September 9, 1959

Dr. Joshua Lederberg
Genetics Department
Stanford University
Stanford, Cal.

Originally prepared August 15,1959

Dear Dr. Lederberg:

Since my last letter much progress has been made. I have met and spoken with Gordon Allen, George LaFevre, Miss Shapard and Miss Tolkan, the last three of NSF. On my next visit to Washington I hope to see Bernard Cohen of the NRC. Sid Bernhard of NIH tole me that Bill Consolazio of NSF might also be interested.

Here is a rundown on my discussions with Gordon Allen. To really demonstrate the value of a citation index we should, somehow, come up with as complete a citation index as possible to a selected list of journals and/or articles. Compiling a citation index to a selected list of articles would increase problem of scanning the bibliographies and references in articles from which citations would be taken. For example, if a paper in "Nature" is included in our sample we would have to carefully examine citations to Nature. However, this would offset the cost of handling a larger quantity of citations in non-genetic articles.

Another sampling approach, a sort of compromise, would be to scan for citations to articles by a few particular authors. The relative speed and costs should be tested. I have done some samplings during the past week (see below).

We agreed that some mechanical method must be developed for copying the citations. This has been acted on already. About two years ago I discussed this problem with the National Library of Medicine. They have a special microfilm camera for copying references. (For about $1,000 a camera can be built precisely for our work) I am enclosing a few sample citations copied on the NIM camera. The citation is just as it appeared in the bibliography of the article. A "mask" is used so that the citation for the citing article is repeated every time. Since time was short we wrote in the reference by hand. It could have been typed. By using a camera of this type we can copy citations at a maximum cost of 2¢ each. More than likely, it will be less than 1¢. This means we can copy 1,000,000 citations at a cost of $10,000 to $20,000 and includes the cost of: camera, labor, film, paper, processing. It does not include supervision, overhead, or intellectual work. (See below)

We also discussed the question of specifying the "kind" of citation involved. Here is where we get into "intellectual" problems. I believe that citation index research will pay off handsomely in the future in that this research will characterize all the different ways in which people "cite" the earlier literature. We will then be able to provide editors with a guide to standardized citation practices. Further, they might be influenced to adopt a natation or terminology that would indicate to the reader and the bibliographer the "nature" of the citation. In this project we would try to characterize, for a selected list of articles, each citation as to whether it was: 1. Review article (Rev.); 2. Communication (Comm.); 3. Editorial (Edit.); 4. Errata (Err.); 5. Translation (Tr.); 6. Abstract (Ab.); 7. Book (Bk.); 8. Discussion (Disc.); 9. Summary (Summ.); 10. Bibliography (Bibl.); 11. Book Review.

I have purposely left out: refutation, confirmation, etc. I have also left out any mention as to whether pertinent portion of citing paper is experimental, theoretical, introductory or whether it is a use of method cited or use of "material" cited. These are points to be investigated later on.

The next problem for discussion was whether or not to include the page number on which a citation is made. This would speed up locating the pertinent statements. In those journals which use a numbering system we would include the reference number. (See enclosed samples). In those articles with a bibliography arranged alphabetically by author we would not make a special attempt to locate exact page. However, if we attempt to provide one of the codes mentioned above it would not be difficult to add the page number in certain cases. Obviously it is not difficult to decide that something is a short communication, translation, or summary. To state if it is a confirmation, refutation, etc., is another thing.

Prepared September 3, 1959

To obtain some basic figures I did several sample test runs with various journals. Two independent analyses show that the average number of references per article is approximately 15. I have tabulated below the two separate tests I ran. Note that there is considerable variation from journal to journal.

Test #1

No. ARTICLES	JOURNAL	NO. REFERENCES	AVERAGE
20	Genetics	372	19
45	Sch. Z. f. Allg. Path. Bact.	721	15
104	" " " " "	1406	13
80	J. Antibiot.	424	5
128	J. Bact.	1670	13
94	JBC	2015	22
72	J. ACS	1207	17
38	J. AMA	285	8
50	B. M. J.	199	4
82	Bull. Exp. B. M.	683	9
101	J. Physiol.	1182	12
814		10,165	13

Test #2

108	Naturwiss.	589	5
58	Science	488	8
71	Exper.	645	9
38	J. Endoc.	743	20
62	J. Bact.	731	12
27	J. Gen. Physiol.	550	20
35	J. Exp. Biol.	663	19
100	J.B.C.	2127	21
161	J. Org. Chem.	2010	13
49	Arch. Intl. Ph.	727	15
17	J. Parasit.	98	5
66	Arch. Biochem.	1174	18
792		10,545	14

Incidentally, in checking six months of the Journal of Bacteriology I found 13 references to genetics journals. I did not continue to compile such figures for the other journals as I was primarily concerned with the amount of time required to scan journals. The time required to record the references found would be low compared to the total time required for scanning.

Scanning 1500 articles took about 15 hours in five sessions of three hours duration. I could sometimes scan as much as 200 articles per hour. It was never lower than 100 per hour. Depending upon motivation skilled clerks could scan at an average rate of about 100 articles per hour.

It was not difficult to scan for several types or combinations of information. I had no difficulty looking for references containing the abbreviation "Gen.", but found that this included the word "general" in titles like the "J. Gen. Microbiol." as well as "genetics" in other titles. Scanning for one or more individual authors was easy too. As you know I have sent you several references to your articles. In fact, scanning was made easier and more enjoyable as the criteria for searching became more complex. The degree of complexity to be allowed would depend upon the people employed.

The first test sampling covering 814 articles (10,165 references) was done to ascertain the feasibility of compiling a Citation Index to Genetics (or any other specific field). Clearly, the cost of scanning a very large volume of literature would be reasonable. To scan 2,000,000 references in over 100,000 articles (coverage of Current Contents) would involve approximately 1500 man hours. This scanning could include searching for a specified list of genetics journals and/or other journals. It could also include specific authors and/or specific articles.

In the second test sampling of 792 articles with 10,545 references I tested the ability to search for references to a list of general science journals. I had no difficulty keeping track of references to Nature, Science, Naturwisenschaften, Proc. Royal Soc., Proc. National Acad., Comptes Rendus, Doklady, Experientia, etc.

In the average article one of these general science journals is cited and 50% of the articles contain none. Those articles that do contain references to general science journals contain two such references. There is considerable variation from journal to journal. For example, the Journal of Organic Chemistry is much different than Arch. Biochem. or JBC. It does not contain many references to such general science journals. However, when an article does contain such references the average is three. The average for all articles in J. Org. Chem. is 1/2. The Archives and JBC are typical of the average. 40% of the articles have no references to Nature, Science, etc. 60% have 2.

As others have found (Brown "Scientific Serials") the citation practices for each journal are colored by many factors. British journals cite the Proc. Royal Soc. more often than American journals. However, I don't go along with the popular idea that this necessarily reflects nationalistic provincialism. The authors in JBC come from all over the world. However, their reading might appear to be concentrated in the JBC. In fact, as the journals become more international in character you might expect their citations to be more international, but this does not necessarily follow.

As regards specific countries I know that the Russian journals contain a high percentage of references to Russian journals but they also contain many references to non-Russian journals. However, the Doklady contains few references to the western general science journals but do contain many references to the Doklady and other specialized western journals.

The conclusion to be drawn from test #2 is that we can compile a complete citation index to all the general science journals, making the sample not only interdisciplinary but permanently useful when it is finished. The work will not be wasted. This ties in beautifully with another idea I had and discussed with Gordon Allen in which we would abandon the concept of a unified Citation Index to all science journals and prepare, instead, individual Citation Indexes for each journal. At the end of each year we could send to each journal editor a citation index for his own journal. Periodically the individual Citation Indexes could be accumulated. This would be similar to the practices followed for legal citation indexes. One is prepared for each state and they are cumulated quarterly, yearly, every five years and 15 years.

One of the things that intrigues me most about the idea of a citation index to Nature, Science, etc., is that it proves to be a statistically significant way of permeating the entire literature of science, since, on the average, every article has at least one reference to a general science journal. Since any searcher could then trace the bibliographies in the articles listed in the Citation Index, the general science CI would give him access to a total number of references exactly equal to the total references that appear in the entire literature. It will be extremely interesting to do comparative literature searches based on using the CI to gen. sci. journals as a starting point and the bibliographies in the articles so located as a follow up.

The citation index for any pertinent additional references so located could then be checked providing a continuous chain reaction.

In conclusion, pending comments from you, G. Allen and others, I feel that a revised proposal to NSF should be based on the following three part program of Citation Index research.

1. Mechanically (photographically) pick up all references found in a specified list of genetics journals and articles, the latter based on some well known genetics bibliography. From these eliminate undesired references.

2. Scan all Current Contents journals for references to all articles appearing in a specified list of genetics journals and a specified list of articles or authors.

3. Scan a large list of journals from all representative scientific disciplines for references to general science journals including Nature, Science, Proc. Natl. Acad., etc.

From the above we would obtain

1. Complete and permanently useful citation index to a specified list of genetics journals.

2. Complete and permanently useful citation index to a specified list of genetics articles published in non-genetic journals.

3. Complete citation index to all articles that appeared in general science journals including the genetics articles.

The scanning should cover at least the last five years of the literature, preferably more. I would prefer to cover fewer journals over longer period of time. In part 1 if we assume that we limit work to 30,000 articles (450,000 references) the maximum cost is $9,000 using one full time camera operator. This could cover large part of if not the entire genetics literature.

Part 2 would turn up an unknown quantity of references. However, if we assume that we will scan 150,000 articles per year at the rate of 75 articles per hour this will require 2,000 hours per year. To cover a five year period would require 5 man years or about 5 X $3,500 = $17,500. As work progresses we can determine actual operational speeds and whether we have to cut down on number of journals covered or whether we can increase the number of years or journals covered. I think it is safe to assume that the number of references found will be less than 1% of those scanned or less than 10,000 references to genetics articles.

Part 3 would produce about one million references to all general science journals. The time to sort and collate these references would require about one man year of clerical time. If we estimate that there are 10,000 articles per year published in the general science journals, that 90% of the citations are to articles published in the last 10 years, then each article would be cited an average of 10 times.

For a journal like Science, its own individual citation index would include about 100,000 references to 10,000 different articles. The CI itself would be a book of about 300 pages. This implies the use of full citations giving authors, journal abbreviations, volumes, pages and years. I would not propose to use any numerical code for journals.

Every year a supplement of about 50 pages could be issued. Every five years a new cumulation could be printed.

The figures given above would make it possible for us to conduct this program on the two year budget of $59,000 originally requested. It would also allow additional funds for "testing" the value of the Cita tion Index. I believe there are a number of "comparisons" that could be made within our research budget, but I would prefer to determine the "value" of a citation index on the basis of users comments. To obtain this information copies of Citation Indexes should be placed in the hands of geneticists and various libraries.

At the completion of the research program for the first year I would suggest that the Citation Index to several individual genetics journals be published as individual journal articles or supplements. If deemed more useful we could publish a single combined "Citation Index to Genetics".

There are so many different ways in which the usefulness of the CI, once compiled, could be tested that it would be too time consuming to consider this in detail at this time. I have already spent more time on the "preparation" for this proposal than I can really afford. I am hopeful that this letter, your comments and those of others, will enable me to proceed to a relatively simple proposal for an NSF or NIH grant.

Sincerely,

Eugene Garfield

CC: Gordon Allen
Katherine Wilson
Connie Tolkan
George LeFevre

Encls.