

April 15, 1980

FOR: F.B. Giller
FROM: A. Komendantov
SUBJECT: F. Colby Analysis of FEB 1980 IPDB RED Documents

In his recent letter, Dr. Colby raises some important issues regarding our processing of high-priority (RED) articles. I have reviewed the relevant IPDB printout and the cited articles and have concluded that there is a need to document or further clarify some of our activity in this area.

My general observations are as follows:

- Less than 50% of the documents in the February 1980 printout are new input. Most of the documents which appear were being re-entered into the IPDB for various maintenance reasons.

- We have backlogs at various stations which affect the entry time of documents into the IPDB.

- A small number of articles selected each year (usually reviews) cannot be loaded into the Permanent Data Base because the routine indexing that is required far exceeds the descriptor storage capability of the LRD system for a single document. These articles must be mechanically or intellectually reprocessed with the assignment of additional document numbers to carry the excess descriptors.

- The number of additional 'documents' generated in this manner accounts for one-half to one percent of our annual productivity.

Since Dr. Colby's letter goes into considerable detail, I have also provided detailed comments which follow his order of presentation (Please see attached materials).

A. Komendantov

April 15, 1980

Comments re F. Colby Letter (3.20.80)

1. I have confirmed that copies of the RED document printouts are not being sent to all users. Dr. Colby received this printout as a result of a specific SDI request that he initiated earlier this year. Printouts for Dec., Jan., Feb., and Mar. have been sent out to-date (I understand that these printouts are addressed to Max Crohn).

2. I believe Dr. Colby's goal with respect to these printouts is to monitor our ongoing monthly identification and input of high-priority articles ("...the most important 'new documents' entered by LRD into the IPDB..."). If this is the case, then we are sending him the wrong product. In addition to listing all "new" input for the month of February, the printout also reflects a considerable amount of Library and Editorial Services maintenance activity with respect to articles which have been in the IPDB for some time (this activity will be described in detail below). When this activity is accounted for, the actual numbers for new documents entered in February are:

1979 - 13

1980 - 11

These numbers do not seem out of line for this time of year. In order to avoid this sort of confusion in the future, I recommend that subsequent RED printouts either be edited to reflect only the new input or annotated in some manner to make this product more meaningful to the recipients.

I am not certain about the thrust of the comment on page length of 1980 documents. I have identified and reviewed six articles of 3 pages or less which appear in the printout. Two are helpful editorials, four are important research papers (see attachment, p.5). I do not believe there should be any question regarding their status as high-priority documents nor should their page length be used as an index of triviality, if only with respect to processing effort. A very high percentage of important, high-priority articles comes from these journals (Lancet, Science, Nature) and these articles are, as a rule, under 3 pages in length. Nonetheless, the abstracting/indexing effort required is comparable to that for articles of greater page length, published in less important journals. In any case, I do not see how these six articles demonstrate any problem with respect to our identification and input of RED documents.

Dr. Colby also refers to 15 1978 documents in the printout. The 1978 documents fall into three groups:

(a) A 19-part symposium on carcinogenicity testing. This item was ordered by us as a library reference in 1979. Since it proved to be very useful in connection with a recent user request for orientation in this area, we decided to fully-process this item and gave it a RED designation even though it was not "new."
(T#099602-20)

LG 2024244

(b) Three articles which could not be loaded into the Permanent Data Base because their indexing exceeded the LRD system capabilities. These required extensive artificial manipulation (splitting) before further processing could occur and were part of an ongoing program of reloading into the IPDB (this problem will be described in detail below). (T#099716-23, 100057-61, 100151-53)

(c) An article which originally presented processing problems at the scanning/accessioning stage and was backlogged for a period of time. It was re-entered as a RED to avoid any further processing delay. (T#100177)

3. Dr. Colby has calculated an average processing time ("delay") of 9 to 10 weeks for RED documents to be loaded into the IPDB. I am not certain how this was calculated from the printout data but these figures may very well be accurate. Our current processing time is related to two factors:

(a) the well-documented backlog situation and
(b) the new, expanded coding procedure which delays input initially, but significantly increases the retrievability of a document when it is loaded into the IPDB.

4. A bit of historical background is required to properly answer Dr. Colby's objection to the splitting of certain documents.

In the past, we have occasionally encountered the problem of a document being rejected at the Preliminary Update stage of input into the Permanent Data Base (i.e., "bombing") due to an excessive number of descriptors. When this occurred, we remedied the problem by choosing the most appropriate of three available options: (a) Deleting the excess descriptors and reloading as a single document (this is only possible when the excess is less than about 25 terms).

(b) Maintaining the abstract intact but splitting off one or more classes of descriptors and loading them in a related but separate document or documents.

(c) Splitting the original article into a number of logical components (sub-chapters, major headings, etc.), providing specific abstracting and indexing for each component, and reloading the article as a series of documents.

In late 1979, we began a major effort to eliminate the backlog of RED documents. In doing so, we found that many of the older RED documents (primarily extensive reviews and monograph materials dealing with carcinogenesis) were being rejected at the Preliminary Update stage because the number of descriptors significantly exceeded the document storage capability (by 100-200 % in some cases, so that option (a) was not available to us and (b) did not solve the problem readily either).

We were then faced with the task of reprocessing about 20 documents both retrospectively (those that had already been abstracted and indexed, but bombed) as well as prospectively (those that had

not been abstracted and indexed but which would, based on our experience, definitely or very probably bomb). Since this task involves a great deal of time and effort, it was staggered over a period of about five months to minimize its impact on normal processing. This activity was reflected in the RED IPDB printouts for those months, along with the routine input of new RED documents.

It is important to note that this splitting is not done to allow for "more thorough indexing and closer pinpointing of the indexing" but to provide a mechanism for our routine indexing standards to be applied to important papers whose user-relevant information content is extremely high.

It must be admitted that there is some "excessive" splitting when option (c) is selected (splitting the original article) because we try to follow a logical structure in addition to simply estimating the number of additional parts needed to handle the total number of descriptors (for example, if an article needs to be split into 3 parts to load all descriptors but is composed of 5 distinct sub-headings, we will generally split it into a series of 5 documents). We feel that this is preferable to a purely mechanical approach which could lead to searching and retrieval problems.

5. The first specific example of excessive splitting given by Dr. Colby refers to a series of documents in which the commentaries have been "separated from the document, even though such a commentary comprises only 1 or 2 pages."

This is a situation unrelated to the discussion directly above in that this is not a case of an excessive number of descriptors. It is, however, related to an abstracting/indexing problem which was detected in a number of 3i and early LRD documents, both by our own staff and by Dr. Colby, where a formal commentary or panel discussion section following an article was mechanically combined with the original article and not given its own bibliographic identity. In many of these cases, the opinions of the various discussants contradicted the findings and opinions expressed in the original articles, but this was either lost or obscured when the commentary of other authors was abstracted and, more importantly, indexed as part of the original article. As a result of this experience, we have been, for well over five years, conscientiously treating separate commentary and discussion sections as individual documents, integrally related to the original articles by a formal "SERIES" designation, even in cases where these sections comprise only 1 or 2 paragraphs. The validity of this procedure is even more readily apparent when the commentary section follows two or more articles or presentations and refers to all of them. In the specific example cited, furthermore, the commentary sections stand alone and have their own designation in the table of contents. This is a good indication that they will be cited separately in secondary sources as well. (T-099602-20)

6. The last series of examples is a good illustration of the types of articles we have been reprocessing (splitting) in order to load them into the Permanent Data Base. All five items noted by Dr. Colby are major reviews dealing with carcinogenesis:

- (a) Occupational carcinogens (Schottenfeld); 24 pgs.
- (b) Vitamin C and cancer (Cameron&Pauling); 19 pgs.
- (c) Chemical carcinogenesis (Miller); 18 pgs.
- (d) Occupational carcinogens (Schottenfeld); 13 pgs.
- (e) Air pollutants and cancer (Cederlof&Doll); 10 pgs.

None of these items can in any way be considered "relatively short." In addition, the user-relevant information content of all five of these reviews is so high that we are almost forced to index every line of text with several descriptors.

I would like to stress once again that these articles simply cannot be loaded into the Permanent Data Base as single documents, using our routine indexing standards.

7. As far as productivity calculations are concerned, it should be stressed that the total number of articles which require splitting is very low. Recent printouts which show a relatively high number of these items are not typical, but are reflecting a current special effort to fully process all backlogged RED documents. On a prospective basis, the number of such articles is about 10 per year. The additional documents generated by splitting these 10 articles have a negligible effect on productivity calculations; their contribution to the annual productivity figures is between one-half and one percent, depending on whether the splitting is being done mechanically or by an editor.

A. Komendantov

Attachment, Comments re F. Colby Letter (3.20.80)

099621 THRODDAHL, M
THE PENCIL PROBLEM--1990 *** CHEM ENG NEWS, VOL 58(13) P5
, 1980 CATEGORY-020,203, 209, 239 MONSANTO CO, NEW
YORK, NY

099642 JONES, JG/ MINTY, BD LAWLER, P HULANDS, G CRAWLEY, JOW
VEALL, N
INCREASED ALVEOLAR EPITHELIAL PERMEABILITY IN CIGARETTE
SMOKERS *** LANCET, VOL 1(8159) P66-68 , 1980
CATEGORY-020,200, 201, 204, 212, 239 MED RES COUNC
CLIN RES CENT, LONDON, UK/ MED RES COUNC CLIN RES CENT,
LONDON, UK

099646 MOSSMAN, RT/ CRAIGHEAD, JE MACPHERSON, RV
ASBESTOS-INDUCED EPITHELIAL CHANGES IN ORGAN CULTURES OF
HAMSTER TRACHEA (INHIBITION BY PETINYL METHYL ETHER) ***
SCIENCE, VOL 207(4428) P311-313 , 1980 CATEGORY-020,
202, 220, 227, 239 U VERM COLL MED, BURLINGTON, VT/ U
VERM COLL MED, BURLINGTON, VT

099869 HOPKIN, JY/ EVANS, HJ
CIGARETTE SMOKE-INDUCED DNA DAMAGE AND LUNG CANCER RISKS
*** NATURE, VOL 282(5745) P388-390 , 1980
CATEGORY-020,200, 202, 239 MED RES COUNC CLIN POP
CYTOGENET UNIT, EDINBURGH, UK WEST GEN HOSP, EDINBURGH, UK
U BIRM MED SCH, BIRMINGHAM, UK U BIRM MED SCH, BIRMINGHAM,
UK/ MED RES COUNC CLIN POP CYTOGENET UNIT, EDINBURGH, UK
WEST GEN HOSP, EDINBURGH, UK

099871 SOLDATOS, CR/ KALES, JD SCHARF, MB RIXLER, FD KALES, A
CIGARETTE SMOKING ASSOCIATED WITH SLEEP DIFFICULTY ***
SCIENCE, VOL 207(4430) P551-553 , 1980 CATEGORY-020,
200, 211, 239 PENNS STATE U COLL MED, HERSHEY, PA/ PENNS
STATE U COLL MED, HERSHEY, PA

099989 ANONYMOUS
WHY THE AMERICAN DECLINE IN CORONARY HEART-DISEASE. ***
LANCET, VOL 1(8161) P183-184 , 1980 CATEGORY-020,200,
205, 239 NONE