done. On the other hand, getting down to a finer level of detail, such as the subfield, is probably riskier because the inclusion of a single journal, for example, can unbalance the content's homogeneity.

Because we worked on equal-sized sets of publications for each subperiod studied, our study makes it difficult to perceive the changes and the enthusiasm of the 2000s for data and more precisely for *big data*. As this phrase hardly emerges from our corpus, our analysis has the advantage of showing that, when it comes to looking at research as a whole, the use of the word and its context of occurrence have not radically changed over the period under study. Data scholarship is an intrinsic part of research activity, both in Social sciences and Physical sciences, and this regularity as well as the frequency of occurrences of the word in the two subject areas reflect this clearly.

## 6. CONCLUSION

To conclude, the results of our corpus-based study demonstrate that there is no point in trying to define the word *data*, as researchers do it themselves in the course of their writing, taking Firth's idea to its fullest extent: "You shall know a word by the company it keeps" (Firth, 1957). The word *data* serves as a rhetorical base and draws on the context for its meaning, relying on the properties conveyed by adjectives and verbs associated to it. Adjectives and verbs accompanying the noun *data* turn out to be even more important than *data* itself in specifying what data are at stake. And to echo Gitelman (2013), we can say that while data can never be raw, the word *data* is, and it only serves as a rhetorical basis, as long as the context and mainly adjectives have not contributed to achieve its potential with the properties they convey.

## AUTHOR CONTRIBUTIONS

Frédérique Bordignon: Data curation, Investigation, Methodology, Validation, Visualization, Writing—Original draft, Writing—Review & editing. Marion Maisonobe: Conceptualization, Investigation, Methodology, Validation, Writing—Original draft, Writing—Review & editing.

## COMPETING INTERESTS

The authors have no competing interests.

## FUNDING INFORMATION

The research project received no funding by third parties.

## DATA AVAILABILITY

The data used in this study are available in a repository (Bordignon & Maisonobe, 2022).

## REFERENCES


Bazerman, C. (1981). What written knowledge does: Three examples of academic discourse. *Philosophy of the Social Sciences*, *11*(3), 361–387. https://doi.org/10.1177/004839318101100305

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*, *67*(1), 164–177. https://doi.org/10.1002/asi.23367

Bordignon, F., & Maisonobe, M. (2022). Data for "Researchers and data. A study through the use of the word data in scholarly articles." *Zenodo*. https://doi.org/10.5281/zenodo.5873829

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2015). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology*, *66*(3), 616–633. https://doi.org/10.1002/asi.23184

Chen, P. C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Dazy, A. (2014). ISTEX: A powerful project for scientific and technical electronic resources archives. *Insights*, *27*(3), 269–273. https://doi.org/10.1629/2048-7754.157

Demarest, B., & Sugimoto, C. R. (2015). Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*, *66*(7), 1374–1387. https://doi.org/10.1002/asi.23271

Denis, J., & Goëta, S. (2017). Rawification and the careful generation of open government data. *Social Studies of Science*, *47*(5), 604–629. https://doi.org/10.1177/0306312717712473, PubMed: 28633611

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, *5*(1).

Elsevier. (2021). *What are the most used Subject Area categories and classifications in Scopus?—Scopus: Access and use Support Center*. https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/

Firth, J. R. (1957). *A synopsis of linguistic theory, 1930–1955*. Oxford: Basil Blackwell.

Gitelman, L. (Ed.). (2013). *'Raw data' is an oxymoron*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM: Une plateforme logicielle open-source pour la textométrie—conception et développement. *10th International Conference on the Statistical Analysis of Textual Data—JADT 2010*, *2*(3), 1021–1032.

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data access, ownership, and control: Toward empirical studies of access practices. *Knowledge*, *15*(4), 355–372. https://doi.org/10.1177/107554709401500401

Hsiao, T.-K., & Schneider, J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies*, *2*(4), 1144–1169. https://doi.org/10.1162/qss_a_00155, PubMed: 36186715

Huang, Y., Schuehle, J., Porter, A. L., & Youtie, J. (2015). A systematic method to create search strategies for emerging technologies based on the Web of Science: Illustrated for 'Big Data'. *Scientometrics*, *105*(3), 2005–2022. https://doi.org/10.1007/s11192-015-1638-y

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

Knorr, K. D., & Knorr, D. W. (1978). *From scenes to scripts: On the relationships between laboratory research and published paper in science*. Wien: IHS.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge, MA: Harvard University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400820412

Law, J. (2004). *After method: Mess in social science research*. London: Routledge. https://doi.org/10.4324/9780203481141

Leonelli, S. (2020). Learning from data journeys. In S. Leonelli & N. Tempini (Eds.), *Data journeys in the sciences* (pp. 1–24). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-37177-7_1

Longino, H. E. (2020). Afterword: Data in transit. In S. Leonelli & N. Tempini (Eds.), *Data journeys in the sciences* (pp. 391–399). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-37177-7_20

Maisonobe, M. (2022). The future of urban models in the Big Data and AI era: A bibliometric analysis (2000–2019). *AI & SOCIETY*, *37*(1), 177–194. https://doi.org/10.1007/s00146-021-01166-4

Markham, A. N. (2013). Undermining 'data': A critical examination of a core term in scientific inquiry. *First Monday*. https://journals.uic.edu/ojs/index.php/fm/article/view/4868

Martin-Scholz, A. (2017). *Communiquer et organiser en échafaudant un 'big data' manuel. Le cas d'un projet de formation en aménagement du territoire* [Phd, Université de Toulouse, Université Toulouse III - Paul Sabatier]. https://thesesups.ups-tlse.fr/3803/

Master, P. (1991). Active verbs with inanimate subjects in scientific prose. *English for Specific Purposes*, *10*(1), 15–33. https://doi.org/10.1016/0889-4906(91)90013-M

Morgan, M. S. (2020). The datum in context: measuring frameworks, data series and the journeys of individual datums. In S. Leonelli & N. Tempini (Eds.), *Data journeys in the sciences* (pp. 103–120). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-37177-7_6

Mullins, N., Snizek, W., & Oehler, K. (1988). The structural analysis of a scientific paper. In A. F. J. Van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 81–105). Elsevier. https://doi.org/10.1016/B978-0-444-70537-2.50008-8

Perret, A., & Le Deuff, O. (2019). Documentarité et données, instrumentation d'un concept. *12ème Colloque International d'ISKO-France: Données et Mégadonnées Ouvertes En SHS: De Nouveaux Enjeux Pour l'état et l'organisation Des Connaissances?*

Plantin, J.-C. (2019). Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values*, *44*(1), 52–73. https://doi.org/10.1177/0162243918781268

Rosenberg, D. (2013). Data before the fact. In L. Gitelman (Ed.), *'Raw data' is an oxymoron* (pp. 15–40). Cambridge, MA: MIT Press.

Rosenberg, D. (2018). Data as word. *Historical Studies in the Natural Sciences*, *48*(5), 557–567. https://doi.org/10.1525/hsns.2018.48.5.557

Ruppert, E. S., & Scheel, S. (2021). *Data practices: Making up a European people*. Goldsmiths Press.

Salager-Meyer, F. (1999). From "Mr. Guthrie is profoundly mistaken..." to "Our data do not seem to confirm the results of a previous study on...": A diachronic study of polemicity in academic writing (1810–1995). *Ibérica, Revista de La Asociación Europea de Lenguas Para Fines Específicos*, *1*, 5–28.

Smolczewska Tona, A. (2021). Enquête sur les représentations discursives des temporalités de la donnée à l'œuvre dans des articles scientifiques. In B. Simonnot, E. Broudoux, & G. Chartron (Eds.), *Humains et données: Création, médiation, décision, narration—Actes du colloque 'Document numérique et société', Nancy, 2020*.

Swales, J. (1988). Discourse communities, genres and English as an international language. *World Englishes*, *7*(2), 211–220. https://doi.org/10.1111/j.1467-971X.1988.tb00232.x

Tenopir, C. (2016). Big data, little data, no data: Scholarship in the networked world by Christine L. Borgman. Cambridge, MA: MIT Press, 2015. 400 pp. (ISBN 9780262028561). *Journal of the Association for Information Science and Technology*, *67*(3), 751–753. https://doi.org/10.1002/asi.23626

Terrier, C. (2011). The value of the geographical data. *L'espace géographique*, *40*(2), 103–108. https://doi.org/10.3917/eg.402.0103

Walford, A. (2013). *Transforming data: An ethnography of scientific data from the Brazilian Amazon* [PhD thesis]. University of Copenhagen.

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, *58*(4), 479–493. https://doi.org/10.1002/asi.20508