Essays of an Information Scientist: Journalology, KeyWords Plus, and other Essays, Vol:13, p.295, 1990
Current Contents, #32, p.3-7, August 6, 1990

# Current Comments®

EUGENE GARFIELD
INSTITUTE FOR SCIENTIFIC INFORMATION®
3501 MARKET ST PHILADELPHIA PA 19104

## *KeyWords Plus*: ISI's Breakthrough Retrieval Method. Part 1. Expanding Your Searching Power on *Current Contents on Diskette*

Number 32 August 6, 1990

*KeyWords Plus*℠, a new search capability, will be added to the upgraded *Current Contents on Diskette*®. *KeyWords Plus* provides search terms extracted from the titles of papers cited in each new article listed in *Current Contents*®. *KeyWords Plus* substantially augments title-word and author-keyword indexing. Also to be included in *Focus On: Global Change*™, *KeyWords Plus* is being developed for use with other ISI® products.

Almost two years ago, ISI® released *Current Contents on Diskette*® (*CC-on-Diskette*™), the electronic version of the print *Current Contents*® (*CC*®). Initially released for the Apple Macintosh, *CC-on-Diskette* soon made its debut in versions for the IBM PC and, most recently, for NEC computers used in Japan.[1-3] Keeping pace with rapid changes in computer technology and the information marketplace, and with invaluable responses and suggestions from our customers, ISI has continued to refine and improve *CC-on-Diskette*. We are pleased to announce a new upgrade that, in addition to various enhancements and improvements, will contain a truly exciting new capability.

### *KeyWords Plus*

In a comprehensive survey of *CC* readers, we learned that the enhancement most requested by readers is expanded keyword indexing. Consequently, *CC-on-Diskette*'s new upgrade will feature an enhancement that promises to lend a whole new dimension not only to *CC-on-Diskette*, but to other ISI products as well: a powerful new search capability called *KeyWords Plus*℠. In the first installment of this two-part essay, we'll discuss *KeyWords Plus* and how it expands the versatility and depth of a literature search. In the conclusion we'll look at some of the other enhancements that are improving the ease and convenience of using *CC-on-Diskette* to stay current with the latest literature.

*CC* readers and users of the previous editions of *CC-on-Diskette* are already familiar with the ability to search each issue by title word. However effective a title may be, it seldom will provide the depth of indexing possible from exploring the full article. *KeyWords Plus* goes beyond title-word indexing. Using a technology that is unique to ISI's database, *KeyWords Plus* supplies additional search terms extracted from the titles of articles cited by authors in their bibliographies and footnotes. To reiterate, the *KeyWords Plus* algorithm identifies recurring words or phrases appearing in a paper's list of cited references. *CC-on-Diskette*, in the "Basic" search mode, pools the title words with those in the *KeyWords Plus* terms. When you search *CC-on-Diskette* with *KeyWords Plus*, you will usually retrieve a greater number of relevant papers—regardless of whether your search term appears in their titles.

But there is even more. The enhancement words and phrases generated by *KeyWords Plus*, when strung together, provide a brief condensation of the major and minor themes discussed. The *KeyWords Plus* terms will be part of the full-record display for each article located by a search. The terms will also be included in the data files when you

export records using the "long-record" option. The combination of terms not only augments the perspective provided by the article's title, but also may alert you to aspects or angles in the material that might have escaped your attention otherwise. In this way *KeyWords Plus* can serve to provide new ideas and avenues, helping to focus your search.

Figure 1 provides a sample illustration of how *KeyWords Plus* can enhance a search. The source title, "A commentary on the interpretation of *in vitro* biochemical measures of brown adipose tissue thermogenesis," along with the author keywords, provides a reasonably detailed picture of the article's subject. The *KeyWords Plus* terms, however, which have been generated independently of the title or author keywords, go into far more detail, describing the article's contents with greater depth and variety.

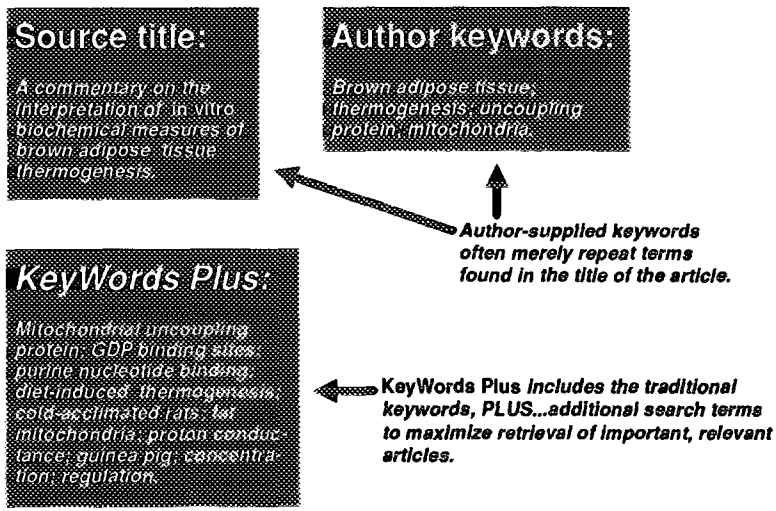
Like the "related records" feature in the CD-ROM versions of the *Science Citation Index®* (*SCI®*) and *Social Sciences Citation Index®*,[4,5] which uses cited references to highlight relationships between seemingly unrelated articles, *KeyWords Plus* is an inventive application of the unique power of citation indexing. Although simple in practice, the capacity to extract meaningful terms from cited references requires a combination of complex computer operations and high-speed data processing.

## Years in Development

*KeyWords Plus* is the result of many years of thought and experimentation by many people. The first informal experiments were conducted by me and Robert L. Hayne[6] when we both worked at Smith, Kline and French Labs (SK&F) in the 1950s. Subsequently, in the 1970s, K.L. Kwok, Queens College, City University of New York, Flushing, reported experiments on the idea of using titles of cited articles as indexing entries as well as in the automatic classification of the citing article.[7,8] These ideas, which were further developed in the 1980s, also have their roots in the early 1960s work of Michael M. Kessler, Massachusetts Institute of Technology, Cambridge, on bibliographic coupling.[9]

In view of its practical potential, this idea has also been the subject of intense study at ISI for many years. Thanks to key contributions from Irv Sher, director, Development and Quality Control—as well as George Vladutz, manager of basic research, Corporate Research Department; Robert Griffith, systems analyst/consultant, Corporate

**Figure 1: Sample illustration of *KeyWords Plus*™ terms.**



Information Management Division; and others—the idea has been removed from the realm of the theoretical and made to work in an actual production environment.

*KeyWords Plus*, it should be noted, represents only the latest of Sher's many contributions to ISI. Irv came to SK&F at about the time Hayne and I were doing first experiments in using the titles of cited articles to develop indexing terms. Irv and I quickly discovered a shared interest in various aspects of information science and began a collaboration that goes back more than 30 years. Having come to ISI in 1961, he was indispensable in the development of such key ISI products as the *SCI*, the *Permuterm® Subject Index*, *Research Alert℠* (formerly *ASCA®*), and many others. Irv's wide-ranging interests and accomplishments in many fields have made him an invaluable source of information on numerous past essays, including those on voice-recognition systems and shorthand.[10,11] Quality-control procedures at ISI, which Irv played a large role in developing, were discussed in a 1983 essay.[12]

The operation behind *KeyWords Plus*, to describe it in the briefest possible terms, begins with the standard initial processing that ISI performs on the publications covered in its database. In addition to the complete bibliographic information for each new source paper (authors, title, subtitle, journal, volume, issue, pages, etc.), condensed citations for all of the article's cited references are also keyed. When the accuracy of the input has been confirmed, ISI's voluminous back files are next searched for each of the cited references in order to recall that article's title, if processed earlier as a source item. The *KeyWords Plus* program then pools the inventory of available older titles, parses words or phrases, groups them by frequency, and, ultimately, identifies the most significant terms. The minimum threshold for inclusion is at least two occurrences of a term. The program also analyzes a variety of orthographic, grammatical, and syntactical attributes. After this processing the accepted terms are sequenced by their likely general utility. The list of new terms is then added to the database record of the appropriate source article.

As mentioned above the procedure can greatly expand the effectiveness of literature searching. In the course of developing and testing, we found that the combined use of *KeyWords Plus* and title words substantially increased the number of relevant articles retrieved. In the life sciences, for example, *KeyWords Plus* typically doubles the number of articles retrieved, although the improvement in article retrieval is not quite as marked in the other editions of *CC-on-Diskette*. The improvement in retrieval is also subject to considerable variation depending upon the type of material indexed. The enhanced searching power for a review article with a pithy title can be enormous. The enhancement for a brief letter may be small but significant.

## More Terms, More Detail

Examining detailed examples will help illustrate how *KeyWords Plus* can enhance a search. Consider the 1989 *Annals of Thoracic Surgery* article entitled "Long-term survival after postinfarction bypass operation: early versus late operation," by H.S. Floten and colleagues, the Heart Institute at St. Vincent Hospital and Medical Center and the Division of Cardiopulmonary Surgery, Oregon Health Sciences Center, Portland.[13] The authors address the controversy concerning the relative risk and appropriate timing of bypass surgery for patients who have had myocardial infarction. Some scientists believe that the operation should be performed as soon as possible, while others find it more appropriate to delay the surgery until the patient is stable. The authors compared long-term survival rates with various time delays. They concluded that the length of delay does not significantly affect the long-term survival rate.

Among the *KeyWords Plus* for this article, SURGICAL MANAGEMENT, CLINICAL COURSE, PROGNOSIS, and SURGERY are closely related to the article's general theme. The issue at hand is timing of SUR-

GERY, or SURGICAL MANAGEMENT. CLINICAL COURSE refers to the progression of a patient's condition over time. A PROGNOSIS is a prediction of the CLINICAL COURSE of a disease. Other *KeyWords Plus* terms generated for this article are ACUTE MYOCARDIAL INFARCTION, CORONARY-ARTERY BYPASS, EXTENSION, REVASCULARIZATION, and ANGINA.

A 1989 paper from the *Journal of Applied Ecology* demonstrates how *KeyWords Plus* terms act as a surrogate "abstract." These terms display additional information about the article that is not discernible from the title alone. In the case of "Environmental effects of air pollution in Britain," by S.J. Woodin, Nature Conservancy Council, Peterborough, UK,[14] the *KeyWords Plus* terms include NITROGEN DEPOSITION, SOIL ACIDIFICATION, FOREST DECLINE, ACID RAIN, GROWTH, VEGETATION, STREAMS, and SULFUR. These terms, as we'll note below, are insightful in highlighting the article's content.

First, the author gives a brief history of air pollution in Britain, which, in the nineteenth century, consisted of SULFUR dioxide and smoke in the air. The recurring terms "air pollution" and "Britain" have been excluded by *KeyWords Plus* since they already appear in the title. The first air pollution inspector found that the rain contained SULFURic acid and coined the term "ACID RAIN." Currently, reports Woodin, the main pollutants are SULFUR dioxide, hydrocarbons, NITROGEN oxide, and ammonia. The author then gives the main ecological effects of air pollution. These include freshwater acidification (STREAMS, lakes, etc.), due to ACID RAIN. Other effects concern FOREST DECLINE due in part to SOIL ACIDIFICATION and various consequences due to NITROGEN DEPOSITION on VEGETATION. The remainder of the article covers the effects of air pollutants on crop GROWTH, and the status of international and domestic pollution control.

*KeyWords Plus* can also elucidate key methodologies or techniques used by authors. Consider "Tissue-specific effects of maize *bronze* gene promoter mutations induced by *Ds1* insertion and excision," a 1989 paper from *Developmental Genetics* by Thomas D. Sullivan and colleagues, Laboratory of Genetics, University of Wisconsin, Madison.[15] The following *KeyWords Plus* terms were generated: POLYMERASE CHAIN-REACTION, TRANSPOSABLE ELEMENTS, MOLECULAR ANALYSIS, FLAVONOID GLUCOSYLTRANSFERASE, ENZYMATIC AMPLIFICATION, FINE-STRUCTURE, ZEA-MAYS, LOCUS, DNA, and others. Two of the terms, POLYMERASE CHAIN-REACTION and ENZYMATIC AMPLIFICATION, describe the all-important development that was applied in this research. The authors' goal, as they noted, was "to analyze the DNA sequences in the region of the *Ds1* insertion site in the three derivatives of *Bz-wm*. This analysis was greatly facilitated by the recently developed technique for the *in vitro* AMPLIFICATION of DNA called 'POLYMERASE CHAIN-REACTION.' "[15]

*KeyWords Plus*, as mentioned above, is entirely unique to ISI and its products. In addition to its availability on *CC-on-Diskette*, *KeyWords Plus* will also be included with *Focus On: Global Change*™. This current-awareness product, introduced this past spring, is also delivered on diskette and monitors research on the array of forces—physical, chemical, social, political, and so forth—that are causing global environmental change.[16] And we are currently investigating the possibilities of harnessing *KeyWords Plus* to more of our products.

While *KeyWords Plus* will be the most striking enhancement in the new version of *CC-on-Diskette*, other notable improvements have been made to version 1.3 for IBM and NEC computers and to version 2.1 for the Macintosh. In the next part of this essay, we will examine these features in detail.

* * * * *

*My thanks to Elise Arle, Kathy Barna, Christopher King, Bill Michael, Gary Schwartz, Irv Sher, and Anita Wagner for their help in the preparation of this essay.*

## REFERENCES


1. **Garfield E.** Introducing *Current Contents on Diskette*: electronic browsing comes of age. *Current Contents* (39):3-8, 26 September 1988.
2. ----------. *Current Contents on Diskette* for the IBM PC: on a screen near you, electronic browsing, searching, and retrieval and expanded coverage. *Current Contents* (49):3-9, 5 December 1988.
3. ----------. *Current Contents on Diskette*: new software for the Macintosh and Japanese NEC computers; journal coverage extended to *CC/Physical, Chemical & Earth Sciences* and *CC/Agriculture, Biology & Environmental Sciences*—2,800 journals and still growing. *Current Contents* (42):3-11, 16 October 1989.
4. ----------. Announcing the *SCI Compact Disc Edition*: CD-ROM gigabyte storage technology, novel software, and bibliographic coupling make desktop research and discovery a reality. *Current Contents* (22):3-13, 30 May 1988.
5. ----------. Expanding the searching power of CD-ROM: ISI's new *Social Sciences Citation Index Compact Disc Edition* is compatible with the *Science Citation Index* on compact disc; new software streamlines searching. *Current Contents* (37):3-10, 11 September 1989.
6. ----------. To remember my brother, Robert L. Hayne. *Essays of an information scientist.* Philadelphia: ISI Press, 1980. Vol. 3. p. 213-4.
7. **Kwok K L.** The use of title and cited titles as document representation for automatic classification. *Inform. Process. Manage.* 11:201-6, 1975.
8. ----------. On the use of bibliographically related titles for the enhancement of document representations. *Inform. Process. Manage.* 24:123-31, 1988.
9. **Kessler M M.** Bibliographic coupling between scientific papers. *Amer. Doc.* 14:10-25, 1963.
10. **Garfield E.** From Vocoder to Vocalock—speech recognition machines still have a long way to go. *Op. cit.*, 1981. Vol. 4. p. 579-85.
11. ----------. Is shorthand the route to success in science or anything else? Parts 1 & 2. *Ibid.*, 1986. Vol. 8. p. 1-20.
12. ----------. Quality control at ISI: a piece of your mind can help us in our quest for error-free bibliographic information. *Ibid.*, 1984. Vol. 6. p. 144-51. (Reprinted from: *Current Contents* (19):5-12, 9 May 1983.)
13. **Floten H S, Ahmad A, Swanson J S, Wood J A, Chapman R D, Fessler C L & Starr A.** Long-term survival after postinfarction bypass operation: early versus late operation. *Ann. Thorac. Surg.* 48:757-63, 1989.
14. **Woodin S J.** Environmental effects of air pollution in Britain. *J. Appl. Ecol.* 26:749-61, 1989.
15. **Sullivan T D, Schiefelbein J W & Nelson O E.** Tissue-specific effects of maize *bronze* gene promoter mutations induced by *Ds1* insertion and excision. *Develop. Genetics* 10:412-24, 1989.
16. **Garfield E.** *Focus On: Global Change*—A new current-awareness service tracking the health of planet Earth. *Current Contents* (14):3-9, 2 April 1990.