Literature Overview for:

A Literature Review on Methods for the Extraction of Usage Statements of Software and Data

Frank Krüger David Schindler

July 26, 2019

Table 0.1: Key aspects of all identified articles. Results hierarchical sorted by approach, artifact and year. *Artf. (Artifact)*: describes which artifact type was the target (D=dataset, DB=database, S=software, P=package, F=function). *Restr. (Restriction)*: indicates that the approach is limited to specific artifacts. *Appr. (Approach)*: classification of the used approach (T=term search, M=manual, R=rule-based, S=supervised learning). *Score*: Metric (P=precision, R=recall, F=FScore, κ=cohen's, A=percentage agreement) and evaluation Score (e=exact (match), p=partial, s=sentence score, d=document score, o=overestimated, *=no gold standard eval.). *Algo. (Algorithm)*: more specific description of the used approach (CA=content analysis, RE=regex matching, IB=iterative bootstrapping, DI=dictionary-based, DIST=(word-)distance measure, SVM=support vector machine, NB=naïve bayes classifier, TS=term search, RB=rule-based scoring, CRF=conditional random field, RF=random forest classifier, *=based on method). *Dom. (Domain)*: domain of the article corpus used (S=social sciences, L=life sciences, N=natural sciences, I=interdisciplinary, E=earth sciences, G=engineering sciences). *Sec. (Section)*: article section examined for research artifacts (F=full text, S=supplements, E=experiment description, M=methods section, A=abstract, R=references, *=selects multiple sections). *Obj. (Objective)*: purpose for which the extraction was performed or for which it is intended (P=(citation) practice analysis, J=(journal) policy enforcement, S=(artifact) sharing analysis, U=usage analysis, I=impact measurement, M=mapping of artifacts, O=other specific purposes). *Source*: how the article was identified for the review (O=unstructured search, Q=review query, C=citation in identified literature, S=survey article, R=reviewer suggestion).

ID	Artf.	Restr.	Appr.	Score	Algo.	Dom.	Sect.	Obj.	Source
Major [2011]	D	✓	T	-	TS	E	F	I	S
Piwowar et al. [2011]	D	✓	T	-	TS	L	F	U	O,Q
Coppin [2013]	D	✓	T	-	TS	I	F	I	S
Belter [2014]	D	✓	T	-	TS	E	F	I	S
Pepe et al. [2014]	D		T	-	TM+RB	N	F	S	S
Henderson and Kotz [2015]	D	✓	T	-	TS	G	F	P	S
Li et al. [2016a]	D	✓	T	-	TS	L	F	U	Q
Servilla et al. [2016]	D	✓	T	-	TS	E	F	I	Q
Kirlew [2011]	DB	✓	T	-	TS	L	A	I	S
Huang et al. [2015]	DB	✓	T	-	TS	L	F	P	S
Yu et al. [2015]	DB	✓	T	-	DI+TM	L	M,R	I	S
Russell et al. [2018]	S	✓	T	-	TS	L	A	O	S,R
Sieber and Trumbo [1995]	D	✓	M	-	CA	S	F	P	S
Ochsner et al. [2008]	D	✓	M	-	CA	L	F	J	C
Alsheikh-Ali et al. [2011]	D		M	-	CA	I	F,S	S	S
Mooney [2011]	D	✓	M	-	CA	S	F	P	S
Mooney and Newton [2012]	D		M	-	CA	I	F	P	O,Q
Zenk-Möltgen and Lepthien [2014]	D		M	-	CA	S	F	S	S
Mahrholz et al. [2015]	D	✓	M	-	CA	S	f	P	Q
Park and Wolfram [2017]	D		M	-	CA	L	F,S	U	Q
Yan and Weber [2018]	D	✓	M	κ n/a	CA	I	F	I	Q
Zhao et al. [2018]	D		M	κ=.86	CA	L	F	P	O,Q
Howison and Bullard [2016]	S		M	A=.68–.83	CA	L	F	P	O,Q,S
Li et al. [2016b]	S	✓	M	-	DI	G	F	P,U	Q
Nangia and Katz [2017]	S		M	-	CA	L	F	U	Q
Pan et al. [2018]	S	✓	M	κ=.73–1.	CA	I	F	I	Q
Allen et al. [2018]	S		M	-	CA	N	F	M	S,R
Li et al. [2017]	S,P,F	✓	M	κ=.67–.87	CA	L	F	P,I	O,Q
Prlic et al. [2010]	D	✓	R	-	RE	L	F	M	c
Haeussler et al. [2011]	D	✓	R	-	RE	L	F	O	S
Boland et al. [2012]	D		R	P=.97–1.,R=.24–.29 e*	IB	S	F	O	O,Q,S
Kafkas et al. [2013]	D	✓	R	F=.74–.96 o	RE+DI	L	F	P,M	S,C
Singhal and Srivastava [2013]	D		R	F=.75-.85 d	DIST	G	E	M	O
Kafkas et al. [2015]	D	✓	R	F=.75–.96 o	RE+DI	L	S	M	Q,S
Ghavimi et al. [2016a]	D	✓	R	F=.84 d	DIST	S	F	M	O,Q
Ghavimi et al. [2016b]	D		R	F=.84 d	DIST	S	F	M	O,Q
Zhang et al. [2016]	D		R	F=.49 s	IB	G	E*	M	O,Q
Grechkin et al. [2017]	D	✓	R	P=.97 e	RE+DI	L	F	J	Q
Li and Yan [2018]	P	✓	R	F=.86 e	RB	L	F	O	O,Q
Greuel and Sperber [2014]	S		R	-	RB	N	A	M	C
Pan et al. [2015]	S		R	F=.58 e	IB	L	M	I	Q
Pan et al. [2016]	S		R	F=.58 e	IB	L	M	I	O,Q
Chrapary et al. [2017]	S		R	-	RB	N	A	M	C
Duck et al. [2013]	S,DB		R	F=.53 e	DI+RB	L	F	M	Q
Névéol et al. [2011]	D		S	F=.80 s	SVM,NB	L	F	J	Q,S
Lu et al. [2012]	D		S	F=.69–.75 p	SVM*	G	E	M	O,Q
Duck et al. [2015]	S,DB		S	F=.63 e	CRF	L	F	O	Q
Duck et al. [2016]	S,DB		S	F=.67 e	RF	L	F	U	Q

Bibliography

A. Allen, P. J. Teuben, and P. W. Ryan. Schroedinger's code: A preliminary study on research source code availability and link persistence in astrophysics. *The Astrophysical Journal Supplement Series*, 236(1):10, may 2018. doi: 10.3847/1538-4365/aab764.

A. A. Alsheikh-Ali, W. Qureshi, M. H. Al-Mallah, and J. P. A. Ioannidis. Public availability of published research data in high-impact journals. *PLoS ONE*, 6(9):e24357, sep 2011. doi: 10.1371/journal.pone.0024357.

C. W. Belter. Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS ONE*, 9(3): e92590, mar 2014. doi: 10.1371/journal.pone.0092590.

K. Boland, D. Ritze, K. Eckert, and B. Mathiak. Identifying references to datasets in publications. In *International Conference on Theory and Practice of Digital Libraries*, pages 150–161. Springer, 2012.

H. Chrapary, W. Dalitz, W. Neun, and W. Sperber. Design, concepts, and state of the art of the swmath service. *Mathematics in Computer Science*, 11(3):469–481, Dec 2017. ISSN 1661-8289. doi: 10.1007/s11786-017-0305-5. URL `https://doi.org/10.1007/s11786-017-0305-5`.

A. Coppin. Finding science and engineering specific data set usage or funding acknowledgements. 2013. doi: 10.5062/f4cv4fp0.

G. Duck, G. Nenadic, A. Brass, D. L. Robertson, and R. Stevens. bioNerDS: exploring bioinformatics' database and software use through literature mining. *BMC bioinformatics*, 14(1):194, 2013.

G. Duck, A. Kovacevic, D. L. Robertson, R. Stevens, and G. Nenadic. Ambiguity and variability of database and software names in bioinformatics. *Journal of biomedical semantics*, 6(1):29, 2015.

G. Duck, G. Nenadic, M. Filannino, A. Brass, D. L. Robertson, and R. Stevens. A survey of bioinformatics database and software usage through mining the literature. *PloS one*, 11(6):e0157989, 2016.

B. Ghavimi, P. Mayr, C. Lange, S. Vahdati, and S. Auer. A semi-automatic approach for detecting dataset references in social science texts. *Information Services & Use*, 36(3-4):171–187, 2016a.

B. Ghavimi, P. Mayr, S. Vahdati, and C. Lange. Identifying and improving dataset references in social sciences full texts. *arXiv preprint arXiv:1603.01774*, 2016b.

M. Grechkin, H. Poon, and B. Howe. Wide-open: Accelerating public data release by automating detection of overdue datasets. *PLoS biology*, 15(6):e2002477, 2017.

G.-M. Greuel and W. Sperber. swMATH – an information service for mathematical software. In H. Hong and C. Yap, editors, *Mathematical Software – ICMS 2014*, pages 691–701, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg. ISBN 978-3-662-44199-2.

M. Haeussler, M. Gerner, and C. M. Bergman. Annotating genes and genomes with DNA sequences extracted from biomedical articles. *Bioinformatics*, 27(7):980–986, feb 2011. doi: 10.1093/bioinformatics/btr043.

T. Henderson and D. Kotz. Data citation practices in the CRAWDAD wireless network data archive. *D-Lib Magazine*, 21(1/2), jan 2015. doi: 10.1045/january2015-henderson.

J. Howison and J. Bullard. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, 67(9): 2137–2155, 2016.

Y.-H. Huang, P. W. Rose, and C.-N. Hsu. Citing a data repository: A case study of the protein data bank. *PLOS ONE*, 10(8):e0136631, aug 2015. doi: 10.1371/journal.pone.0136631.

Ş. Kafkas, J.-H. Kim, and J. R. McEntyre. Database citation in full text biomedical articles. *PloS one*, 8(5):e63184, 2013.

Ş. Kafkas, J.-H. Kim, X. Pi, and J. R. McEntyre. Database citation in supplementary data linked to europe pubmed central full text biomedical articles. *Journal of biomedical semantics*, 6(1):1, 2015.

P. W. Kirlew. Life science data repositories in the publications of scientists and librarians. 2011. doi: 10.5062/f4x63jt2.

J. Li, S. Zheng, H. Kang, Z. Hou, and Q. Qian. Identifying scientific project-generated data citation from full-text articles: An investigation of TCGA data citation. *Journal of Data and Information Science*, 1(2):32–44, 2016a.

K. Li and E. Yan. Co-mention network of R packages: Scientific impact and clustering structure. *Journal of Informetrics*, 12(1):87–100, 2018.

K. Li, X. Lin, and J. Greenberg. Software citation, reuse and metadata considerations: An exploratory study examining LAMMPS. *Proceedings of the Association for Information Science and Technology*, 53(1):1–10, 2016b.

K. Li, E. Yan, and Y. Feng. How is R cited in research outputs? structure, impacts, and citation standard. *Journal of Informetrics*, 11(4):989–1002, 2017.

M. Lu, S. Bangalore, G. Cormode, M. Hadjieleftheriou, and D. Srivastava. A dataset search engine for the research document corpus. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1237–1240. IEEE, 2012.

N. Mahrholz, A. Reinhold, and M. Rittberger. Data citation quantity and quality in research output of a large-scale educational panel study. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, page 31. ACM, 2015.

G. R. Major. Impact of NASA EOS instrument data on the scientific literature: 10 years of published research results from Terra, Aqua, and Aura. 2011. doi: 10.5062/f4cc0xmj.

H. Mooney. Citing data sources in the social sciences: do authors do it? *Learned Publishing*, 24(2):99–108, apr 2011. doi: 10.1087/20110204.

H. Mooney and M. P. Newton. The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship & Scholarly Communication*, 1(1), 2012.

U. Nangia and D. S. Katz. Understanding software in research: Initial results from examining nature and a call for collaboration. *arXiv preprint arXiv:1706.06527*, 2017.

A. Névéol, W. J. Wilbur, and Z. Lu. Extraction of data deposition statements from the literature: a method for automatically tracking research results. *Bioinformatics*, 27(23):3306–3312, 2011.

S. A. Ochsner, D. L. Steffen, C. J. Stoeckert Jr, and N. J. McKenna. Much room for improvement in deposition rates of expression microarray datasets. *Nature Methods*, 5(12):991, 2008.

X. Pan, E. Yan, Q. Wang, and W. Hua. Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics*, 9(4):860–871, 2015.

X. Pan, E. Yan, and W. Hua. Disciplinary differences of software use and impact in scientific literature. *Scientometrics*, 109(3):1593–1610, 2016.

X. Pan, E. Yan, M. Cui, and W. Hua. Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools. *Journal of Informetrics*, 12(2):481–493, 2018. doi: 10.1016/j.joi.2018.03.005.

H. Park and D. Wolfram. An examination of research data sharing and re-use: implications for data citation practice. *Scientometrics*, 111(1):443–461, 2017.

A. Pepe, A. Goodman, A. Muench, M. Crosas, and C. Erdmann. How do astronomers share data? reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLoS ONE*, 9(8):e104798, aug 2014. doi: 10.1371/journal.pone.0104798.

H. A. Piwowar, J. D. Carlson, and T. J. Vision. Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–4, 1 2011. ISSN 1550-8390. doi: 10.1002/meet.2011.14504801337. URL `https://doi.org/10.1002/meet.2011.14504801337`.

A. Prlic, M. A. Martinez, D. Dimitropoulos, B. Beran, B. T. Yukich, P. W. Rose, P. E. Bourne, and J. L. Fink. Integration of open access literature into the rcsb protein data bank using biolit. *BMC bioinformatics*, 11:220, Apr 2010.

P. H. Russell, R. L. Johnson, S. Ananthan, B. Harnke, and N. E. Carlson. A large-scale analysis of bioinformatics code on GitHub. *PLOS ONE*, 13(10):e0205898, oct 2018. doi: 10.1371/journal.pone.0205898.

M. Servilla, J. Brunt, D. Costa, J. McGann, and R. Waide. The contribution and reuse of LTER data in the provenance aware synthesis tracking architecture (PASTA) data repository. *Ecological informatics*, 36:247–258, 2016.

J. E. Sieber and B. E. Trumbo. (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics*, 1(1):11–20, mar 1995. doi: 10.1007/bf02628694.

A. Singhal and J. Srivastava. Data extract: Mining context from the web for dataset extraction. *International Journal of Machine Learning and Computing*, 3(2):219, 2013. doi: 10.7763/IJMLC.2013.V3.306.

A. Yan and N. Weber. Mining open government data used in scientific research. In *International Conference on Information*, pages 303–313. Springer, 2018.

Q. Yu, Y. Ding, M. Song, S. Song, J. Liu, and B. Zhang. Tracing database usage: Detecting main paths in database link networks. *Journal of Informetrics*, 9(1):1–15, jan 2015. doi: 10.1016/j.joi.2014.10.002.

W. Zenk-Möltgen and G. Lepthien. Data sharing in sociology journals. *Online Information Review*, 38(6):709–722, sep 2014. doi: 10.1108/oir-05-2014-0119.

Q. Zhang, Q. Cheng, Y. Huang, and W. Lu. A bootstrapping-based method to automatically identify data-usage statements in publications. *Journal of Data and Information Science*, 1(1):69–85, 2016.

M. Zhao, E. Yan, and K. Li. Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, 2018.