**Theories of Citations**

**Data Citation**

**Software Citation**

**Moravcsik, M. J., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, *5*(1), 86-92. doi:10.1177/030631277500500106**

"1. Is the reference conceptual or operational? In other words, is the reference made in connection with a concept or theory that is used in the referring paper, or is it made in connection with a tool or physical technique used in the referring paper? The distinction is not meant to be a value judgment, and is not to be taken as synonymous with judging the importance of the paper referred to.

"2. Is the reference organic or perfunctory? In other words, is the reference truly needed for the understanding of the referring paper (or to the working out of the content of that paper), or is it mainly an acknowledgment that some other work in the same general area has been performed?

"3. Is the reference evolutionary or juxtapositional? In other words, is the referring paper built on the foundations provided by the reference, or is it an alternative to it?

"4. Is the reference confirmative or negational? In other words, is it claimed by the referring paper that the reference is correct, or is its correctness disputed? Incorrectness need not be claimed through an actual demonstration of an error in the paper referred to, but could also be established, for example, through inferior agreement with experimental data."

--------------------

**Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, *40*(4), 284–290. Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(198907)40:4%3C284::AID-ASI10%3E3.0.CO;2-Z/abstract**

Authors of 42 technical papers (published in two journals of the American Society of Civil Engineers) categorized the citations in their own papers using Moravcsik classification model (above).

"The research subjects reported multiple use of cited documents. This was in conflict with the model since citations were paired within the presumed dichotomous categories. In particular there were 29 instances of cited documents reported to have both a conceptual (theoretical) and an operational (instrumental) nature."

"On the basis of the data collected, perfunctory and negational citations (first citation type group), were given a low utility level and exhibit similar distribution shapes, Fig. l(a). However, in spite of the reportedly low utility level exhibited by the perfunctory type, research subjects reported it nonetheless, to be the most frequent category (26%)."

"Citations tended to be concentrated in three areas of the paper defined as: beginning section (up to the 15 percentile), middle section (from the 20th to the 75th percentile), and end (from the 80th percentile). The largest concentration was located in the first 15% of the paper; this agrees with citation densities reported by Voos and Dagaev [lo]. Perfunctory citations comprised the largest category of citations in this section (33%). Organic citations (32%), comprise the largest category in the middle section of the paper. This section shows a decline around the twentieth per cent mark of the paper and minor peaks occur near the fortieth and sixtieth per cent. Interestingly, organic citations also comprise the largest category at the end of the paper (4 1%) . Overall, citation frequency steadily declined after the fifteenth per cent location mark."

"Citations seem to be more heavily concentrated in the first 15% of the technical papers studied, corresponding roughly to their introductory sections. However, over one third of the citations in this section were of a perfunctory nature and of very little use to the research subjects in their production of novel information. This result might lead to the hypothesis that citations located at introductory sections of technical papers represent a mere “setting of the stage” and have very little informational utility to the authors of the papers. This does not mean that citations in these sections are worthless in information content to all readers, Merely that to elite scientists citations in introductory sections might present exclusively the background literature in relation to which their new contribution is to be examined."

"The proposition concerning the visibility of intellectual influence as reflected by citations has recently been contested by MacRoberts and MacRoberts [ 131. Their study points out the inability of citations to capture intellectual influences embedded in inarticulated knowledge. In other words, citations can only reflect influences that are specifically acknowledged as such, by the authors of a document; citations do not reflect influences that belong to the “intellectual baggage”, accumulated by years of intellectual pursuit. In other words, citations do not capture information or knowledge that has been integrated into the mental constructs of a writer to the extent that it has become inarticulated, or tacit,"

----------------------------------------------------------------------------

**Lipetz, B.-ami. (1965). Improvement of the Selectivity of Citation Indexes to Science Literature Through Inclusion of Citation Relationship Indicators. *American Documentation*, *16*(2).**

TABLE 1: Citation Relationships in Science Literature

GROUP ONE. Original Scientific Contribution or Intent of Citing Paper

1. Description of observed phenomena

2. Data transformation

3. Explanation

4. Hypothesis or theory

5. Calculation from theory

6. Prediction

7. Definition or notation

8. Statement of experimental technique

GROUP TWO. Contribution of Citing Paper Other than Origi- nal Scientific Contribution

9. Review article

10. Bibliography

11. Data cumulation

GROUP THREE. Identity or Continuity Relationship of Citing Paper to Cited Paper

12. One or more authors in common

13. Same text

14. Abstract or condensation

15. Erratum

16. Continuation

17. Precursor

18. Inclusion

GROUP FOUR. Disposition of the Scientific Contribution of the Cited Paper in the Citing Paper

19. Noted only

20. Distinguished

21. Reviewed or compared

22. Applied

23. Improved or modified

24. Replaced

25. Changed the precision (plus or minus)

26. Changed the scope of applicability (plus or minus)

27. Questioned

28. Affirmed

29. Refuted

------------------------------------------------

**Pham, S. B., & Hoffmann, A. (2003). A New Approach for Scientific Citation Classification Using Cue Phrases. In *AI 2003: Advances in Artificial Intelligence*(pp. 759-771).**

In this paper, we used four citation types, namely Basis, Support, Limitation and Com-parison. In [9] 15 different citation types were distinguished. Our citation types appear to be the most relevant among those 15 types for the purpose of a citation map that is designed to support obtaining an overview of a field of research.

– Basis: One work is based on another work.

– Support: One work is supported by another work.

– Limitation: One work has been criticized to have some limits or weaknesses.

– Comparison: Two approaches are compared.

These citation types arguably reflect the most important relationships between papers that users are interested to know. They had been chosen before we acquired the corpus used in our experiment.

-------------------------------------

**Teufel, S., & Tidhar, D. (2006). Automatic classification of citation function. *Computational Linguistics*, (July), 103-110.**

Category Description

Weak -- Weakness of cited approach

CoCoGM -- Contrast/Comparison in Goals or Meth- ods(neutral)

CoCo -- Author’s work is stated to be superior to cited work

CoCoR0 -- Contrast/Comparison in Results (neutral)

CoCoXY -- Contrast between 2 cited methods

PBas -- Author uses cited work as basis or starting point

PUse -- Author uses tools/algorithms/data/definitions

PModi -- adapts or modifies tools/algorithms/data

PMot -- This citation is positive about approach used or problem addressed (used to mo- tivate work in current paper)

PSim -- Author’s work and cited work are similar

PSup -- Author’s work and citedwork are compat- ible/provide support for each other

Neut - Neutral description of cited work, or not enough textual evidence for above cate- gories, or unlisted citation function

"As expected, the distribution is very skewed, with more than 60% of the citations of category Neut.5 What is interesting is the relatively high frequency of usage categories (PUse, PModi, PBas) with a total of 18.9%. There is a relatively low frequency of clearly negative citations (Weak, CoCo-, total of 4.1%), whereas the neutral– contrastive categories (CoCoR0, CoCoXY, CoCoGM) are slightly more frequent at 7.6%. This is in concordance with earlier annotation experiments (Moravcsik and Murugesan, 1975; Spiegel-R¨ using, 1977)."

"Our categories are as follows: One category (Weak) is reserved for weakness of previous re- search, if it is addressed by the authors. The next four categories describe comparisons or contrasts between own and other work. The difference be- tween them concerns whether the contrast is be- tween methods employed or goals (CoCoGM), or results, and in the case of results, a difference is made between the cited results being worse than the current work (CoCo-), or comparable or bet- ter results (CoCoR0). As well as considering dif- ferences between the current work and other work, we also mark citations if they are explicitly com- pared and contrasted with other work (i.e. not the work in the current paper). This is expressed in category CoCoXY. While this is not typically annotated in the literature, we expect a potential practical benefit of this category for our applica- tion, particularly in searches for differences and rival approaches.

"The next set of categories we propose concerns positive sentiment expressed towards a citation, or a statement that the other work is actively used in the current work (which we consider the ulti- mate praise). We mark statements of use of data and methods of the cited work, differentiating un- changed use (PUse) from use with adaptations (PModi). Work which is stated as the explicit starting point or intellectual ancestry is marked with our category PBas. If a claim in the liter- ature is used to strengthen the authors’ argument, or vice versa, we assign the category PSup. We also mark similarity of (an aspect of) the approach to the cited work (PSim), and motivation of ap- proach used or problem addressed (PMot).

"Our twelfth category, Neut, bundles truly neutral descriptions of cited work with those cases where the textual evidence for a citation function was not enough to warrant annotation of that category, and all other functions for which our scheme did not provide a specific category."

**Gambardella, A., & Hall, B. H. (2006). Proprietary versus public domain licensing of software and research products. Research Policy, 35(6), 875–892. doi:10.1016/j.respol.2006.04.004**

“Production of public knowledge good” operationalized as research software or database

Focusing on impact of GPL use on scientific software in the production of public good

Specific setting looks at econometric software used by academic researchers and private firms

Particularly concerned with ideas of investment, pricing, and incentives

Sees some cases of the production and distribution of scientific software as occurring counter to a “Republic of Science” ideal

Concerned with the incentives, market, and non-market impacts on the production and distribution of “knowledges” such as scientific software

Considering a tradeoff between protecting “upstream” knowledge (the investment in the production of the software) and “downstream” knowledge (the scientific outputs of disseminating the software)

Sees scientific software as more likely to be privatized than scientific research in general because:

* market demand
* weaker reputation effects produce weaker norms
* more users who are not themselves producers
* possibility of simultaneous public and private development
* diluted benefits to the original producers

Drawbacks of privatization of scientific software

Possibility of GPL solving the problem

\

**Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine, 13*(3/4). Retrieved from** [**http://papers.ssrn.com/sol3/papers.cfm?abstract\_id=1081955**](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1081955)

Outlines importance of having a common standard for data citation, based on the importance of standard citation forms for other publication types

Lists types of inconsistencies and resulting problems from unregulated data citation practices - particularly in finding the dataset ever again (assumption is that interfering with access then risks possibility of replicability)

Minimal citation standard has six required components:

* author(s)
* date published
* data set title
* unique global identifier
* universal numeric fingerprint
* bridge service

Also has optional elements

Discusses possibility of “deep citation” of datasets - standards for citing parts of a dataset

**Polhill, I., & Edmonds, B. (2007). Open access for social simulation. *Journal of Artificial Societies and Social Simulation,* *10*(3)**.

Concerned with software licensing in science - the default isn’t appropriate for the “scientific domain”

Outline the rights that should be used in licensing for scientific software (doesn’t name a particular license)

Focusing on scientific simulations - concerned with their opacity, seeming transparency, and persuasiveness (therefore, if they aren’t open, science is at risk)

Goes beyond concerns of replicability - also concerned with comparing simulations and making incremental improvements and equality of access

Simulations seems to be a type of scientific software in which the bias or view of the creator is more obvious/accepted - they discuss how simulations are “not value-free” but encode the authors’ assumptions

Their recommendation includes the following rights:

* unrestricted right to run the software
* right to re-implement the model
* right to modify the source code
* right to distribute the modified version
* copyleft protection

Also have recommendations for documentation of code

**Mayernik, M. S. (2012). Data citation initiatives and issues. *Bulletin of the American Society for Information Science and Technology,* 38(5), 23–28. doi:10.1002/bult.2012.1720380508**

Concerned with lack of individual incentives and “pervasive cultural inertia” acting against recognized importance/impact of data citation

Linked data provides additional leverage towards data citation practices

Presents optimistic account of current state of data citatio

**Konkiel, S. (2012). Tracking citations and altmetrics for research data: Challenges and opportunities.**

Pointing out the promising effect of altmetrics on facilitating data citation and making datasets findable

For greater access/findability, need to standardize data sets and system architecture

**Borgman, C. (2012). The conundrum of sharing research data.**

Presents four reasons to share research data:

* to reproduce/verify research
* to make results of publicly funded research available to the public
* to enable others to ask new questions of extant data
* to advance the state of research and innovation

Maps each of these on axis of beneficiaries (data producers or data users) and the nature of the argument (research- or public-driven)

Each rationale (and its relationship to beneficiaries and arguments) produces different policies and practice

(How many of these map onto sharing scientific software? Last one is particularly relevant to the use open source and modifiable programs)

Responding to “the data deluge”

Section on the urgency of data sharing (can map these onto software)

* requirements of funding agencies, in US & UK
  + identifies NSF data management requirements as “the tipping point”
* requirements of journals
  + *Science* introduces more rigorous requirements, including computer code (Hanson et al., 2011)
* movements towards standardization of data citation & attribution (CODATA task group report, 2010)

Outlines the intellectual and physical labour involved in the production, organization, and analysis of data

Section on stakeholders’ motivations/incentives to share data

Worth noting - points out that humanities also have data sharing issues, but the sciences and social sciences “are on the front lines of current policy debates” (1067)

Discusses each of the rationales in detail

Identifies software use as an additional barrier to replication - the data are transformed in many ways by software and proprietary software used is not always available

Summary of issues arising from software use: “*Numerical data are of little value without the software associated with the data collection, analysis, and processing technologies. That software may be proprietary or it may be crafted locally as part of the research project. In either case, the software necessary to interpret the data may not be sharable. Even if the data were produced with common tools, those data may be unreadable after a few years because of changes in hardware and software, unless they have been curated well and migrated to new technologies*” (1073)

Brings up the issue of what is meant by “the data” - the entire dataset, the cleaned data, a subset of the data - has parallels with software, where specific algorithms or versions might be relevant

**Simons, N., Visser, K., & Searle, S. (2013). Growing institutional support for data citation. *D-Lib,***

Characterizes data citation as “an emergent practice” rather than a norm, a “relatively new concept”

Article is a case study of a project by Griffith University and the Australian National Data Service to provide DOI infrastructure for data produced by the university.

Describes purpose of data citation as: “identification, retrieval, replication, and verification of data underlying published studies”

Names the Dryad Data Repository (framework for making datasets available alongside journal articles) “more the exception than the rule” in terms of necessary infrastructure for data citation and access

Other than the functions enabled by dataset DOIs, they also point out the respect that they add: When applied to data, they indicate that data is to be treated with the same respect as publications — to be well managed, persistently available and cited over the scholarly lifecycle of the research.”- value beyond citation “functions”

Project brought up issues of versioning, granularity, metadata requirements, and citation format

Difficult to gauge impacts of the framework without wide mandates for data deposits and novelty of data citation as a practice - also requires that researchers and librarian receive training

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

Good parallel article for our analysis - similar approach, possibly more optimistic results (with regards to how many journals had relevant policies & how many articles gave the amount of information necessary for citation functions)

Approaches data citation as necessary for a reward structure, which is in turn necessary for data sharing

Study is a content analysis of journal articles, author instructions, style manuals, data publishers

Looking for disciplinary and cultural norms regarding data citation

Identifies “multiplicity of data types” and “lack of awareness regarding existing standards” as causes of the results

Notes that the data sharing movement has revealed inadequacies in technological infrastructure and cultural practices

Lit review includes historical references - articles from 70’s and 80’s noting lack of rewards for data sharing (in terms of reputation, CV); also cites more recent studies showing the same barriers exist

Updating a similar study from more than 15 years prior - doesn’t find much improvement

Developed a Data Citation Adequacy Index from existing standards to measure citation practices against best practices for citation functions - scores based on “completeness of the data citation given”; gives weights to different possible elements of a citation

Sample of articles across Humanities, Social Science, and Scienc databases (WilsonWeb) - only those which had some mention of a dataset

Most DCAI scores below 5 - only had 2-3 elements of a complete (36-point) citation (usually the name /title of the dataset in the body of the article, sometimes a date in the reference list)

Only half of the journals had any mention/instruction for data citation (in style manual or otherwise)

References a chicken/egg dilemma with data citation and data sharing as cultural or disciplinary practices