DOCUMENT RESUME

ED 024 571 SE 004 900

By- Herschman, Arthur

Information Retrieval in Physics.

American Inst. of Physics, New York, N.Y.
Spons Agency- National Science Foundation, Washington, D.C.
Pub Date Apr 67
Grant- NSF-GN-549
Note- 22p.
EDRS Price MF-$0.25 HC-$1.20
Descriptors- *College Science, *Information Dissemination, *Information Retrieval, *Information Science, *Physics, Scientific Research
Identifiers- American Physical Society, National Science Foundation

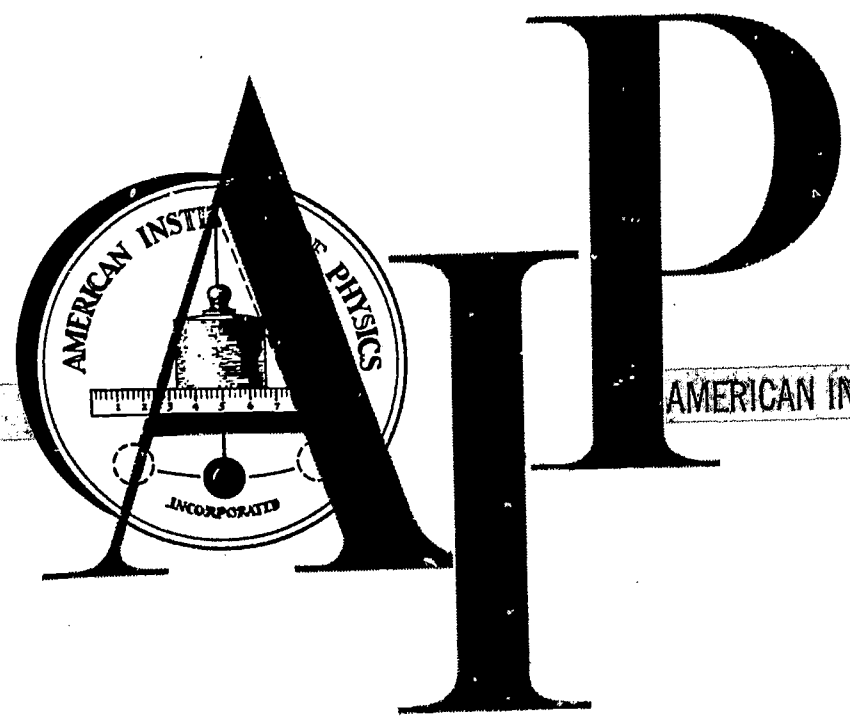
Discussed in this paper are the information problems in physics and the current program of the American Institute of Physics (AIP) being conducted in an attempt to develop an information retrieval system. The seriousness of the need is described by means of graphs indicating the exponential rise in the number of physics publications in the last few years. The AIP program in the past has been one of studying documents, but now need dictates an expanded program going beyond document analysis and moving toward information retrieval. The present undertaking has as its immediate goal a pilot model of a physics information system, the heart of which is an information store. The AIP system development plan has five stages: (1) classification scheme, (2) computer retrieval capability, (3) computer composition and input for store, (4) user need studies, and (5) information network studies. These studies are expected to yield the means by which journals and abstract journal indexes can be efficiently produced and a system of information stored. This paper explains the classification scheme that is under study and illustrates its use with two sample documents. Various ways in which the output of the system might be used are also discussed. (DH)

Received in RSP 4/17/68
No. of copies 7
Grant (Contract) No. GN-549

IARD 67-1
(October 1967)

INFORMATION RETRIEVAL IN PHYSICS

Arthur Herschman



Information Analysis and Retrieval Divisi

AMERICAN INSTITUTE OF PHYSICS · 335 East 45 Street · New York, N.Y.

This program supported by the
National Science Foundation
under Grant No. NSF-GN-549

INFORMATION RETRIEVAL IN PHYSICS*

Arthur Herschman

American Institute of Physics

Two of the previous speakers have given you a picture of the scope of the national and international problem in scientific information. What I would like to do in this talk is to cover the information problems in physics and the program of the American Institute of Physics.

Some years ago, a very emminent physicist made the following remark:

> In science, by a fiction as remarkable as any to be found in law, what has once been published, even though it be in the Russian language, is spoken of as known, and it is too often forgotten that the rediscovery in the library may be a more difficult and uncertain process than the first discovery in the laboratory.

This was said by Lord Rayleigh in 1884. Since 1884 a great deal has happened and the increase in scientific information has been astounding.

Figure 1 shows the increase in the number of United States scientific journals and the number of world's principal physics journals since 1900. These are the journals that essentially contribute approximately 90% of the physics literature. The first curve is almost an exponential function and the second is really exponential with a doubling time of 17 years. Figure 2 shows the growth of Physics Abstracts and of The Physical Review for the same time period; again we see essentially exponential functions, if all the dips etc. are averaged over; these have a doubling time of about 10.3 years. Figure 3 shows the increase in the number of pages published by The Physical

* Invited paper DB4 presented at the American Physical Society Meeting in Washington, 25 April 1967.

Review, by the American Institute of Physics, and of the number of abstracts covered by Physics Abstracts since just after the war. Again the increase is essentially exponential. There is a great deal of structure in these curves and one can read a lot of interesting history into some of the dips and rises; however, this is not our interest at this time. The problem is that the literature is increasing and increasing very rapidly; something we all know. Figure 4 shows the same information on a linear plot. In addition to this, there is also an increase in the number of people involved in physics, and Figure 5 shows the increase in the membership of the American Physical Society, the fellows of the APS, and the AIP members. The top curve is a somewhat poor one, since some people were counted more than once because of multiple membership.

The reaction of the physicist to this increase is something we can also document. The percentage of those members of the American Institute of Physics who subscribe to any one of our journals has been decreasing, although the absolute number of subscriptions has increased. If one were to look at the percentage of the members of the American Physical Society who subscribe to The Physical Review, there is also a significant decrease, which has only now begun to level off and increase slightly - something we attribute to the fact that people can subscribe to sections of the Physical Review, that is, they don't have to scan the entire literature to find what they want. The increase in the amount of material published, the number of abstracts given at APS meetings, and other measures of a similar kind are all indications of the growing problem in information in physics.

Before proceeding further, let me define this nebulous word information. Samuel Johnson made a remark a long time ago: "knowledge is of two kinds, we know the subject ourselves or we know where we can find the information upon it." It is in the sense of "where we can find" that we use the word information in our discussions. The retrieval of information from our point of view is not so much the retrieval of fact but the retrieval of the reference to the place where that fact can be found! Sometimes it is just finding the knowledgeable colleague who might know it, sometimes it is finding the compilation or data center where the data exists, and sometimes the appropriate book to gain background on the subject. But for most of us and at most times, it is the problem of finding which article discusses the point at interest. For most purposes the retrieval of information is reduced to the problem of a retrieval of a document, and in most instances it would suffice to simply retrieve an adequate reference to the document so that the library or one's own shelf will supply the document. The information within the document must be retrieved by the ultimate user.

The AIP's program, over the years, has been one of studying documents: how they are to be represented, what is the best way to classify them, what is the best way to be able to indicate what their contents are, how do people use them, and so on. We expect to continue our work in documentation research and are in the process of expanding it into the larger problems of the physics information system.

The current program at the AIP goes beyond merely the analysis of document. It has a part which is involved in an attempt to understand other forms of communi-

cation in physics, beyond document analysis. This involves studies of methods of informal communication, person-to-person contact; of the best way to structure meetings; of the problems of pre-published and unpublished information and what services one can give to these areas, and in what ways one can improve the utility of such forms of communication. It also worries about interconnection between any system which the American Institute of Physics is setting up and other systems which border on it, both from the point of view of discipline, e.g., chemistry and engineering, which relate to physics, and from the point of view of various missions, e.g., government programs, which relate to physics; the inter-relationship between these to prevent duplication; and the inter-relationship between the AIP's Information Program and those abroad, particularly the program of the Institution of Electrical Engineers, the publishers of Physics Abstracts. This whole picture is being studied and will continue to be studied until enough hard facts can be obtained so we can make sensible proposal of that manner in which the AIP can contribute to the improving of the national physics information handling problem.

However, the prime focus of our present activities relates to the constructing of a pilot model of a physics information system at the Institute. Since such a model revolves around an information store, one must establish a computer file of document representations which has enough material about the various documents published in physics so that these documents can be retrieved for research and other purposes. Initially, we imagine such a store would contain only AIP generated documents which is about $\frac{1}{4}$ of the world's total

(or 1/3 if one includes the Russian translations). Ultimately, any real store would contain the representation for all documents published in physics.

Essential to the whole picture is a computer store, not simply a file of documents on a shelf in the library. It is for this and several other purposes that the Institute has initiated a study of the computer production of journals. As many of you are well aware, there has been a terrible delay in our journal publication. Standard hot-type printing methods have fallen down under the tremendous tide of new publications in physics. It has become necessary to turn to the modern technology of automation in printing to resolve this problem. As a by-product of this technology we expect to obtain computer representation for the documents.

In the program which the Institute has initiated, manuscripts will be keyboarded into a computer by means of magnetic tape or paper tape to magnetic tape input. The computer would then have in addition to the terms keyboarded, the various specific type styles that these terms go into, etc. It would then operate a photo-composition device, what might be called an optical transducer, a device for producing photographic images of these terms in the form of a printed page. The photographic image would then be transformed into an off-set plate and then printed by conventional printing means through off-set processes. This procedure will allow us to speed up the operations of journals since the largest time-lag is in the composition phase and we would be able to do composition by means of a modified typewriter.

Also, at some point along the line, we would have the whole document on

magnetic tape and therefore it can serve as an input for our system. It is the nature of this input, the kind of store that would hold it, and the nature of the output such a store would have, which is the main purpose ofthat part of the Institute's program which I will discuss here.

Figure 6 shows the overall plan for the development of the AIP system; it indicates the five essential points of our current activities which would contribute to the overall program. The first box is the development of a classification scheme for physics. This classification scheme is a means of dividing up physics into enough individual separate subject areas so that each one can be handled essentially as a separate entity. It is a faceted classification which I will discuss in more detail later. The second stage is the development of computer retrieval capability. The third phase is the computer composition phase which generates the input for the store, which I described above.

The fourth stage is the user needs studies which will determine what kind of material should go into the store. For example, is it only journal literature? Should we include patents and similar material? What kind of output is most desirable? What format is the most desirable? What kind of services should we supply to users? In this connection, I would point out that I believe one cannot really answer many of these questions until you have a workable pilot that people can see, use and find out how it would respond to their needs. The last box, Information Network Studies, relate again to the kind of user needs; in this case, the needs of the largest

class of users that is the national and the international needs in physics. This also relates to needs by other disciplines and the inter-relationships between our system and those of other disciplines: whether we would get input from other organizations, supply them with direct input, what kind of output distribution problems would be involved, should the output be duplicated among several services or should each service divide up the pie in a specific way, etc.

We expect as output from these initial studies a means of producing journal and abstract journal indexes, and also the development of an information store, which would contain what we call the unit records of the documents. The special services we would expect to supply users from the AIP's program are, for example, searches of the store for answers to specific questions, something which we call group SDI (Selective Dissemination of Information) which means that we will package the information in the store, by limited subject or mission interests, for special groups; special laboratories doing work in certain areas, a group of users who are geographically dispersed but have the same interests, like all molecular spectroscopists, and it might be direct input to other scientific services or government missions. In addition we might produce a current awareness service in the form of a titles journal. Finally, in the nebulous future, we have the physics information system which has as input, the various forms of physics information -- from journal to store, together with all the necessary interconnections and "switching" procedures from one source to another. It will also include the interfacing between our own system and all of the other systems that would exist on a national and international basis and therefore would be the final product of all the operations that we are engaged in and are proposing. Such a system

would not be better than a pilot program for several years and would not be operational for quite a bit longer than that.

Now to return to the classification scheme; the classification system we are working on is a faceted classification. Figure 7 shows the kind of facets we are talking about. Rather than elaborate on each of these terms individually, let me indicate the way in which they would be used. We would imagine that the document under consideration has several units of information. These units of information can be considered from several different viewpoints. It might treat of a specific object or process, investigated by use of a certain discipline or by imposing a definite stimulus, and the result may be the production of some phenomenon or the elucidation of some property. One may ask "What was the object a man studied," or "What practices did he study," or "What properties did he find," or "What methodology did he use," and so on. These general characteristics which become the main headings for indexing terms are called facets of the index. One looks for information under several facets, and therefore would classify the information by using items from each of several faceted lists -- thereby classifying by more than the most obvious terms.

In our system we hope that the author would classify his paper prior to submission. He would be given a handbook of terms and rules, similar in size to the AIP style manual and would be asked to pick a term or terms from the first list, the second list, and so on. The hope is that the reviewers would check this classification when they check the paper, and that the editor would approve it when he approves the paper. The classification,

which might be several strings of words, would be printed with the abstract so that when the title page containing the abstract goes to a secondary service such as Physics Abstracts, the intellectual effort for producing its indexes has already been done. It is hoped the index can be produced by automatic means once rules are given on how to convert the faceted lists to index headings.

Figure 8 shows two papers which have been set up for inclusion in the store The aspects of the documents which would be included in the store is what we call the unit record. It has two parts, the bibliographic record and the indexing record. The first part is the coded entry for journal title, volume, and page; followed by the title of the article, the authors and the institution. The second part consists of three sub-parts.

The first is the faceted classification of the indexing record according to, in this case; object, processes, properties, phenomena, and disciplines. Other facets might be considered more useful and we are busy now working on the problem of determining what facets are the most useful in what parts of physics and constructing appropriate term lists. Lists have been constructed on a zero order approximation, if you like, in chemical physics, nuclear physics, solid state physics and in plasma physics. We are presently engaged in revising all of these lists and in developing lists in fluids, atoms and molecules, and lasers and masers. All the present ones are of different degrees of detail and use different numbers of facets. We are also studying the current headings being used by Physics Abstracts to determine what facets are best for the optimum choice of terms usable for the whole of physics.

In addition to the faceted classification is a set of keywords. These

are the free terms chosen from the text and also from series of lists which might be appended in the computer's memory, to the various possible classification headings. I will elaborate on this in a minute. The final item in this example is the citations of these papers to other papers which would also be included as additional terms for searching.

Let us go back a little and ask ourselves how we would use the output from the computer production system. Again this is partly conjectural because we have no output and we are only now studying the problem, but the computer production system at one point of its operation has a tape of the total journal article. I imagine that the information store might make an edited tape of this in the following manner: first include bibliographic material, the journal, volume, page, title, authors, and institutions. Next pick up the classification headings, which the editor has approved for this paper, and which appear right near the abstract. Then the computer would look in its memory for an appropriate set of keywords for each of these headings. These words might have been developed previously by a panel of experts in each of the subject areas involved. So that the computer could scan the rest of the paper for those keywords, which are on the list appropriate to the classification and also in the text of the paper. Various problems arise here about the appropriate linguistic forms and so on but programs being run by other organizations will help us in analyzing this problem when the time comes. The last item on the edited tape would be all the citations. The whole procedure is automatic and readily produces the record of the paper which would be entered in the store.

One can search the store on many items: on authors, on various classification groups, by means of citations (using methods developed for example at MIT, Project TIP). The importance of the classification scheme for the information store is that it permits the store to be broken down into a number of relative small subject areas, so that a search can be made within a limited region for most purposes, and thereby greatly expedite the efficiency of the search. For example, a person might ask for everything on light nuclei. After seeing the number of documents to be retrieved, he may want to limit his search by specifying certain nuclei, or only mirror nuclei, or only papers involved with scattering from light nuclei, and so on. He can further limit the search, if he wants more specific information than that of the finest subject classification by saying that he wants papers that are only published in certain journals, or in such and such a period, or by certain authors, or by a particular institution, or finally, if this is not the manner in which he wants to limit his material, after he has got a subject matter breakdown, he can say "well probably this paper mentions the following terms, or they might have cited so and so, and not so and so", etc. All these things become a possible means of conducting the search. Our reason for estimating that the unit record would have this form is that we have determined that these are the various kinds of handles that people put on information. We of course are receptive to further developments of other types of handles which might be put in and which would simplify retrieval.

Now as I mentioned earlier the principal product that we have in mind at this time for such a store, would be a current awareness titles periodical.

This would contain all the titles of material put into the store since the last issue. (Presumably the journal would be a semi-monthly). If one makes some reasonable estimate as to size of this, one could estimate about 3500 items per half month in a format of 150 to 200 pages, a rather slim issue, about the size of a thick copy of Physical Review Letters. This journal would probably contain the titles arranged by various convenient subject matter groupings; it might also contain an author index to these titles, or even further indexes to the same titles by special interest classifications. The purpose of this journal would be to allow people to know what has been done, in the most recent period, in areas of interest to them.

Several styles of current awareness journals are being published. These include:

Current Papers in Physics -- published by the Institution of Electrical Engineers, the same organization that publishes Physics Abstracts, in a newspaper format; it comes out twice a month and is eight pages in size; the classification used is that of Physics Abstracts.

Current Contents -- a commercial publication from the Institute of Scientific Information in Philadelphia, consists of the contents pages of various journals that have come out within the period between issues.

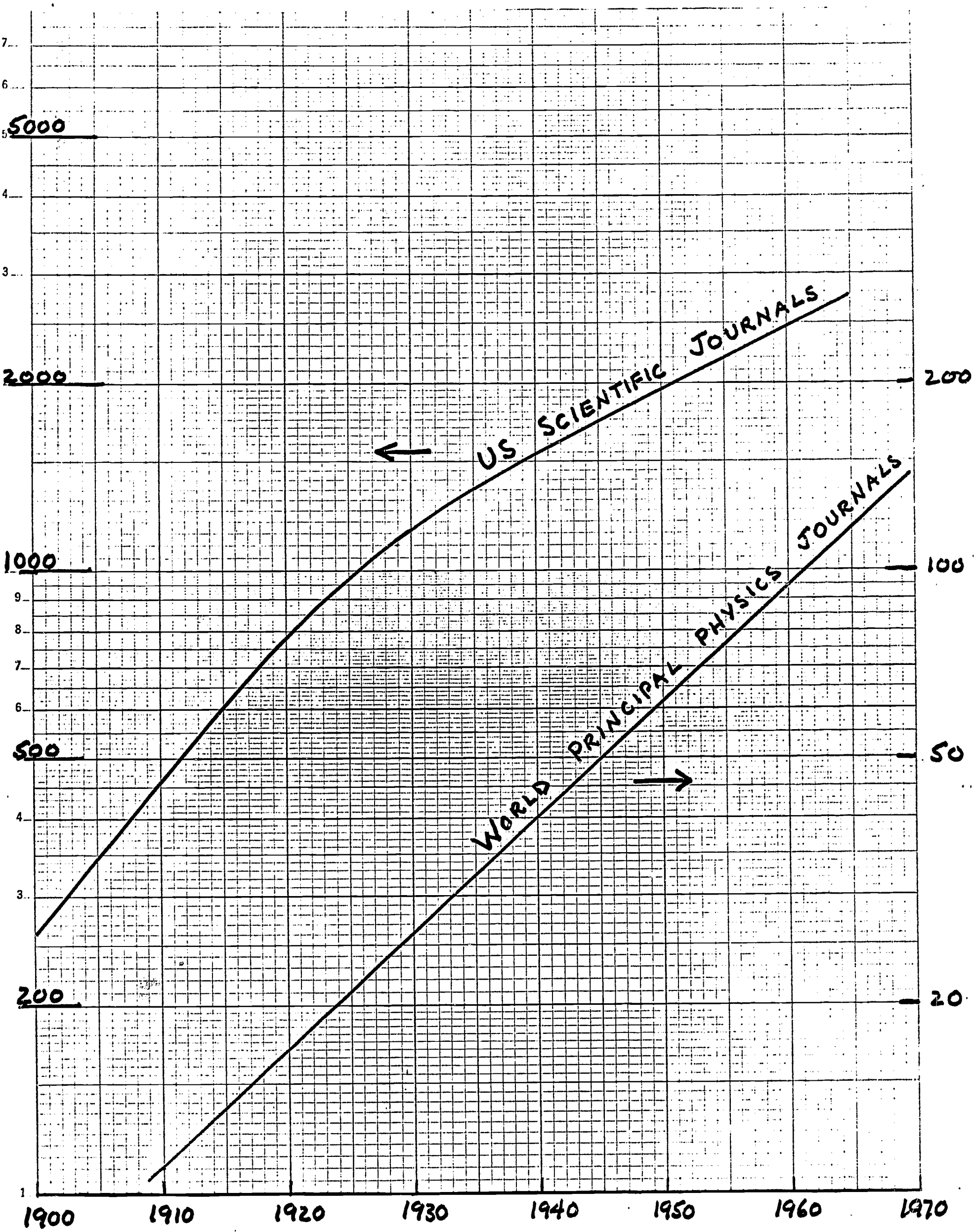
Chemical Titles -- published by Chemical Abstracts Service, which is in the form of a so-called KWIC index, that is the titles are permuted on all of the possible key terms; one looks down the list for an appropriate key term, sees what the rest of the title is to determine whether it is acceptable. This isn't too bad if the titles are well written but not all titles are, so

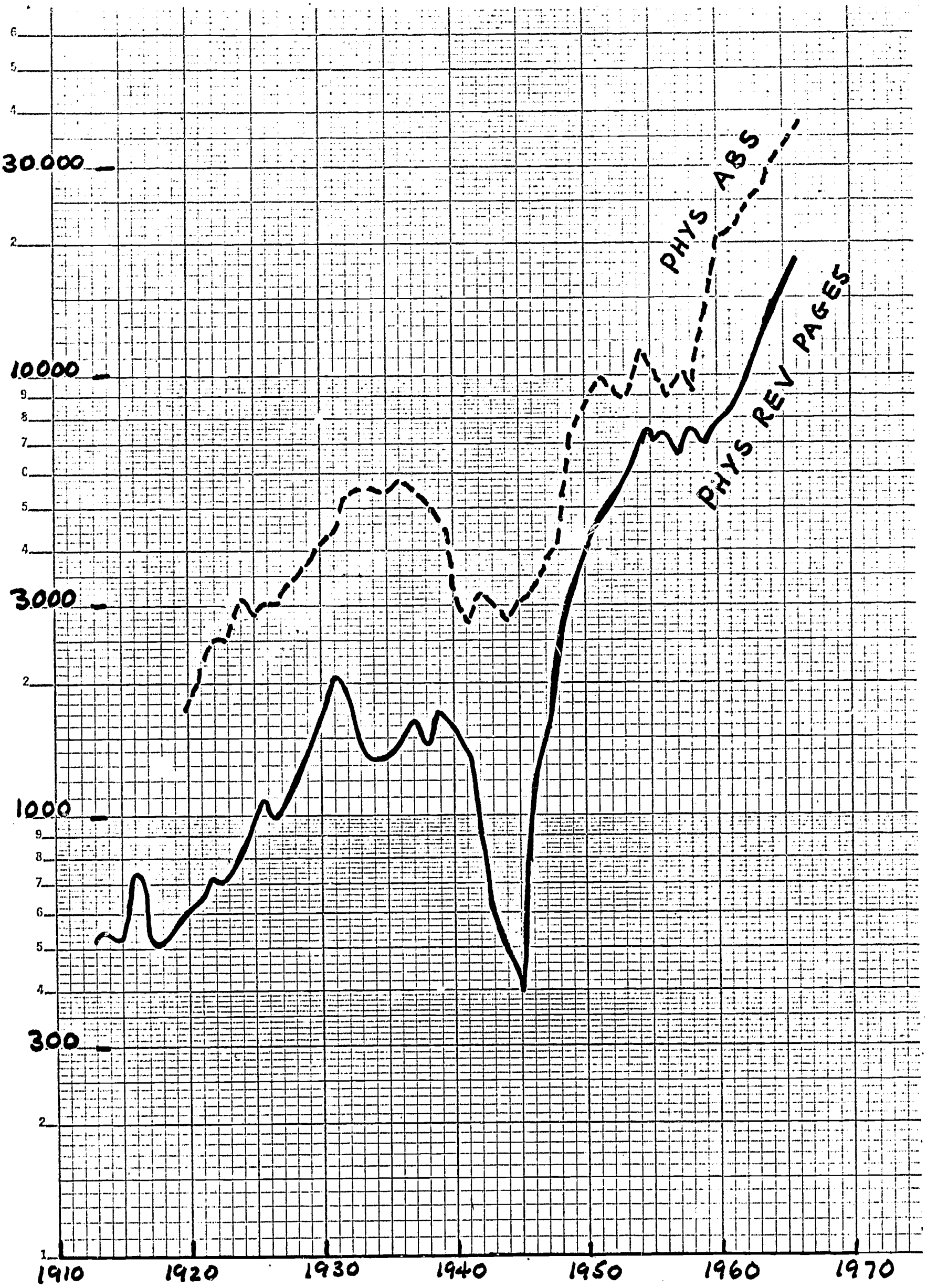
that sometimes the results are misleading.

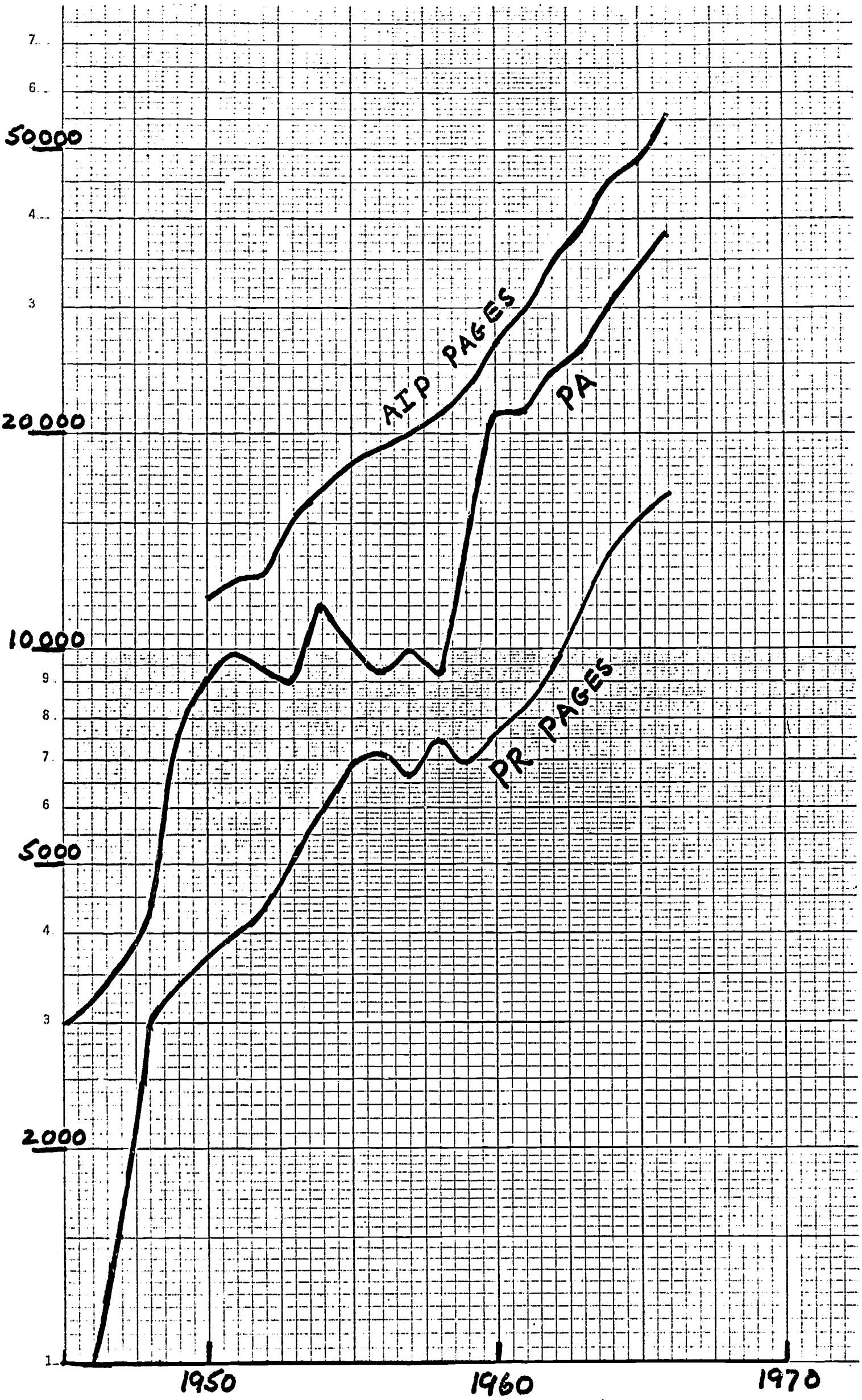
Index Medicus -- published by the National Library of Medicine, from a bio-medical information store called MEDLARS. The journal is issued monthly containing a classified titles list, very similar to what I have described as our aims in physics. In fact the medical library has been more of a model for us than any of the other existing services.

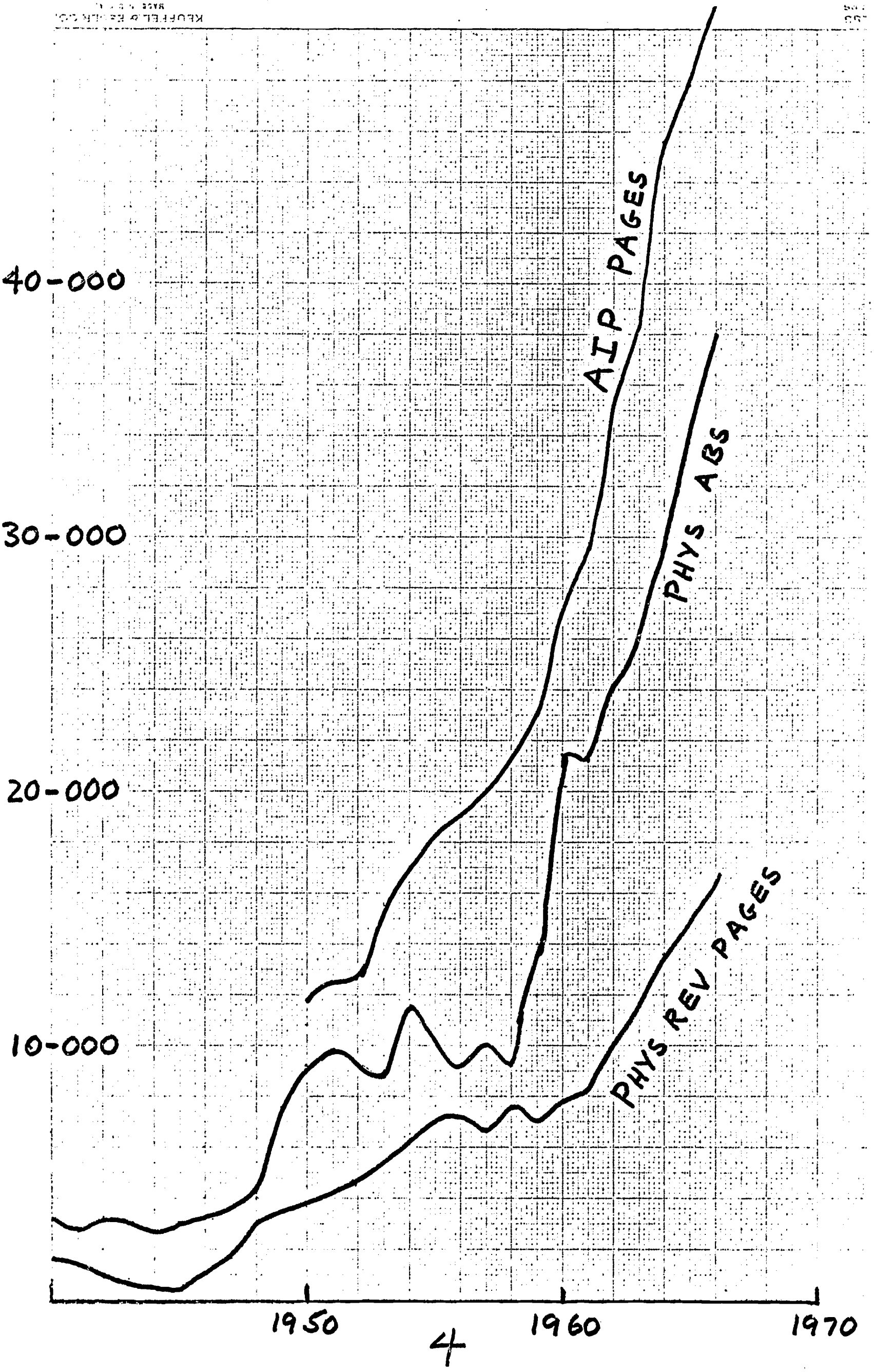
Now what form we would choose is not clear at this point. It may be much more sensible for us to simply attempt to improve Physics Abstracts by agreement with the Institution of Electrical Engineers whereby we would share input and share output distribution of this journal, and simply have one journal produced. The production and distribution aspects of this journal would also have to be extensively studied. It might very well be a modified version of CPP, produced at the IEE from input supplied by the AIP as well as the IEE, and distributed in this country by AIP. One of the things we could add in addition to titles is the full classification terms for each paper, and this would serve as a micro-abstract of the paper and tell with more precision what this particular paper is about.

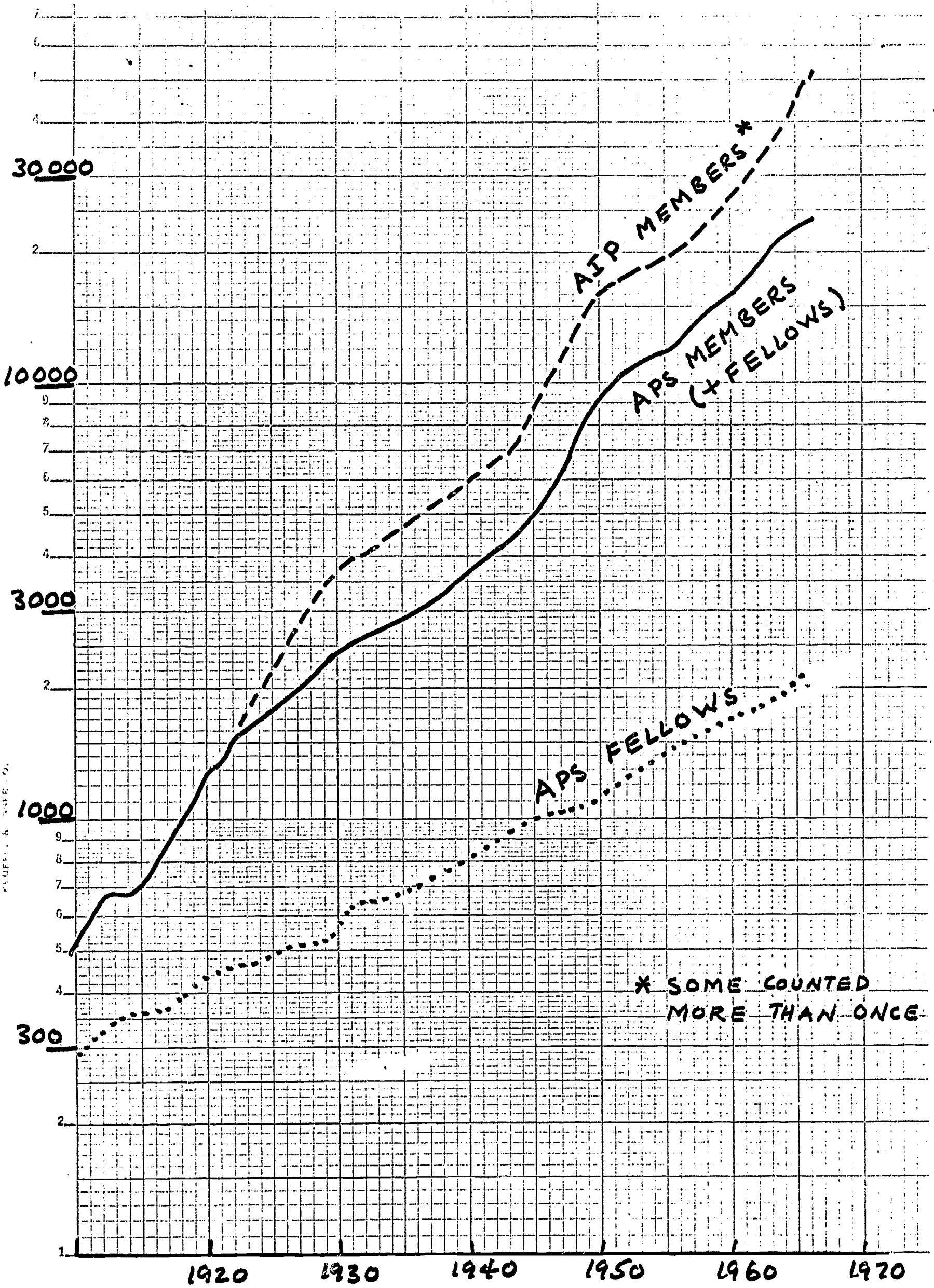
This is then the overall picture of what we are trying to do at the American Institute of Physics to face the problems of supplying services to the physicists who, themselves, are facing the problem of an exploding literature.


5000
2000
1000
500
200
200
100
50
20
US SCIENTIFIC JOURNALS
WORLD PRINCIPAL PHYSICS JOURNALS
1900
1910
1920
1930
1940
1950
1960
1970

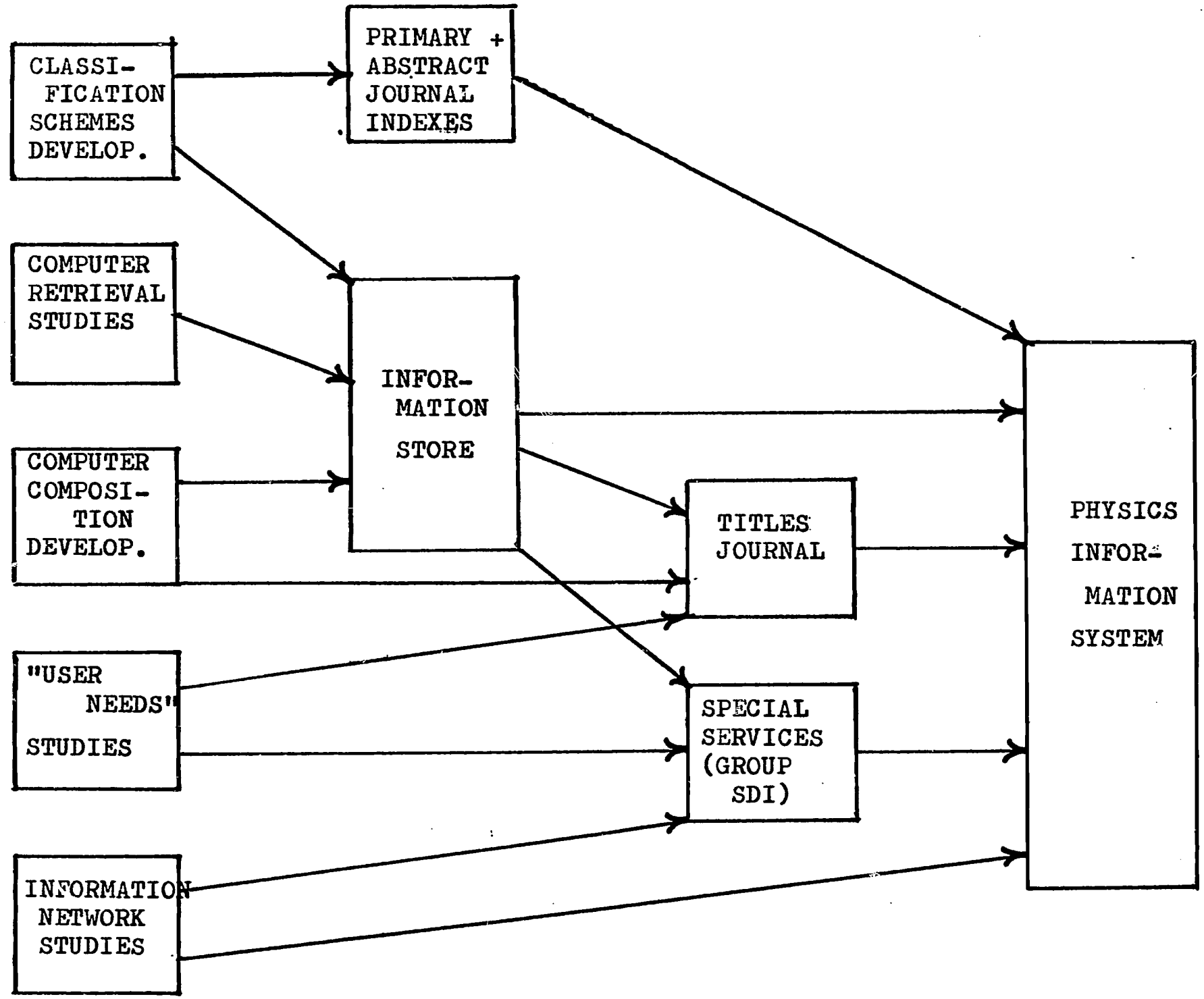

PHYS ABS
PHYS REV PAGES
30,000
10000
3000
1000
300
1910
1920
1930
1940
1950
1960
1970
SEMI-LOGARITHMIC 46 5490
3 CYCLES X 70 DIVISIONS MADE IN U.S.A.
KEUFFEL & ESSER CO.


AIP PAGES
PA
PR PAGES
50000
20000
10000
5000
2000
1000
1950
1960
1970

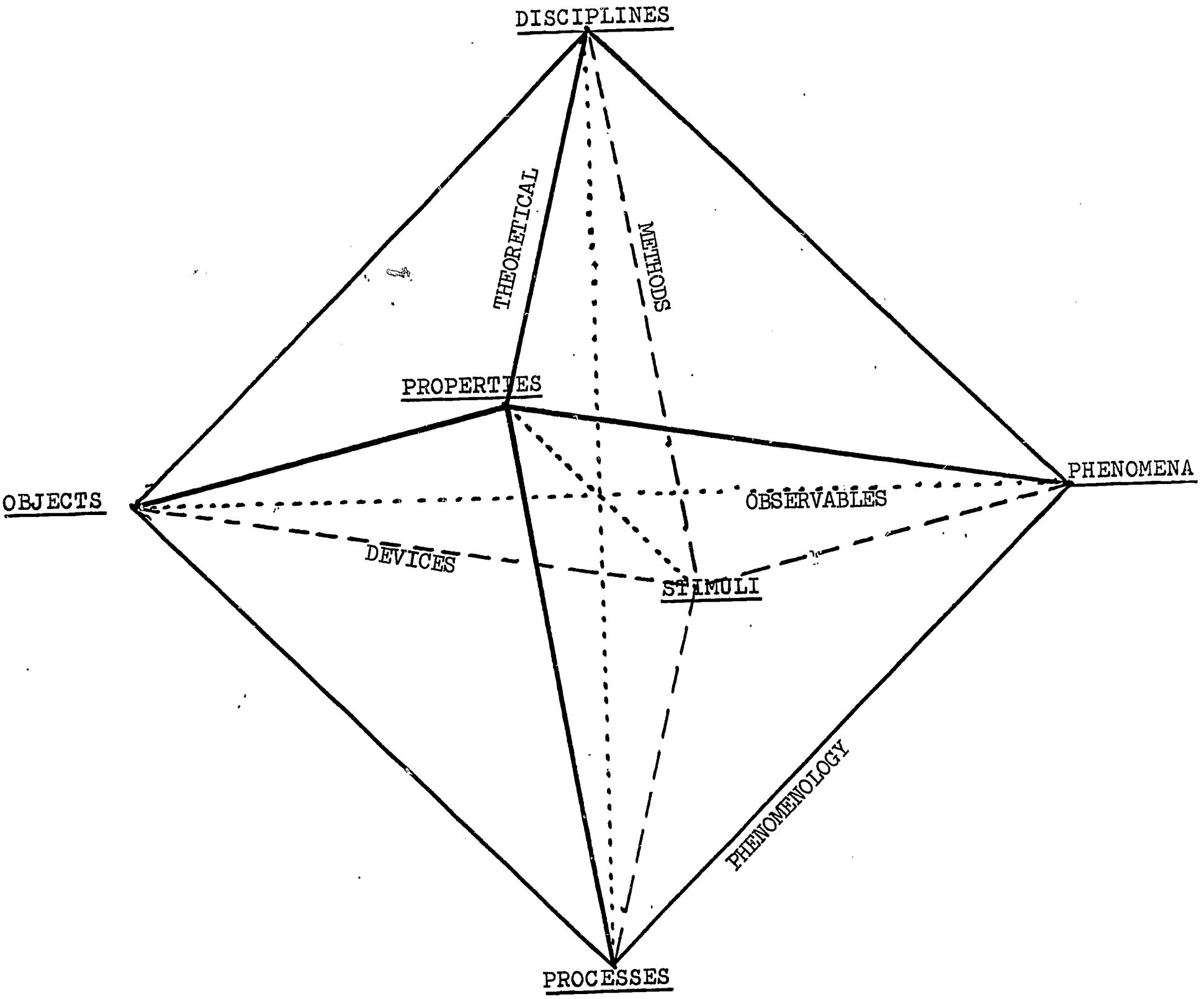

40-000
30-000
20-000
10-000
AIP PAGES
PHYS ABS
PHYS REV PAGES
1950
1960
1970


30000
10000
3000
1000
300
AIP MEMBERS *
APS MEMBERS (+FELLOWS)
APS FELLOWS
* SOME COUNTED MORE THAN ONCE
1920
1930
1940
1950
1960
1970

AMERICAN INSTITUTE OF PHYSICS INFORMATION PROGRAMS



INTER-RELATIONS AMONG THE FACETS



Lines represent aspects which facets have in common

EXAMPLES OF POSSIBLE UNIT RECORDS

BIBLIOGRAPHIC RECORD

PHRVA0900482 (53)
EFFECT OF SHORT RANGE REPULSION ON LOW ENERGY SINGLET NUCLEON NUCLEON SCATTERING
M H HULL JR, A HERSCHMAN
PHYS D, YALE U

JCPSA0260678 (57)
MICROWAVE SPECTRUM AND STRUCTURE OF PROPIONITRILE
R G LERNER, B P DAILEY
CHEM D, COLUMBIA U

INDEXING RECORD

OBject: NUCLEON, ProPerty: POTENTIAL
DiScipline: NONREL QM
ProCess: SCATTERING, PP: SCATT FN

OB: ORGANIC COMP (C_2H_5CN)
PP: STRUCTURE
PC: INTERNAL ROTATION
DS: MICROWAVE SPECTROSCOPY

KeyWords: HARD CORE, LOW ENERGY, CHARGE INDEPENDENCE, SINGLET, SQUARE WELL, YUKAWA

KW: TORSIONAL FREQUENCY, BARRIER HEIGHT, ISOTOPIC SUBSTITUTION

CiTations: PHRVA0810165, PHRVA0770441, etc.

CT: JCPSA0230184, PHRVA0790054, RSINA0210120, etc.