IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 001, LIR010
DOI: https://doi.org/10.52731/lir.v001.010

Characteristic Analysis of Data Description in Highly Cited Research Data

Naoto Kai [*], Toshiki Shimbaru [†]

Abstract

The publication and utilization of not only evidence data of research results but also many other data associated with research activities are attracting attention as a new indicator for improving research efficiency and evaluating researchers. With the spread of open science, data repositories and data journals for research data have gradually begun to be recognized, and an environment that encourages the reuse of such research data will be further developed in the future. In this study, we focused on data description, which is important for getting an overview of research data. By clarifying the characteristics of data descriptions of highly cited research data, we hope to increase the number of citations, which will lead to the evaluation of researchers and, ultimately, to the evaluation of universities. Specifically, we will analyze the part-of-speech composition ratio of data descriptions and compare the characteristics of the highly cited research data with the low-cited research data.

Keywords: Research data, Reuse, Data description, Citation

1 Introduction

Research data grows day by day, and research data management, including storage and distribution, is a major issue due to the worldwide trend toward open science. Its importance is also beginning to be recognized.

The reuse of research data is expected to improve research efficiency and promote the fusion of different fields of research. In this context, data repositories and data journals are becoming popular worldwide. In order to promote the reuse of research data, metadata (especially data descriptions) that facilitate the understanding of research overviews are important. Although data descriptions are important for explaining a wide variety of research data, they have not yet been fully discussed.

2 Related Work

Jinfang Niu argued that the lack of data description hinders the promotion of data reuse [1]. The cause of the lack of data description is that the interests of data creators and data users dampen the willingness of data creators to provide information. Jinfang also stated that the decrease in communication makes it necessary to prioritize necessary information, but tacit information from the data creator has a lower priority, and that the way to overcome these problems is to increase the peripheral knowledge of the data users and the data creators and data

[*] University Library, Osaka University, Osaka, Japan
[†] Faculty of Commerce, Seinan Gakuin University, Fukuoka, Japan

users. Jinfang also explains that these can be overcome by increasing communication channels between data creators and data users.

Ui Ikeuchi mentions a metadata case study in research data management at the University of Edinburgh [2]. She proposes that in sharing research data, metadata should be limited to basic elements as in the case of the University of Edinburgh and that freely describable data descriptions should be enhanced.

According to Ayoung Yoon's research, researchers spend a lot of time deciding whether or not to reuse data, and as a result, they often give up on data reuse.[3] As a result, researchers are likely to lose access to more data by wasting their time. In order to promote data citation, it is necessary to reduce the time required for data search and the time required for grasping the data summary. She insists that not only many of the metadata items are filled in, but also the richness of data descriptions that are important for understanding the data overview effectively reduces the time for researchers to match research data.

Although these studies mention the importance of data description, a specific analysis of the data description itself is considered necessary. This study analyzed actual data descriptions. A more detailed analysis was conducted to clarify the characteristics of data descriptions of research data with a high number of citations from the viewpoint of part-of-speech composition.

3 Data Description Analysis Method

In this study, we focus on the data descriptions that researchers first see when deciding whether or not to reuse research data, and analyze the relation between part-of-speech and the number of citations. Specifically, data descriptions are extracted from data repositories and data journals, and morphological analysis of parts of speech is performed using KH coder (morphological analysis engine “Stanford POS tagger”) [4],[5].

Data descriptions are contained in data repositories and data journals. First, Edinburgh DataShare was selected from the data repository. The reason for the selection was that it is one of the world's first data repositories to be established. Although it is difficult to investigate the number of citations of data in data repositories, data repositories have expanded earlier than data journals and were selected to confirm the current status of data descriptions. For data journals, the following four journals were selected from the Scopus bibliographic database. We selected journals in terms of the number of registrations, two from data journals not restricted to any field and two from data journals specialized in a particular field. “Scientific Data” and “Data in Brief” were selected as data journals not restricted to any field, and “Earth System Science Data” and “Chemical Data Collections” were selected as field-specific data journals. [6],[7],[8],[9],[10].

Although data descriptions are optional for research data registered in the Edinburgh DataShare, data journals have data descriptions as metadata, just like general academic paper submissions, and include text explaining how the research data was collected and how it was reproduced. The data description briefly describes the content of the academic paper or another article and is an important sentence that determines whether the information is necessary for researchers and whether they will read the content in-depth. For example, adjectives and adverbs may increase ambiguity if used too often, but they may also mitigate limitations and catch the attention of a wider audience. Nouns, on the other hand, are also important words, and there are cases where daring to use technical terms may convey things in detail.

4 Result of Analysis

A. Data repository Edinburgh DataShare

First, we analyzed the research data registered in Edinburgh DataShare, a data repository operated by the University of Edinburgh. The Edinburgh DataShare is operated by the IS (Information Service) service, which was established as a result of the Data Information Specialists Committee-UK (DISC-UK) DataShare Project, a data-sharing project conducted by the Universities of Edinburgh, Oxford, and Southampton from 2007 to 2009. Data registration started in 2008, and it is one of the pioneering data-sharing repositories in the world. It is currently managed by the Research Data Support team within Information Services at the University of Edinburgh. The number of research data registrations in the last five years is shown in Table 1.

There were 298 datasets registered during 2021, and we investigated their data descriptions. The datasets were categorized by each academic field, and the data were restricted to those fields in which more than 5 data sets were registered. We first investigated the number of words in the data descriptions, because we found that the number of words varied greatly depending on the data description. Figure 1 plots the number of words in each field of study in descending order of the number of words described in the data. The number of words describing the data varied considerably even within the same discipline. Some of the data descriptions were abstracts of the articles in which the data were published, while others were detailed descriptions of the background of the data.

Table 1: Number of Registration for Each Year

Registration year	2017	2018	2019	2020	2021
Number of Registrations	555	205	265	219	298

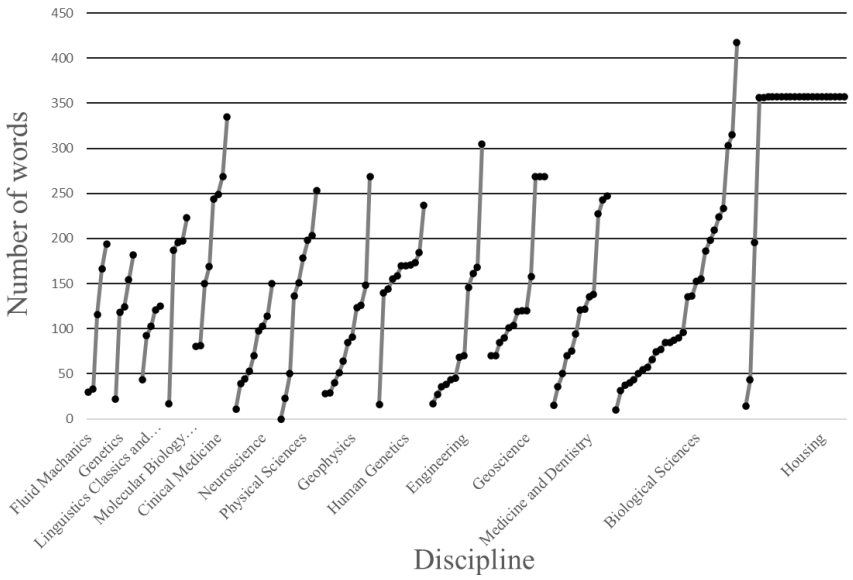

Figure 1: The Number of Words in Data Description

It was also found that the data descriptions were the same despite the different data. One reason for this may be that data repositories, unlike data journals, are also places where evidence data published in scholarly articles are made public, and multiple data in articles are registered. However, if the goal is not merely to publish evidence data but to reuse those data, it is important to enhance the data descriptions.

B. Data journal

Data journals began to be published in earnest around 2014 and take a form similar to previous journals that go through a peer-review process. In this section, we analyze the journals "Scientific Data", "Earth System Science Data", "Data in Brief" and "Chemical Data Collections". First, an analysis of the part-of-speech structure of data descriptions registered in "Scientific Data", "Earth System Science Data", "Data in Brief" and "Chemical Data Collections" is shown in Figure 2.

While there were no significant differences among nouns, verbs, and adjectives, a different trend was observed for proper nouns, with the percentages decreasing for "Chemical Data Collections" "Earth System Science Data" "Data in Brief" and "Scientific Data" in that order.

Table 2 summarized the number of citations and the average number of citations per registered data for "Scientific Data", "Earth System Science Data", "Data in Brief" and "Chemical Data Collections".

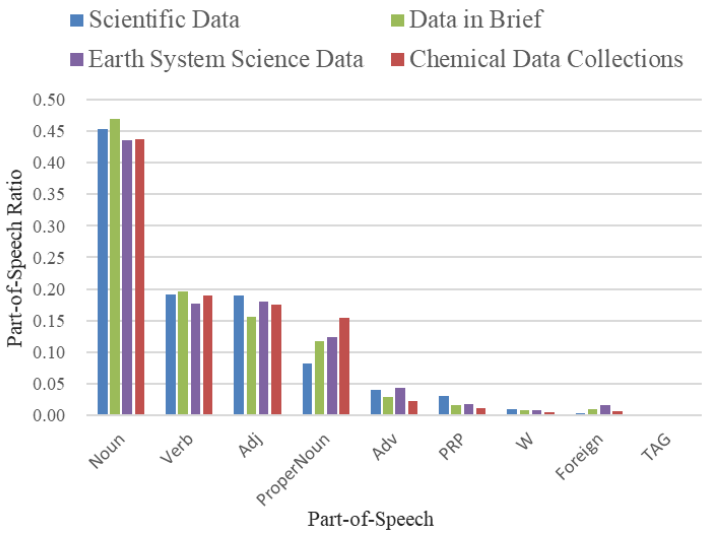

Figure 2 : Part of Speech Composition in Data Description. (4 data journal)

Table 2: Number of Registration for Each Year

Name of Journal	Number of citations	Total of paper	Number of citations per paper
Scientific Data	48,590	1,946	24.97
Earth System Science Data	25,265	949	26.62
Data in Brief	24,763	7,555	3.28
Chemical Data Collections	3,071	749	4.10

When focusing on the number of citations per registered data, it can be classified into "Scientific Data" and "Earth System Science Data" which have a large number of citations per registered data, and "Data in Brief" and "Chemical Data Collections" which have a small number of citations per registered data. The part-of-speech structure of the registered data descriptions was analyzed in detail as shown in Figure 3 and 4.

The analysis shows that "Data in Brief" and "Chemical Data Collections" do not show significant changes in the slope of the decrease of verbs and adjectives and the slope of the decrease of adjectives and proper nouns concerning the part-of-speech composition ratio. However, "Scientific Data" and "Earth System Science Data" show significantly different slopes of decrease in verbs and adjectives and slopes of decrease in adjectives and proper nouns. From this, it can be inferred that the number of citations per registration data tends to be larger when the slope of the decrease in adjectives and proper nouns is larger than the slope of the decrease in verbs and adjectives. Since these trends are for the journal as a whole, more detailed analysis of trends per article is needed in the future.

In any case, this result shows a different trend from the results analyzed in the data repository Edinburgh DataShare shown in Figure 5. This is thought to be because the data descriptions in Edinburgh DataShare are directly posted as abstracts of academic papers describing the data, and thus the ratio of proper nouns is high. Since the academic papers are submitted to journals that specialize in the field, the abstracts are written on the assumption that the readers are experts in the field. Therefore, it is assumed that many proper nouns are used. However, from the viewpoint of open science, it may be important to avoid the use of specialized proper nouns and to describe data in a plain manner that is easy to understand the outline to promote the utilization of data in more fields. From these comparative results, it can be assumed that the ratio of proper nouns is influenced by differences in the assumed readership.

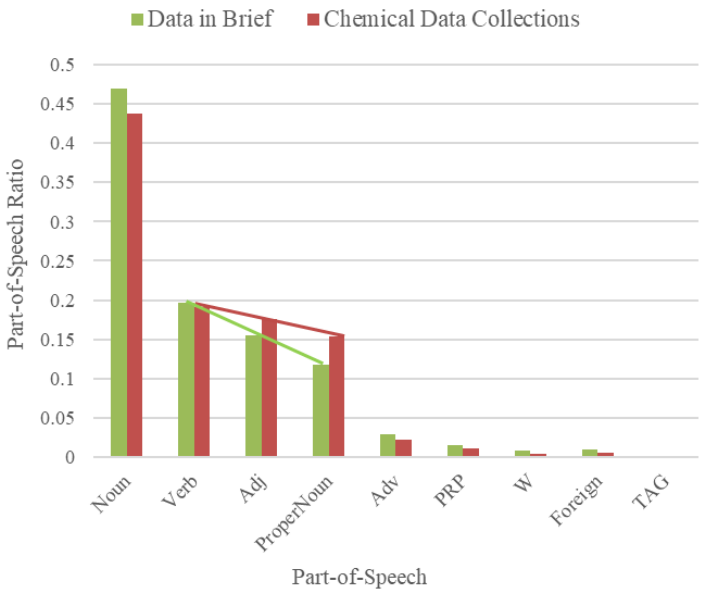

Figure 3 : Part of Speech Composition of Data Description (a)

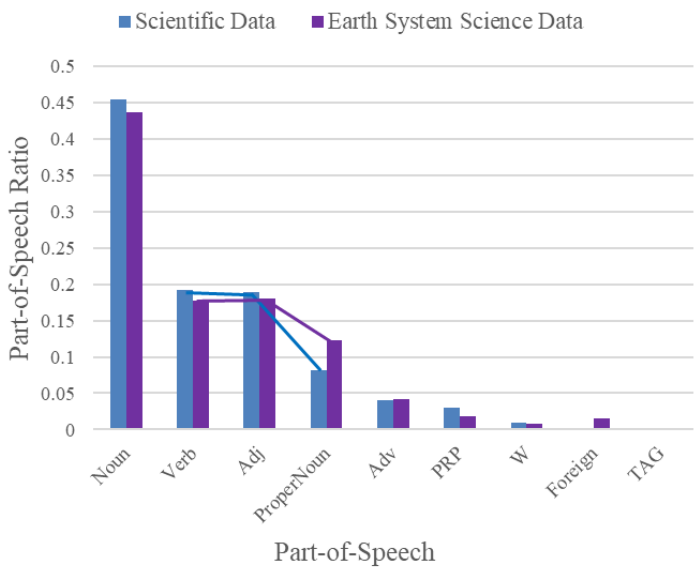

Figure 4 : Part of Speech Composition of Data Description (b)

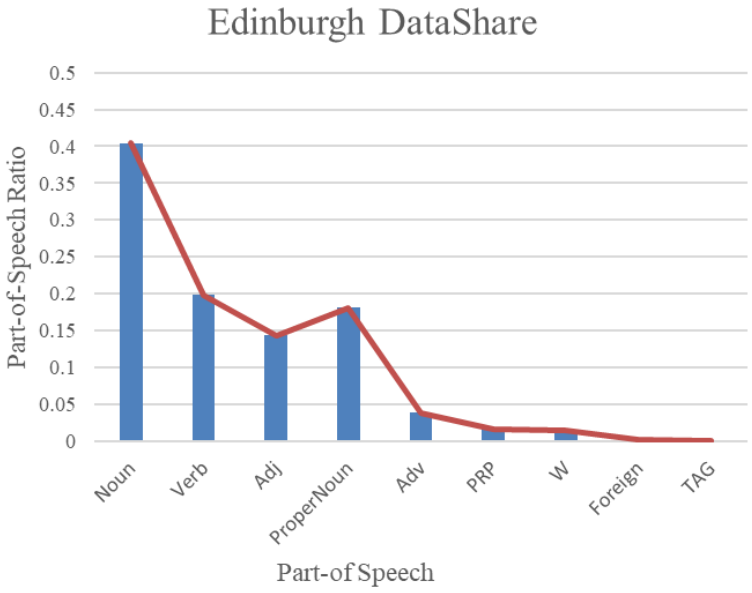

Figure 5 : Part of Speech Composition of Data Description (data repository)

To indirectly examine this assumption, we again focused on “Scientific Data”, a data journal that does not specialize in a field, and analyzed whether there is a difference in the composition ratio of proper nouns between data descriptions that briefly describe the content of the research data and the text that provides a deeper, more specialized description of the research data.

Figure 6 shows the number of citations for each year of research data articles published by “Scientific Data”. Among them, we focused on 12 research data papers (bold line) whose number of citations increased by more than 100 in the first two years after publication. 12 research data papers are shown in Table 3 with their titles, publication years, and the number of citations.

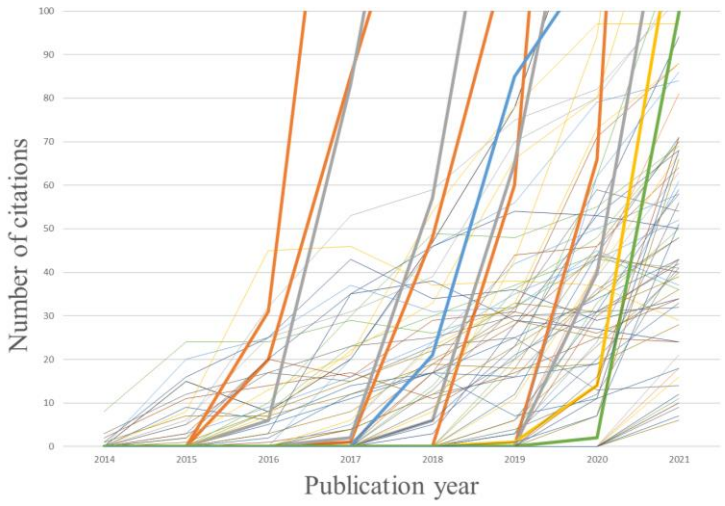

Figure 6 : Number of Citations for Each Year of Data Articles in Scientific Data

Table 3: Articles Increased by 100 or More in The Two Years Since Publication

Title	Publication year	Number of citations
A cross-country database of COVID-19 testing	2020	127
Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset	2020	566
The first high-resolution meteorological forcing dataset for land process studies over China	2020	238
The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data	2020	187
Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions	2018	525
China CO 2 emission accounts 1997-2015	2018	405
Present and future köppen-geiger climate classification maps at 1-km resolution	2018	1095
Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features	2017	846
Climatologies at high resolution for the earth's land surface areas	2017	991
MIMIC-III, a freely accessible critical care database	2016	2145
Comment: The FAIR Guiding Principles for scientific data management and stewardship	2016	3759
The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes	2015	1673

Morphological analysis was performed on the data descriptions and the text of 12 research data articles to check the part-of-speech composition ratio. As shown in Figure 7, the proportion of proper nouns in the text increased compared to that in the data descriptions. This is because “Scientific Data” is not a discipline-specific journal, and its intended audience is considered to be broad. Therefore, it is assumed that researchers unconsciously reduce the ratio of proper nouns in the data description to make the research data content easily understood. On the other hand, for readers who want to know more details, it is assumed that the researcher uses many research-specific nouns in the text to provide in-depth explanations.

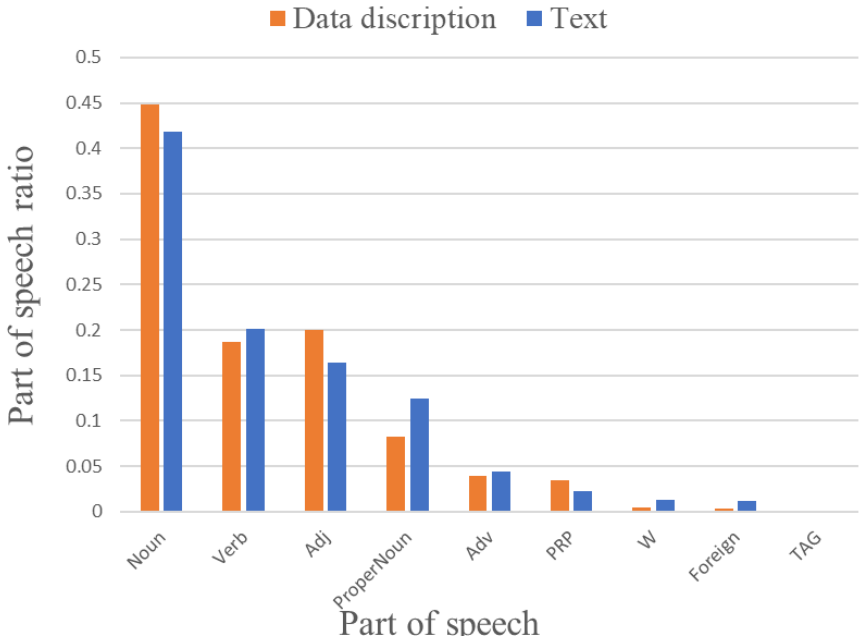

Figure 7 : The Composition Ratio of Test and Data Description(Scientific Data)

5 Conclusion and Further Work

This study focused on the data descriptions of research data articles published in data repositories and data journals to identify the characteristics of data descriptions of research data articles with the highest number of citations. Specifically, the data repository Edinburgh DataShare and the data journals “Scientific Data”, “Earth System Science Data”, “Data in Brief” and “Chemical Data Collections” were targeted. The data were analyzed by morphological analysis of data descriptions. Morphological analysis of data descriptions was performed to examine part-of-speech structure. The analysis revealed that for “Scientific Data” and “Earth System Science Data” which have a large number of citations per paper, the proportion of proper nouns was lower with a large slope compared to verbs and adjectives, which remained flat. On the other hand, for "Data in Brief" and “Chemical Data Collections” which have fewer citations per paper, the slopes of adjectives and proper nouns do not change significantly compared to the slopes of verbs and adjectives.

Focusing again on “Scientific Data” we analyzed whether there was a difference in the com-

position ratio of proper nouns between the data description and the text in the paper. As a result, it was found that the ratio of proper nouns in the data description was smaller than that in the main text in the highly cited research data articles. This suggests that the data descriptions of research data with a high number of citations may be conscious of general explanations by reducing the use of proper nouns, while the main text may be conscious of detailed explanations through the use of proper nouns.

While the analysis in this study has identified overall trends, a characteristic analysis of individual papers has not yet been conducted. In the future, a similar analysis should be conducted on individual papers to analyze their characteristic trends.

References

[1] Jinfang Niu. "Overcoming inadequate documentation. Proceedings of the American Society for Information Science and Technology", vol.46, issue1, pp.1-14, 2010.

[2] Ui Ikeuchi. "Forefront of research data management by university libraries-The case of the University of Edinburgh to strengthen research capabilities- in Japanese", vol.52, No.4, pp.227-236, 2014.]

[3] Ayoung Yoon. "Red flags in data: Learning from failed data reuse experiences," Proceedings of the Association for Information Science and Technology, vol.53, issue1, pp.1-6, 2016.

[4] Koichi Higuchi. "A Two-Step Approach to Quantitative Content Analysis: KH Coder Tutorial Using Anne of Green Gables (Part II)" Ritsumeikan Social Science Review, 53(1): pp.137-147, 2017.

[5] Stanford POS tagger, https://nlp.stanford.edu/software/tagger.shtml, (accessed on 11 April 2022).

[6] Data Share, https://datashare.ed.ac.uk/, (accessed on 11 April 2022).

[7] Scientific Data, https://www.nature.com/sdata/, (accessed on 11 April 2022).

[8] Earth System Science Data, https://www.earth-system-science-data.net, (accessed on 11 April 2022).

[9] Data in Brief, https://journals.elsevier.com/data-in-brief, (accessed on 11 April 2022).

[10] Chemical Data Collections, https://www.sciencedirect.com/journal/chemical-data-collections, (accessed on 11 April 2022).