Data from "Grey Literature citations in the age of Digital Repositories and Open Access"

https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:64280/tab/2

Silvia Giannini, ISTI-CNR Italy
ORCiD 0000-0001-7323-3786
Dataset Creator / Draft Text / Review & Editing
Anna Molino, ISTI-CNR Italy
Review & Editing

Abstract

The data collected is based on a sample corpus built on: a) the bibliographical references of articles in four journals over the years 2012-2014; b) the proceedings of two international conferences held in 2012 and 2014. The full text paper, presented at the International Conference Series on Grey Literature, measures grey citations in the years 2012, 2013 and 2014 and describes the features of GL documents mentioned in the areas *Computational Linguistics* and *Computer Science and Engineering*. The data from the study was collected and arranged in 2016. The original information was extracted directly from the primary sources, i.e. the bibliographical references of the articles published in the selected journals and proceedings. The dataset consists of all the analyzed bibliographical references of the chosen journals and proceedings accompanied by some informative classes processed to perform calculations and provide statistical information. It contains the number of bibliographic references and the number and percentages of usage of "grey" references and "grey" document types in the considered timespan. The data derived from this sample are suitable for different types of reuse and functional for anyone interested in citation analysis.

Keywords: Bibliographic references, Citation analysis, Models of scientific communication.

Subject Area: Information Science.

Methods

- Steps

The dataset has been produced by merging the tables used for data analysis and the tables used for data processing into a single Excel file. It is composed by 11 spreadsheets: the first six (ACM_TOIS, EURASIP, CL, LR&E, EACL, JDCL) contain the bibliographical references accompanied by five informative classes: year, issue number, kind of document – Grey (G) or Published (P), document type and standardized document type. The remaining spreadsheets contain the tables with measurements of the frequency of GL citing, the frequency of GL use and the intensity of GL use accompanied by graphs (Frequency_Intensity). The spreadsheet called "DT_Tables" reports the overall distribution of GL by document type while the spreadsheet called "Journals_DT and Proceedings_DT" reports the distribution of GL documents for each journal and proceeding. The last one includes the graphs measuring the intensity of use of the individual types of document by each resource analyzed (DT_Graphs).

- Sampling strategy

The sample data was selected from journals with an Impact Factor (IF) over the last three years, indexed by Scopus Citation Database and ISI Web of Science, and from two proceedings of international conferences held in 2012 and 2014. The chosen journals are all indexed under the ISI-JCR subject category "Computer Science" (CS), except for the *EURASIP Journal on Advanced in Signal Processing* (EURASIP), which is under the subject category "Engineering, Electrical & Electronic" (E&E). *ACM Transactions on Information Theory* (ACM TOIS) is under the sub-category "Information systems"; *Computational Linguistics* (CL) is under the sub-categories "Artificial Intelligence" and "Interdisciplinary Applications"; *Language Resources and Evaluation (LR&E)* is under the sub-category "Interdisciplinary Applications". Scopus citation database places the journals CL and LR&E in areas related also to the Humanities and Social Sciences: "Language and Linguistics" for CL; "Language and Linguistics", "Education", and "Library and Information Sciences" for LR&E. Indeed, Computational Linguistics is a discipline that draws contributions from different fields of study, such as linguistics, psychology, mathematics and statistics, in addition to computer science. For all these reasons, we considered the selected journals and conference proceedings as belonging to two different scientific communities: "Computer science" and "Engineering and Computational Linguistics".

- Quality Control

The quality of source data is guaranteed by official publishers. When the bibliographic reference is unclear or lacking useful information for its identification, special indexes, catalogs and Google search are used.

Dataset Description

File name:	Data.xlsx
Format name and version:	Excel 2013
Creation dates:	From 2016-03-30 to 2016-04-04
Language:	English
License:	CCO, Open access for registered users
Archive name:	DANS EASY Archive
Publication date:	2016-04-04

Potential Reuse

The original study measured the impact of Grey Literature (GL) on different areas of knowledge by selecting only "grey" citations and described the features of GL cited documents. A further study might suggest a different interpretation of the examined data or compare the results with those of other disciplinary fields. Another study might investigate the entire corpus of citations by identifying the most cited document types, their features, the time coverage as well as the nature of publishers: commercial publishers, associations, foundations. Moreover it may be interesting to analyze the amount of citations in open access journals. In the current digital era bibliographical citations have gained a strategic role within the mechanisms of scientific communication, especially due to the implementation of the citation indexing services. The Impact Factor is based on the count of citations and is the most well-known and used bibliometric indicator. The contents of this dataset are affected by this approach and are strongly oriented to traditional models of scientific communication. The Open Science movement is extending the transmission of knowledge to new documentary typologies and is encouraging the use of alternative metrics. For these reasons, the dataset could also be used for comparative purposes with the future citation models.

The original bibliographical references are freely accessible from the publishers' sites and a copy of the dataset is openly accessible in the DANS EASY Archive.

References

1. Giannini S., Biagioni S., Goggi S., Pardelli G. *Grey Literature citations in the age of Digital Repositories and Open Access*. In: GL17 - Seventeenth International Conference on Grey Literature : A New Wave of Textual and Non-Textual Grey literature (Amsterdam, NL, 1-2 December 2015). Proceedings, pp. 137 - 145. D. Farace and J. Frantzen (eds.). TextRelease, Amsterdam, The Netherlands, 2016.
2. DANS EASY Archive https://easy.dans.knaw.nl/ui/home.
3. Data Papers Project http://www.greynet.org/greyforumseries/datapapers.html.
4. GreyNet Data Paper Template – Version 1.0 http://www.greynet.org/greyforumseries/datapapers.html.