Lee Duss Both

Department of Energy
Washington, DC 20585

FEB 18 1992

Mr. Gerald Cranford
Director, Information Resources Management
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555

Re: Suggested Revisions to TWG Paper

Dear Mr. Cranford:

The following represents revisions and additions per the assignment made at the last meeting of the Technical Working Group.

REVISIONS TO CURRENT TEXT:

Page 5, in the paragraph beginning "DOE has offered . . .":

> ". . . INFOSTREAMS is being designed to use DOE's existing Digital Equipment Corporation VAX/VMS hardware and BASISPLUS software. DOE notes that additional customization would be needed to re-develop relevancy ranking algorithms embedded in INFOSTREAMS' content/relevancy based retrieval software. Furthermore, the Department of Energy's contractor (TRW) has emphasized that scaling requirements are a serious concern to migration of the INFOSTREAMS software to NRC, and DOE agrees with this assessment. Only selected information will be . . .".

First paragraph on page 12:

> "~~The Working Group recommends adoption of this alternative for a number of reasons.~~ Given that INFOSTREAMS can closely mirror LSS processing requirements and will be doing so on a very large scale, it would be most economical for DOE to assume responsibility for the other 10% . ."

then, at the end of the same paragraph,

> ". . . The Working Group believes that adequate quality checks can be instituted by both the LSS Administrator and LSS participants to assure accurate ~~and timely~~ processing of non-DOE material through INFOSTREAMS."

and, then add the following new paragraph,

> "With regard to this alternative, DOE has serious concerns about any plan that would give DOE responsibility for entering other parties' submissions. DOE's access to, and control over, other parties' materials was very contentious during the negotiated rulemaking. Furthermore, there are serious policy and operational

questions related to responsibility for intake prioritization, liability for accuracy and timeliness of entry of other parties' materials, and budgeting and procurement of incremental resources. DOE's level of concern is such that it can not concur in recommending this for any further consideration."

NEW SECTION II:

The earlier section reflects a series of alternatives for reallocating or reducing costs based on reuse of DOE's INFOSTREAMS technology, reduced functionality, reduced availability, and other strategies. Many of the items from the chart entitled LSS Concept/Design Alternatives Summary are interdependent or could be considered in conjunction with others as part of a "package". In this section, we are going to address some discrete strategies.

Sample Strategy #1: Is it feasible that not all documents be included with searchable full text, but rather made available via bibliographic header and bit-mapped images only?

Text conversion is the single most costly element of all LSS processes. It was incorporated into the LSS design as a blanket requirement for all documents: 1) before the header fields were decided, 2) as a response to legal representatives who were familiar with the technology, and 3) recognizing that subject cataloging had inherent deficiencies.

Conversely, the text of some documents adds little or nothing to their retrievability if they have been competently and fully cataloged. A bibliographic header does provide search and retrieval capabilities and is an appropriate level of treatment in some circumstances. By using bibliographic headers, where the associated image is available on line, participants still have access to those materials.

Some situations typical of the LSS document collection are amenable to differential treatment:

> Transmittal memos and letters attached to reports, studies, etc., are often not content rich. Rather, it is the item being transmitted that contains the information of value. So, if attachment relationships and cross-reference fields are properly designed, and if the item attached, itself, is full-text searchable the entire package (transmittal and report) is still eminently retrievable via text search.
>
> Another situation consists of the flip charts and other presentation materials which are attached to textual meeting minutes.

DOE's contribution to the LSS holdings is estimated to be 80% of the low-volume estimate and 86% of the high-volume estimate. Of DOE's contribution for the low volume estimate, 4,320,000 pages (65% of all its documents but only 12% of its pages) will be correspondence (letters, memoranda, telex, etc.). A simple analysis of impact for not including text for all such

correspondence follows:

Reduce all OCR intake labor for this proportion of the material:

OCR Pre-processing ($10,136,000 x .12)	$ <1,216,320>
OCR Cleanup ($16,650,000 x .12)	<1,998,000>
Scan/Text Supervisors ($4,864,000 x .12)	< 583,680>
Disk Storage ($5,120,000 x .12)	< 614,400>
Increment for offsetting increase in hardcopy printout from image ($9,045,000 x .12)	1,085,400
Net True Savings for Bibliographic Header & Image, but no full text for DOE correspondence:	$<3,327,000>

This is a simplistic presentation insofar as storage would not really decrement proportionately, since text from correspondence is less character-dense than equivalent pages of reports and publications. And, it is more palatable if text is omitted only for correspondence that is attached to a text-searchable report, resulting in a smaller percentage reduction. However, it is representative of strategies focused on the peculiarities of the document collection and knowledge of users' retrieval expectations.

Sample Strategy #2: What economies could result if data accuracy requirements reduced to 98% accuracy rather than 99.8% because "intelligent" retrieval software compensates?

DOE's LSS Prototype showed that the most accurate OCR device tested achieved an average character accuracy of 98.6%, which corresponds to 25 errors on an average 1800 character LSS page.[1] In that same prototype, it was found that text accuracy must approach 99.8% or users would lose confidence that they were able to retrieve all critical documents, and thus lose confidence in the LSS itself. Under SAIC's design, documents for which the OCR output accuracy was not in the 95-98% range would require additional editing that would exceed the cost of a complete, manual rekey of the entire page.[2] It was also found that, on average, editing represents from 65-75% of the total cost of text conversion, and that correcting OCR-induced errors constitutes 67% of that total editing cost. The multiple-OCR device approach for intake, reflected in SAIC's final cost estimates, was based on analyses showing that 2/3 of OCR-induced errors could be eliminated by merging and matching streams from multiple OCR devices.

DOE is studying content/concept based retrieval software to augment classic Boolean tools. The most significant aspect of this software is that it profiles a document's content. But, one unanswered question is how much of the document must be "read" before all the relevant terms and topics have been identified?

---

[1] Dickey, Lois. "Operational Factors in the Creation of Large Full-Text Databases", INFOTECH '91. p.41.

[2] Ibid., p. 45.

If, after analyzing the first 20 pages (@99.8%) of a 350 page report, the software "knows" what the document is about, isn't the rest of the document superfluous in adding to our understanding of its content? If after these 20 pages our matrix is already "saturated", we have 330 pages where the input accuracy of the text could be as low as 90% and have no impact whatsoever on our ability to characterize the document.

What if the entire document were 90% accuracy? -- then it may take an additional 10 pages of text to find a "clean" occurrence of terms and topics before the matrix was again saturated. But, again, 320 succeeding pages contribute nothing more to our understanding.

In a way, the "intelligence" of the software compensates for typographical errors by having access to enough bulk ASCII, with enough clean text, to be able eventually to correctly characterize the document. DOE's testing still has to validate the concept. For example, DOE does not know where the "saturation" level is, and if it is affected by the overall length of a document. How much "clean" ASCII is needed? Will it work as well on an eight page letter as it will on the longer report? Will a high percentage of uncorrected ASCII result in an unacceptable level of false characterizations, resulting in associated false drops during retrieval? How would the highlighting of occurrences of terms in text be implemented in a "dirty ASCII" environment? Can the matrix compiled during the filtering of incoming text be added somehow to a simple bibliographic header with associated image, obviating the need for text?

Academic papers about the new software packages which utilize "fuzzy logic" indicate that this approach will work for search and retrieval -- and this would be sufficient because it is the images and not ASCII that are relied on for introduction as exhibits. Will LSS users be satisfied that such intelligent software is able to compensate for typographical errors? If they could be convinced, and, if one of these new software packages is roughly comparable in cost to a current state-of-technology, Boolean-based package such as BASIS+, then the following scenario could apply: we could remove the multiple OCR devices from SAIC's final design, and accept the basic 98% text accuracy with no additional text editing and OCR cleanup.

A simple calculation of savings is as follows:

5 Capture Systems requiring less OCR hardware:	$< 500,000>
Eliminate OCR compare software:	< 100,000>
Eliminate OCR Cleanup Staff:	<16,650,000>
Eliminate OCR Cleanup Supervisors:	< 1,200,000>
1 Correction Station's Text Cleanup Eliminated:	< 1,050,000>
Total Savings for 98% accuracy:	$<19,500,000>

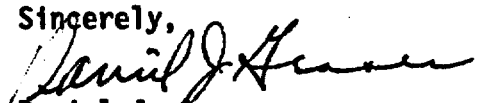
A demonstration of this strategy was made to DOE utilizing commercially available off-the-shelf technology, EXCALIBUR software, which is VAX compatible but does not work in conjunction with DOE's current BASIS+ records management software.

Both of the approaches outlined above would represent major deviations from what the parties agreed to during the Negotiated Rulemaking, and they would all be contentious to varying degrees.

OTHER VARIANTS:

Other strategies are conceivable: if bibliographic headers could be made to store the entire subject-content matrix of every textual document (derived from DOE's expert system software), the OCR and text analysis processes would not decrease, but perhaps no text at all would have to be stored or retrieved. This would have major impacts on the amount of disk storage, the size of the search engine hardware, the organization of databases (no partitioning), database loading and maintenance, the size of the telecommunications lines, etc. The cost ramifications of such strategies would require robust, detailed, and professional feasibility and benefit-cost studies outside the scope of this paper.

Sincerely,



Daniel J. Graser
Program Analyst
Information Management Division
Office of Civilian Radioactive
Waste Management

Copies:

B. Cerny, RW-12
J. Bartlett, RW-1