Beyond funding: What can acknowledgements reveal about credit distribution in science?1

Adèle Paul-Hus[*], Adrián A. Díaz-Faes[**], Nadine Desrochers[*], Rodrigo Costas[***], Maxime Sainte-Marie[*], Benoît Macaluso [****] and Vincent Larivière[****]

[*] adele.paul-hus@umontreal.ca, nadine.desrochers@umontreal.ca, msaintemarie@gmail.com;
École de bibliothéconomie et des sciences de l'information, Université de Montréal
PO Box 6128, Downtown Station, Montreal, Quebec, H3C 3J7 (Canada)

[**] adrian.arias@cchs.csic.es
IFS, Spanish National Research Council (CSIC), Albasanz 26-28, 28037 Madrid (Spain)

[***] rcostas@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. box 905, 2300 AX Leiden (the Netherlands)

[****] macaluso.benoit@uqam.ca; vincent.lariviere@umontreal.ca
École de bibliothéconomie et des sciences de l'information, Université de Montréal
PO Box 6128, Downtown Station, Montreal, Quebec, H3C 3J7 (Canada) and Université du Québec à Montréal, Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Observatoire des Sciences et des Technologies (OST), CP 8888, Succ. Centre-Ville, H3C 3P8 Montréal, Qc. (Canada)

ABSTRACT
Funding acknowledgements found in scientific publications have been used for decades to study the role of funding in science production. However, beyond funding information, acknowledgements convey the indebtedness of authors to individuals, institutions and organizations that contributed, in some way, to the research that lead to publication. The objective of this paper is to explore the different types of contributions acknowledged in WoS funding acknowledgement (FA) texts. The Correspondence Analysis performed in this study reveals that FAs offer a unique window on research and collaborative practices, credit distribution, and how these vary across disciplines. FAs thus contribute to make the traditionally "invisible contributions" visible to the scientific community. Results presented in this study go further in demonstrating that acknowledgements are not confined to credit attribution, as they include disclosures of conflict of interest.

INTRODUCTION
Funding acknowledgements found in scientific publications have been used for decades to study the role of funding in science production (e.g., Crawford and Biderman, 1970; Harten and Hooten, 1992). However, beyond funding information, acknowledgements convey the indebtedness of authors to individuals, institutions and organizations that contributed, in some way, to the research that lead to publication. Acknowledgements reveal the hidden infrastructure that supports scientific research, showing how colleagues, tools, materials, and grants are mobilized in the context of scientific endeavour (Cronin, 2005).

[1] This work was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Fonds de recherche du Québec – Société et culture (FRQSC), as well as by the South African DST-NRF Centre of Excellence in Scientometrics and STI Policy (SciSTIP).

Acknowledgements can therefore be perceived as credit for contributions and could be used to better understand collaboration and the division of labor in the scientific field.

The objective of this study is to explore the different types of contributions acknowledged in WoS funding acknowledgement (FA) texts. More specifically, this study aims to answer the following research questions:

- What types of contributions are acknowledged in FA texts?
- How do these types of contributions vary by discipline?

DATA AND METHODS

Data were retrieved from Web of Science (WoS) Science Citation Index Expanded (SCI-E), which includes FA (i.e., funding information but also the full acknowledgement text provided by authors for all other types of contributions). The FA corpus used in this study was generated by collecting all 2014 articles and reviews from Biology, Biomedical Research, Chemistry, Clinical Medicine, Earth and Space, Engineering and Technology, Mathematics, and Physics totaling 880,809 FA texts for as many papers. Discipline assignation was done using the National Science Foundation field and subfield classification of journals.

A term extraction procedure[i] was performed on FA texts to extract nouns and noun phrases. Since acknowledgements are collected and indexed by WoS only if they include funding source information (Paul-Hus, Desrochers and Costas, under review), the presence of funding information was considered a common denominator. Funding-related terms were thus removed from the acknowledgement text in order to focus the analysis on other types of acknowledgements. Additionally, proper nouns were removed from the corpus using python packages *nltk* (Bird, Klein and Loper, 2009) and *pattern* (De Smedt and Daelemans, 2012).

A frequency score was generated for each extracted lexical items (nouns and noun phrases), providing the number of times the item appears in the corpus. A threshold of a minimum of five occurrences was applied for a remaining corpus of 5,770 distinct lexical items. Finally, a Correspondence Analysis (CA) was applied on the corpus following the procedure described in Díaz-Faes and Bordons (2014) and using a MATLAB package (Vicente-Villardón, 2014).

RESULTS

Table 1 shows that FAs are not evenly distributed across disciplines, both in absolute terms and in relation to the proportion of FA-bearing papers per discipline. The FA ratio varies between 54% (Clinical Medicine) and 82% (Biomedical Research) for an overall ratio of 69%.

Table 1. Presence of FA by discipline for 2014 papers

Discipline	All papers	FAs	% of paper with FA
Clinical Medicine	373,185	202,464	54%
Engineering and Technology	222,263	149,157	67%
Biomedical Research	180,245	148,177	82%
Chemistry	144,020	114,747	80%
Physics	124,619	94,344	76%
Biology	98,893	72,171	73%
Earth and Space	85,305	66,055	77%
Mathematics	50,377	33,694	67%
Total	1,278,907	880,809	69%

Using CA, lexical items were grouped in five clusters using k-means clustering and cosine similarity. A threshold of 700 occurrences was set for visualisation purposes. Retaining five axes, over 85% of variance is explained and all disciplines can be interpreted on plane 1-3 (Figure 1) and plane 2-4 (Figure 2)[ii].

Cluster 1, where authors mainly show their gratitude for the technical help and assistance received, is mostly found in Biomedical Research. Cluster 2 is formed from lexical items related to technical assistance, access to facilities and resources, and discussions associated to research work and projects. This pattern is found in Physics, Chemistry, and Engineering and Technology.

Cluster 3 reflects suggestions and comments but in contrast to Cluster 2, these contributions seem to be more related to manuscript improvement rather than to the research process itself. Furthermore, field work, a specific form of data collection associated with Biology as well as Earth and Space and involving uncontrolled environments, also characterize that cluster.

Cluster 4 gathers lexical items linked to manuscript and editorial assistance as well as data analysis support. Authorship and potential conflicts of interest, which are important concerns in clinical studies given the consequences of fraud and unethical behaviour in that field, also emerge as strong factors in this cluster, led by papers in Clinical Medicine. Cluster 5 appears as the most peripheral one. Mostly pertaining to Mathematics, this cluster is characterized by lexical items referring mainly to authorship and intellectual debts associated to manuscript preparation.

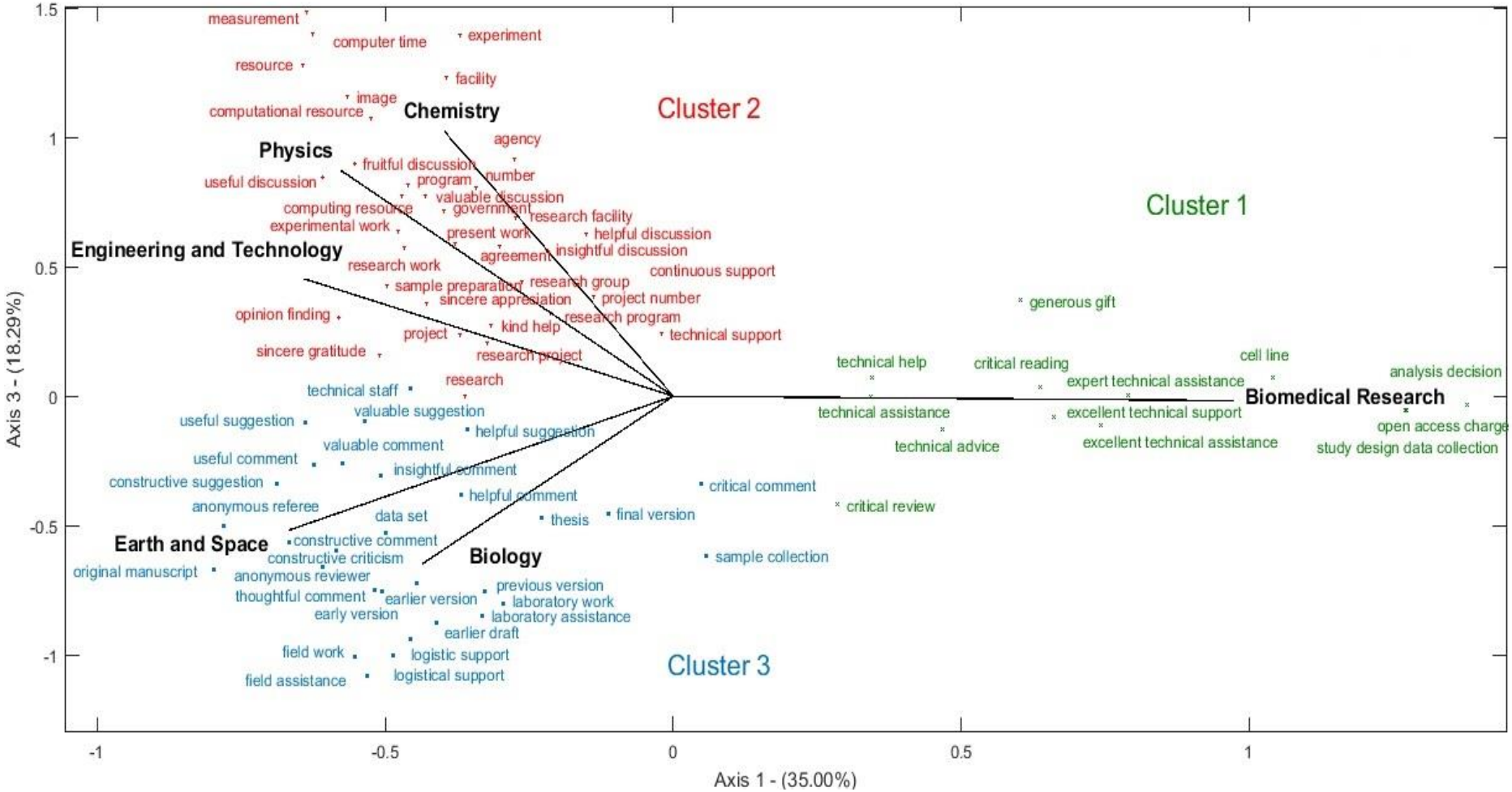

Figure 1. Bidimensional Correspondence Analysis for acknowledgements patterns by discipline (plane 1-3)

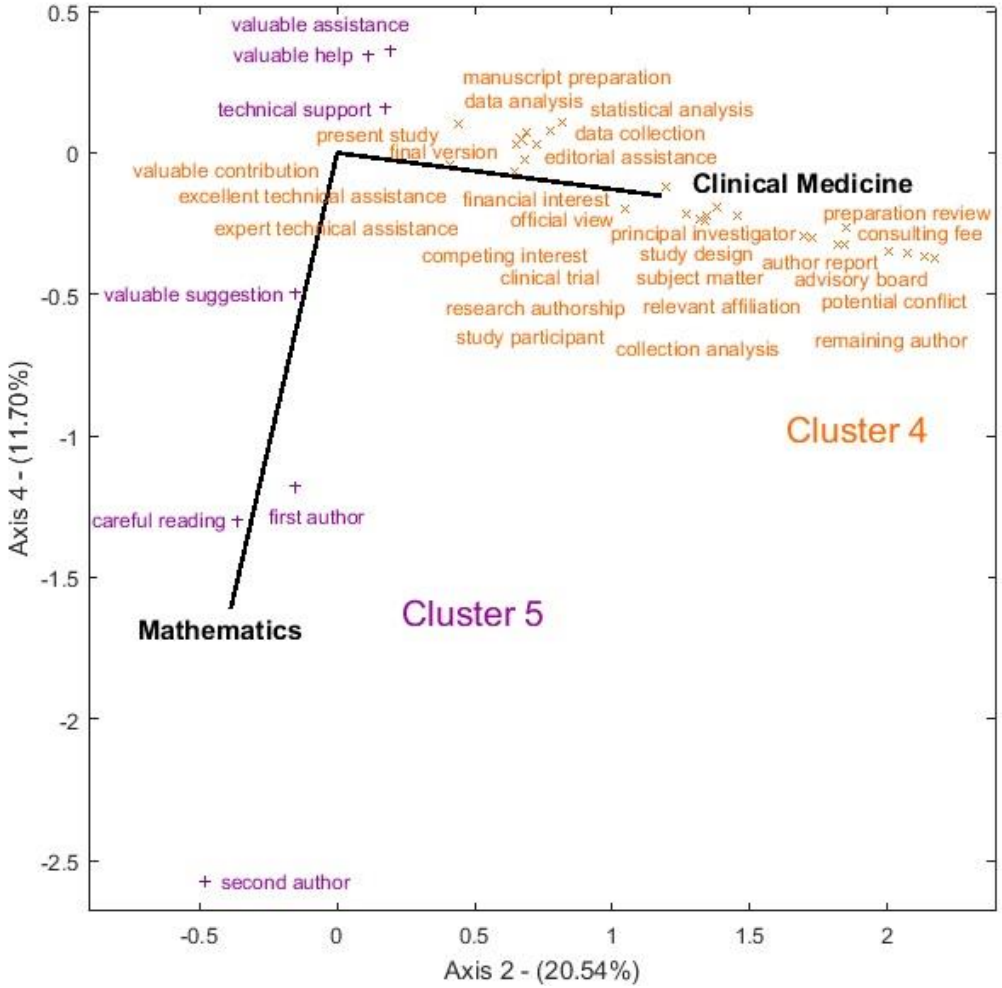

Figure 2. Bidimensional Correspondence Analysis for acknowledgements patterns by discipline (plane 2-4)

Table 2 presents the most frequent lexical items in the corpus. Percentages indicate the relative frequency of a given item for each discipline. Lexical items related to critical reading are among the most frequent, and two broad categories can be distinguished: acknowledgements to reviewers (e.g., "anonymous reviewer") and those made to colleagues (e.g., "helpful discussion"), the latter being most frequently found in Physics. The remaining lexical items related to critical reading can be associated to either reviewers or colleagues (e.g., "helpful comment").

As revealed by the CA (Figure 2), lexical items associated with conflicts of interest are mostly found in Clinical Medicine, in which disclosure of such potential conflicts is made mandatory by most journals' guidelines (e.g., ICMJE, 2015). That being said, caution is required when analyzing terms out of context. For example, two frequent lexical items, "study design data collection" and "analysis decision", do not constitute contribution acknowledgements *per se*, since most of their occurrences come from a conflict of interest disclosure statement found in journal guidelines, such as *PLOS* journals and *PeerJ*. The extracted items must thus be interpreted in the context of the original statement, such as, "The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript."

Table 2. Most frequent FA lexical items by discipline

Acknowledgement terms	Biology	Biomedical Research	Chemistry	Clinical Medicine	Earth and Space	Engineering & Technology	Mathematics	Physics	Total
anonymous reviewer	12.8%	1.7%	0.3%	0.4%	20.9%	2.5%	2.5%	0.8%	3.7%
study design data collection	0.4%	19.2%	0.0%	1.2%	0.0%	0.0%	1.6%	0.0%	3.6%
analysis decision	0.4%	19.0%	0.0%	1.1%	0.0%	0.0%	1.6%	0.0%	3.6%
helpful discussion	1.1%	2.6%	2.7%	0.8%	2.8%	1.2%	1.5%	4.1%	2.0%
technical assistance	4.0%	3.4%	1.2%	2.2%	1.4%	0.7%	0.0%	0.7%	1.9%
helpful comment	4.2%	1.2%	0.3%	0.6%	5.2%	0.8%	3.9%	1.0%	1.5%
valuable comment	2.9%	0.6%	0.3%	0.2%	3.4%	1.2%	3.9%	0.8%	1.1%
technical support	1.8%	1.4%	1.0%	0.9%	1.6%	1.0%	0.0%	1.0%	1.1%
anonymous referee	2.2%	0.3%	0.1%	0.0%	5.4%	0.8%	6.4%	0.8%	1.1%
first author	2.0%	0.3%	0.1%	0.4%	1.8%	1.2%	9.4%	0.4%	1.1%
useful discussion	0.3%	0.5%	0.8%	0.1%	2.3%	0.7%	1.0%	4.6%	1.1%
constructive comment	2.3%	0.4%	0.1%	0.1%	6.5%	0.8%	1.7%	0.3%	1.0%
fruitful discussion	0.3%	0.4%	1.0%	0.1%	1.3%	0.7%	0.6%	2.8%	0.8%
research project	0.8%	0.5%	0.4%	0.4%	1.4%	1.0%	0.8%	0.6%	0.7%
critical reading	1.0%	1.9%	0.3%	0.6%	0.2%	0.1%	0.0%	0.4%	0.6%
data collection	1.4%	0.7%	0.3%	1.2%	0.7%	0.1%	0.1%	0.1%	0.6%
project	0.6%	0.5%	0.8%	0.1%	1.6%	0.5%	1.0%	0.4%	0.6%
useful comment	1.2%	0.3%	0.1%	0.1%	2.6%	0.2%	1.6%	0.9%	0.6%
official view	0.1%	0.8%	0.2%	1.4%	0.3%	0.1%	0.2%	0.1%	0.6%
earlier version	2.6%	0.3%	0.0%	0.2%	1.9%	0.2%	0.8%	0.1%	0.5%
valuable discussion	0.4%	0.4%	0.6%	0.1%	0.7%	0.5%	0.4%	1.5%	0.5%
second author	0.3%	0.0%	0.0%	0.0%	0.3%	0.5%	8.6%	0.2%	0.5%
research work	0.4%	0.3%	0.6%	0.2%	0.6%	1.1%	0.4%	0.4%	0.5%
analysis	0.4%	0.4%	1.2%	0.1%	1.2%	0.5%	0.0%	0.1%	0.5%
measurement	0.0%	0.1%	1.8%	0.0%	0.1%	0.8%	0.0%	0.6%	0.5%

Color based on cell value: from darkest blue (lowest value) to darkest red (highest value)

DISCUSSION AND CONCLUSION

The CA performed in this study reveals that FAs offer a unique window on research and collaborative practices, credit distribution, and how these vary across disciplines. FAs thus contribute to make the traditionally "invisible contributions" visible to the scientific community, but the results here presented go further in demonstrating that acknowledgements are not confined to credit attribution, as they include disclosures of conflict of interest—or of their absence. These disclosures reveal that acknowledgements can also be self-declarations of ethical behaviour and tools to release third parties for any responsibility on the published results. Acknowledgements' role might then be akin to that of contributorship statements and warrant another look in terms of their place in the evaluation of science and scientists.

Furthers steps in this study will include the addition of more disciplines (Social Sciences) and the use of linguistic processing techniques adapted to the idiosyncrasies of the FA corpus. This will allow for a better understanding of how acknowledgements can support the analysis of scientific practices beyond the core concern with funding.

REFERENCES

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python: Analyzing text with the Natural Language Toolkit*. Beijing: O'Reilly Media.

Crawford, E. T., & Biderman, A. D. (1970). Paper money: Trends of research sponsorship in American sociology journals. *Social Science Information*, *9*(1), 50–77.

Cronin, B. (2005). *The hand of science: Academic writing and its rewards*. Lanham, Maryland: Scarecrow Press.

De Smedt, T., & Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research*, *13*(1), 2063-2067.

Díaz-Faes, A. A., & Bordons, M. (2014). Acknowledgments in scientific publications: Presence in Spanish science and text patterns across disciplines. *Journal of the Association for Information Science and Technology*, *65*(9), 1834-1849.

Harter, S. P., & Hooten, P. A. (1992). Information science and scientists: JASIS, 1972–1990. *Journal of the American Society for Information Science*, *43*(9), 583–593.

ICMJE (2015). Uniform requirements for manuscripts submitted to biomedical journals: Ethical considerations in the conduct and reporting of research: Authorship and contributorship. http://www.icmje.org/icmje-recommendations.pdf

Paul-Hus, A., Desrochers, N., and Costas, R. (under review). Characterization and description of funding acknowledgement data in Web of Science: considerations for their use. *Scientometrics*.

Vicente-Villardón, J.L. (2014). *MultBiplot: A package for Multivariate Analysis using Biplots*. Departamento de Estadística, Universidad de Salamanca [Software]. http://biplot.usal.es/ClassicalBiplot/index.html

[i] The term extraction transformation carried out by the SQL Server 2016 software : https://msdn.microsoft.com/en-us/library/ms141809

[ii] Lexical items and disciplines with variance explained below 40% are not displayed.