Scientometrics (2018) 117:637–640
https://doi.org/10.1007/s11192-018-2855-y

Field classification of publications in Dimensions: a first case study testing its reliability and validity

Lutz Bornmann[1]

Received: 6 June 2018 / Published online: 13 July 2018

Abstract
Dimensions is a research data infrastructure and tool, including grants, publications, citations, clinical trials, and patents in one place. An interesting feature of Dimensions is its field classification scheme, which is not based on journal classification systems, as in the Web of Science or Scopus, but on machine learning. Each publication is assigned to at least one field. Using the set of my own publications, I investigated whether they were reliably and validly assigned to fields. The results put in question the reliability and validity of the scheme. Large scale studies seem necessary to investigate the scheme in more detail.

Keywords Bibliometrics · Dimensions · Lutz Bornmann

The Digital Science software product Dimensions (see https://www.dimensions.ai) is a research data infrastructure and tool, including grants, publications, citations, clinical trials, and patents in one place (Bode et al. 2018). The platform focusses on publication data; thus, it is similar to Web of Science (WoS, Clarivate Analytics) and Scopus (Elsevier). However, it also includes further data, such as grants and clinical trials. Dimensions is a very new tool, which was launched at the beginning of 2018. An interesting new feature of Dimensions is its field classification scheme, which is not based on journal classifications, as in WoS or Scopus, but on machine learning. It operates on the level of single publications.

Bode et al. (2018) describe the field-classification scheme as follows: "Technology has developed further. The fields of natural language processing, machine learning and artificial intelligence have all made huge advances in recent years. Dimensions has been able to leverage these technologies to solve a very practical problem requiring a different approach: If you want to consistently categorize grants, patents and clinical trials, a journal proxy is no longer available. The path we have chosen for Dimensions is to use existing classification systems and an AI/machine learning based approach to automatically assign

✉ Lutz Bornmann
bornmann@gv.mpg.de

[1] Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

Table 1 Fields of research for publications by Lutz Bornmann (some example titles of papers are presented which have been assigned to the fields)

	Frequency	Percent of responses	Percent of cases
Applied Economics (e.g. "an empirical look at the nature index")	26	13.07	9.92
Public Health and Health Services (e.g. "sampling issues in bibliometric analysis")	23	11.56	8.78
Information Systems (e.g. "how to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations")	16	8.04	6.11
Psychology (e.g. "what do we know about the h index?")	16	8.04	6.11
Statistics (e.g. "which cities produce more excellent papers than can be expected? A new mapping approach, using Google Maps, based on statistical significance testing")	15	7.54	5.73
Sociology (e.g. "introducing CitedReferencesExplorer (CREXplorer): A program for reference publication year spectroscopy with cited references standardization")	14	7.04	5.34
Historical Studies (e.g. "recent developments in China—U.S. cooperation in science")	12	6.03	4.58
Econometrics	8	4.02	3.05
Literary Studies	8	4.02	3.05
Policy and Administration	8	4.02	3.05
Artificial Intelligence and Image Processing	6	3.02	2.29
Physical Chemistry (incl. structural)	5	2.51	1.91
Clinical Sciences	4	2.01	1.53
Computer Software	3	1.51	1.15
Other Studies in Human Society	3	1.51	1.15
Political Science	3	1.51	1.15
Specialist Studies in Education	3	1.51	1.15
Biochemistry and Cell Biology	2	1.01	0.76
Ecology	2	1.01	0.76
Genetics	2	1.01	0.76
History and Philosophy of Specific Fields	2	1.01	0.76
Neurosciences	2	1.01	0.76
Other Biological Sciences	2	1.01	0.76
Pure Mathematics	2	1.01	0.76
Theoretical and Computational Chemistry	2	1.01	0.76
Atomic, Molecular, Nuclear, Particle and Plasma Physics	1	0.50	0.38
Biomedical Engineering	1	0.50	0.38
Computation Theory and Mathematics	1	0.50	0.38
Ecological Applications	1	0.50	0.38
Environmental Science and Management	1	0.50	0.38
Geomatic Engineering	1	0.50	0.38
Linguistics	1	0.50	0.38
Materials Engineering	1	0.50	0.38
Other Physical Sciences	1	0.50	0.38
Philosophy	1	0.50	0.38
Total	199	100.00	75.95

a consistent set of categories to all documents—regardless of the source" (Bode et al. 2018, p. 4).

As a development partner for Dimensions, I have unrestricted access to the tool. In a first explorative case study, I investigated its field classification scheme. The scheme is especially interesting for bibliometrics, because it is not based on journals (but on single papers). Journal sets have the disadvantage that papers published in multidisciplinary journals cannot be assigned to corresponding fields and small fields are scarcely reflected in the scheme (Bornmann et al. 2008). The case study is based on my own publications. Using the author identifier in Dimensions (which is still in the beta phase), I identified and downloaded my set. With 262 publications in the set, Dimensions seems to have identified my publications more or less validly. My validated list on ResearcherID (see www.researcherid.com) contains 285 papers (ResearcherID number: A-3926-2008). Although I have published many reviews, notes, and letters besides articles, my publications have been assigned only to articles (n = 260) and chapters (n = 2) in Dimensions.

All my publications appeared in the areas of research evaluation, including scientometrics, informetrics, bibliometrics, and altmetrics, as well as grants and journal peer review. Thus, my research is in the area of sociology of science or science of science (see the overview of this area by Fortunato et al. 2018). Some publications are also in the history of science. Using my own publication set, I can immediately see whether the classification of my papers is correct or not. Table 1 shows the fields, which Dimensions assigned to 199 of my 262 publications (the rest are without field assignments). Since some publications have two field assignments, the table presents two percentages: percent of responses and percent of cases. The assignments in the table show that they frequently do not agree with the fields in which I am mostly active: scientometrics, informetrics, bibliometrics, and altmetrics. For fields with more than ten publications in the table, I included the title from a single paper as an example.

The results show that most of the papers seem misclassified. "Applied Economics" and "Public Health and Health Services" are the most frequent fields in the table, which do not agree with my areas of research. Some papers have been assigned to fields, which reflect only a peripheral aspect of the paper. For example, the assigned field for the paper with the title "which cities produce more excellent papers than can be expected? A new mapping approach, using Google Maps, based on statistical significance testing" is not wrong, but does not reflect the main topic of the paper: spatial bibliometrics.

The first studies on publication data from Dimensions were published by Thelwall (2018) and Orduna-Malea and Delgado Lopez-Cozar (2018). Thelwall (2018) concludes that "Dimensions seems to be a plausible alternative to Scopus and the Web of Science for general citation analyses and for citation data in support of some types of research evaluations" (p. 430). Orduna-Malea and Delgado Lopez-Cozar (2018) conclude similarly that "Dimensions is an alternative for carrying out citation studies, being able to rival Scopus (greater coverage and free of charge) and with Google Scholar (greater functionalities for the treatment and data export)". However, "anecdotal evidences" make the authors also "suspect on the reliability and general validity of the subject classification" used by Dimensions.

The results, which I received based on my own publication set and by Orduna-Malea and Delgado Lopez-Cozar (2018) based on some other examples, put in question the reliability and validity of the Dimensions field classification scheme. However, this case study is based on only a very small dataset from very specific fields. Thus, it is necessary to undertake large scale investigations including papers from different fields in future studies. This is especially necessary, because the field-normalized citation scores presented in

Dimensions for each publication are calculated based on the field classification scheme (see https://dimensions.freshdesk.com/support/solutions/articles/23000013157-what-is-the-fcr-how-is-it-calculated-).

Acknowledgements The bibliometric data used in this paper is from Dimensions. The author thanks Digital Science for data access.

References

Bode, C., Herzog, C., Hook, D., & McGrath, R. (2018). *A guide to the dimensions data approach. A collaborative approach to creating a modern infrastructure for data describing research: where we are and where we want to take it*. London: Digital Science.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Use of citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8,* 93–102. https://doi.org/10.3354/esep00084.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., et al. (2018). Science of science. *Science*. https://doi.org/10.1126/science.aao0185.

Orduna-Malea, E., & Delgado Lopez-Cozar, E. (2018). Dimensions: Re-discovering the ecosystem of scientific information. Retrieved May 5, 2018, from https://arxiv.org/abs/1804.05365.

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics, 12*(2), 430–435. https://doi.org/10.1016/j.joi.2018.03.006.