*Information Technology and Libraries* 22(4): 183-187, December 2003

# Mapping the Output of Topical Searches in the *Web of Knowledge* and the Case of Watson-Crick

Eugene Garfield, A.I. Pudovkin and V.I. Istomin

*HistCite*™ is a system that generates chronological maps of subject (topical) collections resulting from searches of the *Institute for Scientific Information Web of* Science® (*WoS*) or *Science Citation Index, Social Sciences Citation Index, and Arts and Humanities Citation Index on CD-ROM.* *WoS* export files are created in which all cited references for source documents are captured. These bibliographic collections are processed by *HistCite,* which generates chronological tables as well as historiographs that highlight the most-cited works in and outside the collection. Articles citing the 1953 primordial Watson-Crick paper on the structure of DNA will be used as a demonstration. Real-time dynamic genealogical historiographs will be shown. *HistCite*™ also includes a module for detecting and editing errors or variations in cited references. Export files of five thousand or more records are processed in minutes on a PC. Ideally the system will be used to help the searcher quickly identify the most significant work on a topic and enable the searcher to trace its year-by-year historical development.

The *HistCite*™ system resulted from a long-term needs assessment of users of bibliographic databases. Librarians and users need to identify the key works on a particular subject, while scholars and editors desire rapid historical reviews of new topics. *HistCite*™ is designed to satisfy both of these requirements, whether for reference service or in writing review articles or historical introductions to new manuscripts. It uses a visual data-mining method based on the analysis of citation links between various documents in an academic library, making it appropriate as a topic for bibliomining.

Eugene Garfield (garfield@codex.cis.upenn.edu) is Chairman Emeritus, Thomson ISI, Philadelphia.
A.I. Pudovkin (aipud@online.ru) is Chief Scientist at the Institute of Marine Biology, Russian Academy of Sciences, Vladivostok, Russia
V.S. Istomin (vi@mall.wsu.edu) is a Systems Analyst at the Center for Teaching, Learning, and Technology, Washington State University

Even before the advent of the Science Citation Index (SCI), the use of citation data was discussed in the 1964 report "The Use of Citation Data in Writing the History of Science."[1] This included a historiograph that sketched the history of DNA from Gregor Mendel in 1865 to Marshall Nirenberg in 1961, through various stages including Avery-McCleod-McCarty in 1944, through Watson-Crick in 1953. Flow charts of the papers were created manually, based solely on the references cited in a set of core source papers identified in a book by Isaac Asimov on the genetic code. This gave rise to speculation on the potential use of citation indexes for historiography.[2] Similar maps were later created by Tony Cawkell.[3]

In an information retrieval course taught at the University of Pennsylvania Moore School of Electrical Engineering, students were required to create similar topical historiographs. It was believed that these historiographs would aid in studying the contemporary history of science. Since history and bibliography were intimately linked, the term "historiobibliography" was coined.[4]

A frequent topic of discussion during the DNA mapping project was the idea of writing computer programs that would create such maps directly from the electronic files of SCI. There was a possibility that this would require random access to ISI's massive files so that cited and citing documents could be retrieved in real time. In the 1960s, however, low-cost gigabyte memories were still a dream. The implementation of real-time mapping had to wait for the time when computer memories were large and cheap enough to handle retrospective files covering many decades of literature. Though the PC had not yet come along, online searches were possible in the 1970s; mapping in real time, however, was still not feasible. Only very recently, when PCs could handle the output of a completely linked large file of thousands of records, did the creation of historiographs in real time become feasible.

There have been many different types of mapping exercises performed on a relatively small scale. In the past, co-citation clustering required mainframe computers and, in most cases, still does.[5] These ideas were later extended to creating small cluster maps online as in the SciMap system developed by Small at ISI. In that system, a starting paper is used to seed the creation of a co-citation cluster map.[6] In spite of the many mapping and visualization reports in the literature, none were applied to the creation of historiographs. Further, none of the many authors on co-citation mapping considered the potential significant relationship between historical displays and the need of reference librarians and users to evaluate the output of literature searches with such sources as SCI, Medline, or Chemical Abstracts.

Until quite recently, the authors thought of creating historiographs primarily by seeding one or two primordial papers. SCI on CD-ROM was used to trace forward in time papers that had cited the starting papers. This is the essence of the now traditional cited-reference search. Since the basic purpose of a historiograph is to display the chronological development of a topic or field year by year from the earliest papers to the most recent, the annual SCI was searched in the same way. Once the initial group of citing papers was identified, further cited-reference searches were done on them, a process sometimes called citation chaining. This process was iterated for as many years of the literature as was necessary, as will be illustrated later in this paper in discussing of the work of Watson and Crick.

## HistCite Program

In HistCite, two types of frequency are distinguished—local citation score (LCS) and global citation score (GCS). LCS (or frequency) is the number of times an item is cited in the retrieved collection. The GCS is the global score, the number of cites in the entire SCI/ Social Sciences Citation Index (SSCI). The record for each source document contains both its LCS and GCS. Once HistCite sorts the papers by citation score, the user will select a group cited above an arbitrarily chosen threshold to be mapped. If there are five hundred source papers, then a 5 percent selection threshold would produce twenty-five core papers. These core papers should be of prime interest especially to a searcher who is not familiar with the subject matter. Ordinarily, one would examine this core list first in reviewing a new topic. The coordinates of these papers are used to create a historiograph of the topic that displays the papers and their citation links chronologically.

### Identifying Core Literatures

The authors initially assumed that the search would begin with one primordial paper. However, it became apparent that groups of papers could be fed in by one or more authors—and by extension, larger clusters of papers by institution or by key word. Thus, the output of any conventional search or a combination of citation and key word searches could be input to the system. Once the input bibliography is created, the core papers on the topic can be rapidly identified.

### Outer References

HistCite also produces frequency-ranked tables of outer references, that is, cited papers and books that fall outside the retrieved original collection. These are works that do not turn up in the original WoS search but, significantly, are cited frequently in the papers that are retrieved. The searcher can examine these candidate references and decide whether to add them to the initial collection.

## Watson-Crick 1953 DNA Paper

The following will illustrate the use of HistCite in a truly historical mapping exercise. This year marks the fiftieth anniversary of the Watson-Crick discovery of the double helix structure of DNA.[7] That 1953 paper was used to conduct a cited reference of SCI. The result of mapping the five years from 1953 to 1958 can be seen in figures 1 through 4.

Figure 1 shows the usual HistCite table for the papers that cite the 1953 Watson-Crick paper, with the addition of a few of the key outer references for Avery and other papers.[8]

Figure 2 represents the table for the 975 papers retrieved by virtue of chaining citations to the 210 papers that cited Watson-Crick. In other words, these are second-generation citations to the citing papers in figure 1.

In Figure 3, the year-by-year map of the twenty-two most-cited papers in the chained indexed file can be seen. Notice that there were nine highly cited papers in 1953. And in 1954, there are five. Using the typical reference citation—that is, only author, volume, page, and year—it is not possible to differentiate the month-by-month progression after the Watson-Crick paper of April 1953. However, the HistCite system can take into account the cover dates of the journals involved if they are included.

In a recent interview, James Watson regretted that he had not cited Avery et al. in the 1953 Nature paper.[9] However, this omission made little difference in the HistCite algorithmic mapping exercise. To emphasize that link, a dotted line has been inserted linking node #2 to node #27.

Figure 4 illustrates how the historiograph changes not just year-to-year but month-by-month. Unfortunately, the Institute for Scientific Information Web of Science (WoS) does not contain cover dates until 1985, so it was necessary to manually insert in the export files the cover dates for the few dozen papers involved in this example.

Outer References Missing Links? Journal list All-Author list Citation Matrix Graphs HistCite Guide

**Articles from 1953-1958 citing Watson and Crick's 1953 paper, "Molecular Structure of DNA" and selected outer references**

Nodes: 210
Sorted by **year, journal, volume, page**.
Page **1**: 1

| # | Cited nodes | Nodes / Authors | GCS | LCS |
|---|---|---|---|---|
| 1 | 0 | 1 1944 JOURNAL OF EXPERIMENTAL MEDICINE 79():137-157<br>**AVERY OT; MACLEON CM; MCCARTY M**<br>*Studies on the Chemical Nature of the Substance Inducing Transformation of Pseumococcal Types Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III* | 0 | 23 |
| 2 | 0 | 2 1952 JOURNAL OF GENERAL PHYSIOLOGY 36(1):39-56<br>**HERSHEY AD; CHASE M**<br>*Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage* | 747 | 23 |
| 3 | 2 | 3 1953 ACTA CRYSTALLOGRAPHICA 6(8-9):673-677<br>**FRANKLIN RE; GOSLING RG**<br>*The Structure of Sodium Thymonucleate Fibres .1. The Influence of Water Content* | 14 | 11 |
| 4 | 3 | 4 1953 ACTA CRYSTALLOGRAPHICA 6(8-9):678-685<br>**FRANKLIN RE; GOSLING RG**<br>*The Structure of Sodium Thymonucleate Fibres .2. The Cylindrically Symmetrical Patterson Function* | 10 | 8 |
| 5 | 1 | 5 1953 ARCHIVES OF BIOCHEMISTRY AND BIOPHYSICS 46(1):12-17<br>**SMITH CL**<br>*The Breakdown of Desoxyribonucleic Acid Under Deuteron and Electron Bombardment* | 5 | 1 |
| 6 | 2 | 6 1953 BIOCHEMICAL JOURNAL 55(5):774-782<br>**WYATT GR; COHEN SS**<br>*The Bases of the Nucleic Acids of Some Bacterial and Animal Viruses - The Occurrence of 5-Hydroxymethylcytosine* | 57 | 8 |
| 7 | 3 | 7 1953 COLD SPRING HARBOR SYMPOSIA ON QUANTITATIVE BIOLOGY 18():123-131<br>**WATSON JD; CRICK FHC**<br>*The Structure of Dna* | 61 | 21 |
| 8 | 1 | 8 1953 COLD SPRING HARBOR SYMPOSIA ON QUANTITATIVE BIOLOGY<br>8():133-134<br>**WYATT GR**<br>*The Quantitative Composition of Deoxypentose Nucleic Acids As Related To the Newly Proposed Structure* | 9 | 4 |
| 9 | 2 | 9 1953 COLD SPRING HARBOR SYMPOSIA ON QUANTITATIVE BIOLOGY 18():171-183<br>**LARK KG; ADAMS MH**<br>*The Stability of Phages As a Function of the Ionic* | 13 | 2 |

**Figure 1.** HistCite Chronological Listing of Papers Citing Watson-Crick 1953

## Conclusion

The authors have described a tool that permits the user to manage the voluminous references produced in a

comprehensive search of the literature. For those who are new to the subject, the mere juxtaposition of the most-cited papers for each five- or ten-year period of the literature will help identify the key literature to be used first. For those who are knowledgeable in the field, the system will help jog the memory to recall the key works that were associated with the development of the field. While the relevance of citing works may be apparent, the collective bibliographic coupling and co-citation of papers in and outside of the basic bibliography should provide a comprehensive structure for completing a synoptic history of the topic.

This tool will prove extremely valuable for reference librarians in academic libraries and those producing bibliographic instruction on an unfamiliar topic. The librarian can use HistCite to produce a graphical view of citation data and can aid a patron in the identification of critical works regarding a topic. Collection development librarians can use the tool to look at the importance of articles from specific journals in order to make purchasing decisions. In addition, librarians doing their own research can use HistCite to understand the history of a research question and to aid in the identification of appropriate work for traditional evidence-based librarianship.

For more information on algorithmic historiography, a paper on this theme has been published in a special issue of the Journal of the American Society for Information Science and Technology on "Visualization of Scientific Paradigms."[10]

To learn more about HistCite and see other examples, visit http://garfield.library.upenn.edu/histcomp.

Outer References Missing Links? Journal list All-Author list Citation Matrix Graphs HistCite Guide

**Articles citing Watson and Crick's 1953 paper, "Molecular Structure of DNA", the articles citing them (1953-1958), and selected outer references**

Nodes: 975
Sorted by **year, journal, volume, page**
Page 1: 1 2

| # | Cited nodes | Nodes / Authors | GCS | LCS |
|---|---|---|---|---|
| 1 | 0 | 1 1938 JOURNAL OF BIOLOGICAL CHEMISTRY 124():425-<br>**SEVAG MG**<br>*[unknown]* | 216 | 37 |
| 2 | 1 | 2 1944 JOURNAL OF EXPERIMENTAL MEDICINE 79():137-157<br>**AVERY OT; MACLEON CM; MCCARTY M**<br>*Studies on the Chemical Nature of the Substance Inducing Transformation of Pseumococcal Types. Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III* | 331 | 43 |
| 3 | 0 | 3 1945 JOURNAL OF BIOLOGICAL CHEMISTRY 161(1):83-89<br>**SCHMIDT G; THANNHAUSER SJ**<br>*A Method for the Determination of Desoxyribonucleic Acid, Ribonucleic Acid, and Phosphoproteins in Animal Tissues* | 696 | 34 |
| 4 | 1 | 4 1945 JOURNAL OF BIOLOGICAL CHEMISTRY 161(1):293-303<br>**SCHNEIDER WC**<br>*Phosphorus Compounds in Animal Tissues .1. Extraction and Estimation of Desoxypentose Nucleic Acid and of Pentose Nucleic Acid* | 952 | 30 |
| 5 | 2 | 5 1946 JOURNAL OF GENERAL PHYSIOLOGY 30(2):117-&<br>**MIRSKY AE; POLLISTER AW**<br>*Chromosin, a Desoxyribose Nucleoprotein Complex of the Cell Nucleus* | 323 | 35 |
| 6 | 0 | 6 1947 JOURNAL OF THE CHEMICAL SOCIETY (SEP):1131-1141<br>**GULLAND JM; JORDAN DO; TAYLOR HFW**<br>*Deoxypentose Nucleic Acids .2. Electrometric Titration of the Acidic and the Basic Groups of the Deoxypentose Nucleic Acid of Calf Thymus* | 70 | 31 |
| 7 | 3 | 7 1951 BIOCHEMICAL JOURNAL 48(5):584-590<br>**WYATT GR**<br>*The Purine and Pyrimidine Composition of Deoxypentose Nucleic Acids* | 276 | 63 |
| 8 | 0 | 8 1951 JOURNAL OF BIOLOGICAL CHEMISTRY 189(2):597-605<br>**MARSHAK A; VOGEL HJ**<br>*Microdetermination of Purines and Pyrimidines in Biological Materials* | 136 | 30 |
| 9 | 0 | 9 1951 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA 37(4):205-211<br>**PAULING L; COREY RB; BRANSON HR**<br>*The Structure of Proteins - 2 Hydrogen-Bonded Helical Configurations of the Polypeptide Chain* | 185 | 26 |
| 10 | 1 | 10 1952 BIOCHEMICAL JOURNAL 52(5):558-565<br>**MARKHAM R; SMITH JD**<br>*The Structure of Ribonucleic Acids .2. The Smaller Products of Ribonuclease Digestion* | 104 | 28 |
| 11 | 0 | 11 1952 JOURNAL OF GENERAL PHYSIOLOGY 36(1):39-56<br>**HERSHEY AD; CHASE M**<br>*Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage* | 206 | |

**Figure 2.** Opening Page of HistCite Collection of "Chained" Citations to Watson-Crick 1953

## References


1. E. Garfield, I. H. Sher, and R. J. Torpie, The Use of Citation Data in Writing the History of Science: Report of Research for Air Force Office of Scientific Research Under Contract AF49 (638)-1256 (Philadelphia: The Institute for Scientific Information, 1964). Accessed Oct. 9, 2003, www.garfield.library.upenn.edu/paper/useofcitdatawritinghistofsci.pdf.

2. E. Garfield, "Citation Indexing, Historio-bibliography and the Sociology of Science Biography," Current Contents M25+, 1971. First published in K. E. Davis and W. D. Sweeney, eds., *Proceedings of the Third International Congress of Medical Librarianship* (Amsterdam: Excerpta Medica, 1969), 187–204. Accessed Oct. 9, 2003, www.garfield.library.upenn.edu/essays/V1p158y1962-73.pdf.

3. A. E. Cawkell, "Acoustic Journals and Acoustic Research Articles," *Current*

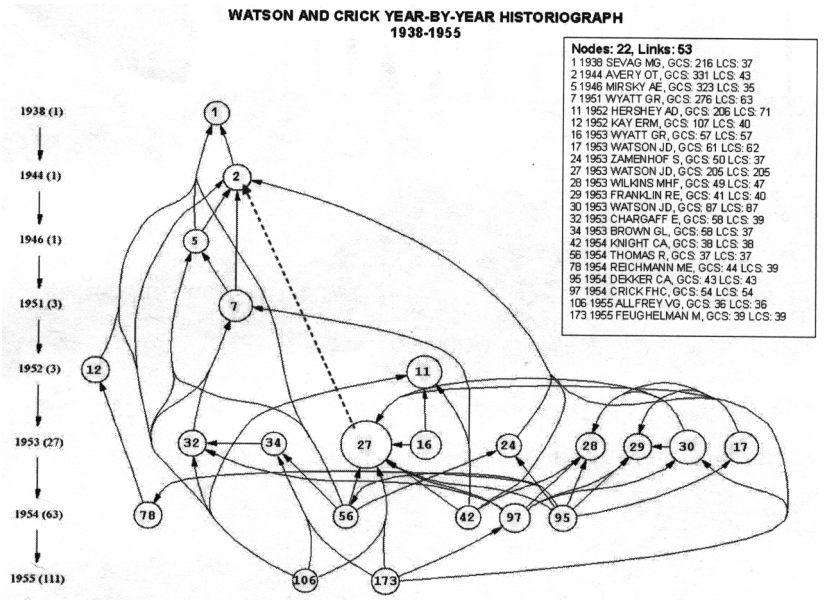


**Figure 3.** Year-by-Year Map of Watson-Crick 1938–1955

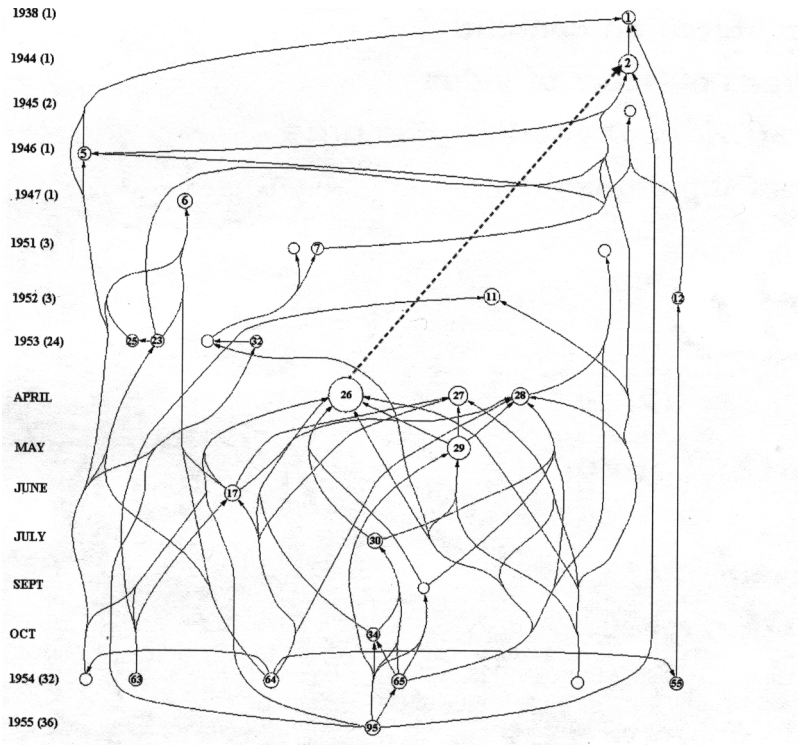


**Figure 4.** Month-by-Month Historiograph Linking Watson-Crick to Avery

*Contents* 44 (1989): 4–15. Reprinted in E. Garfield, *Essays of an Information Scientist*, Vol. 12 (Philadelphia: ISI Pr., 1977). Accessed Oct. 9, 2003, www.garfield.library.upenn.edu/essays/v12p301y1989.pdf. A. E. Cawkell, 2000. "Visualizing Citation Connections," in *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*, eds. B. Cronin and H. Barsky Atkins, (Medford, N.J.: Information Today, Inc.), 177–94.

**4.** E. Garfield, "Citation Indexing, Historio-Bibliography and the Sociology of Science Biography," *Current Contents 15 (1971)*, M25+.

**5.** E. Garfield, "Contract Research Services at ISI—Citation Analysis for Governmental, Industrial, and Academic Clients," *Current Contents* 23 (1992): 5–13. Accessed Oct. 9, 2003, http://garfield.library.upenn.edu/essays/v15p075y1992-93.pdf. H. Small and E. Garfield, "The Geography of Science: Disciplinary and National Mappings," *Journal of Information Science* 11 (1985): 147–59. Reprinted in *Current Contents* 43 (Oct. 27, 1986): 3–14; *Essays of an Information Scientist*, vol. 9 (Philadelphia: ISI Pr.) 324-35. Accessed Oct. 9, 2003, www.garfield.library.upenn.edu/essays/v9p325y1986.pdf.

**6.** H. Small, "A Sci-Map Case Study: Building a Map of AIDS Research," *Scientometrics* 30 (1994): 229–41. H. Small, E. Sweeney, and E. Greenlee, 1985. "Clustering the Science Citation Index Using Co-citations: 2. Mapping Science," *Scientometrics* 8 (1985): 321–34.

**7.** J. D. Watson and F. H. C. Crick, 1953. "Molecular Structure of Nucleic Acids—A Structure for Deoxyribose Nucleic Acid," *Nature* 171, no. 4356 (1953): 737–38.

**8.** O. T. Avery, C. M. Macleon, and M. McCarty, 1944. "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III," *Journal of Experimental Medicine* 79 (1944): 137–57.

**9.** Anonymous, "Genes, Girls and Honest Jim," *Bio-IT World* 2 no. 4 (2003): 28.

**10.** E. Garfield, A. I. Pudovkin, and V.S. Istomin, "Why Do We Need Algorithmic Historiography?" *Journal of the American Society for Information Science and Technology* 54, no. 5 (2003): 400–12. Accessed Oct. 9, 2003, http://garfield.library.upenn.edu/papers/jasist54(5)400y2003.pdf.