情報検索システムの比較評価法について
——Aslib-Cranfield Research Project の意義——

On Comparative Evaluation of Retrieval Systems
——Significance of Aslib-Cranfield Research Project——

中　村　初　雄
Hatsuo Nakamura

Résumé

The writer of this article pointed out in his contribution to a 1968 Festschrift honoring Fujio Mamiya, that there exists in Japan and elsewhere an naive admiration and overestimation towards evaluation technics of indexing methods, and in the essay criticized a recent commentary of M. Takahashi rather severely. However, after reading some of the critical comments on C. W. Cleverdon's Aslib-Cranfield Research Project, the writer reflects that his own severe attitude may serve to discourage rather than to encourage the progress in such research.

The writer has a great esteem for Cleverdon's report of 1962. It is most challenging and epoch-making. Introducing the test results of the Aslib-Cranfield Project and the underlying hypotheses, the writer tries to make clear main issues which are attacked and criticized by P. A. Richmond and other critics. Some of the attacks on these issues are open to refutation, others come from overcautious care or from lack of understanding towards the hypotheses. Though C. N. Mooers has defined in 1959 the exact set of relevant documents in the whole store, it was rather only in ideality. In actual fact Cleverdon's report made more concrete and popular the concepts of recall and relevance to experiment designers, who are in search of so-called objective measures for retrieval systems. According to the discussion session following the 1960 Aslib Conference, Cleverdon himself was aware of unavoidable weaknesses which lead to certain controversies. He is still continuing his tests and is trying to supplement and retouche his report.

The most impressive result of his achievement, the writer thinks, is the analysis of unsuccessful retrievals. This result may have been foreseen by an expert librarian, although this fact does not belittle the credit due him.

His expression of the quantitative evaluation of the system is quite useful for the future designing of indexes as well as for use in training courses.

(School of Library and Information Science)

はじめに

Brownson 女史[1]も指摘しているが，最近の10年間というもの，特に検索システムをいかに客観的に試験し評価してゆくかについての関心が高まってきている。それは当然のことであり，科学・技術上の文献を貯蔵しておき，後日，探索・取りだして利用する新規な方式を考えたり，開発してゆく為には是非必要なことである。従来からあった伝統的な方式に較べて，新方式が，またその機械化がどんな点で優れているかを確認するためには，諸性能についての客観的評価の方法を持たねばならない。いくつかの改善・進歩は既になされてはいるが，しかしこの評価法については，まだ今後に期待しなければならない点が多い。

American documentation の 1955 年 4 月号には，Jesse H. Shera が巻頭論文を書いているが，彼がそこでドキュメンテーション・システムと言っているのは，主として検索システムのことを指しているのであるが，その効率を評点してゆくには，全ての実験結果を綿密周到に評価する必要があると言っている。検索システムは実用的のものであって，[机上の] 論争などで決定されるのではなくて，要求と経費との関連即ち能率といった見地から，その実用性を知的に測定していって決定されるべきものであるとした。

太西洋をへだてた英国では，Aslib 航空学グループ委員会は，1955 年 8 月 25 日に謄写刷の報告ではあるが "Programme for research into information retrieval" を発表して，前述の Shera 論文を引用している。その中で，英国の図書館員は戦術転換を要すると述べているのである。このような状況下にあって，1957-61 に Cleverdon により指導された Aslib-Cranfield 研究調査は行なわれたのである。

その最初の報告に対し Metcalfe[2] は机上の推論が主の比較であって，評価してゆく際に用いるべき規準となるものが設立されたとは言い難い。今後の決定的な実験を待つべきである，と批評している。

それよりも更に鋭い批判をあびせているのは Richmond[3] である。彼のは "Aslib-Cranfield 研究調査は，唯一の比較研究とも思われがちであるが，実はあの 4 システムの効果比較には弱点がある。極言すれば無意味であるともいえよう。検索システムとして用いるものに，大ざっぱにわけて，稀釈的近接と濃縮的近接と 2 種類あるのだが，あの報告では，UDC だとか一般的 abc 順件名標目表によるものも濃縮的近接であるかの如くに扱っている。これではキメのこまかい比較は出来ない筈である" といった論調で問題点を指摘している。

筆者が Cleverdon の第一報を読んだのは 8 年前のことであるが，自然科学者が研究計画をたてるのと似た手法がみられ，感心して読ませられた。即ち Aslib の戦術転換に拍手をおくったことを思いだし，再読検討してみることにした。新らしい手法をとりいれたが故に感心してしまい，無批判に受け容れたのであれば反省を要するし，また Richmond の批判が無理な要求であるかについても検討してみようというのが，この題目を選んだ理由である。

I. Aslib-Cranfield 研究調査の成立

この研究計画は Cyril W. Cleverdon 指導のもとに，下記のものたちによって実施されたものである。この計画に対しては National Science Foundation が資金援助を与え 1958 年 4 月から実施された。

J. Sharp（航空学関係索引専門家）
J. Hadlow（公共図書館から参加）
T. Opatowski（航空関係技術者）
但し 6 ケ月で他の専門分野からの図書館員 Miss B. Warburton と交替

蓄積された文献・資料（文書）

総数 18000 点で，長短とりまぜ，半頁の報告から 200 頁のものまであった。約半数が研究報告(平均30頁)，半数は雑誌論文・寄稿(平均 7 頁)であった。北米合衆国からの文献が約半数，残りはその他の国々からのものである。主題分野で区別すると，大体半分は，高速航空力学関係のもので，他は航空学に関連のある一般主題のものである。

試験の対象となる 4 つのシステム

1. UDC これは，アメリカ，特にヨーロッパで用いられる列挙式分類の例として選んだ。試験実施のた

めに，普通のカード目録の形式で作成した。620と629にも若干展開を行ない，特に532.5と533.6の部分には特別の詳細展開を使用した。UDCの合成法はよく失念されたり，間違いをおこすことが多いので，abc順索引を必らず作っておくことにした。これはSharp氏の責任で，調整を行った。新規番号が使われると必らず言葉からの記入を作成，Sharp氏が公認した上で索引に編成，コピーを索引者全員に配布するということにした。

2. 件名標目表　これは，いろいろの件名標目表を検討したあげく，アメリカ専門図書館協会の A list of subject headings for use in aeronautical engineering libraries. 1946. を用いることにした。これにも勿論欠点はあり，適当な標目がなかったり，あってもあまりにも大ざっぱすぎる包括概念で，細密度が不足していることもあった。標目追加用の指針ともいうべきものがなかったので，E. J. Coates の *Subject catalogues: headings and structure* で補充していった。新件名はみな Hadlow 氏に提出して，既に使っている索引標目とも較べて，基準形の標目を作っていった。参照は，「を見よ」参照だけにとどめておいた。(将来計画では，分類表なども準備した上で，「をも見よ」参照を作る予定。)

3. ファセット分類　これは，主題を分析していって，個々単一の性質（性格の要素とも言えよう）でカテゴリーを表現してゆくのを，約束できめた順序で組合せてゆく方式で，検索を容易にするため，チェーン索引が用いられている。あまり勝手な順序，組合せがおこるのを防ぐために，ファセット分類の原編者 [Vickery と Farradane を指すものならん] と作業者との間には何回かの話しあいを行った。その際特別に追加した約束は，Opatowski 氏が全部を見なおして，必要と承認したものだけを追加して，後日の追加は認めないように，凍結してしまった。

4. ユニターム方式　これは Miss Warburton の責任に於て，新規に採用すべき「ユニターム」の典拠カードファイルを作成・編成して，管理していった。大型（8吋×15吋）の，アスペクト・カードが使用された。ユニターム索引については，高橋正明氏が“科学技術情報管理”1967 の p. 288 ff で紹介しておられる。また筆者も“びぶろす”1956年8月号に，“専門図書館と資料検索のあり方”と題して紹介したことがある。

索引作業（検索システム作成の段階）

18000 の資料を 6000 宛，3 のグループにわけ，それぞれを更に 100 点ずつ 60 の小グループにわけておいた。100点ずつの小グループを，3人の索引者が，それぞれ4種類のシステムで，しかもまた1文書あたり所要時間2分，4分，8分，12分，16分という5種類の時間配当で索引していった。

4つのシステムによって，索引作業をした結果について比較データをあげておこう。これは同時に，それぞれのシステムの索引言語としての性能のある面を推察させることにもなろう。[6]

第1表　(原報告 1.1 表)

システム	標数（標目）ユニターム	備考
UDC	2350	付属索引における言語標目の数 4052
件名目録	2864	副標目 592，参照 1560
ファセット分類	1686	標数 notational elements
ユニターム	3174	(固有名 607 数・量を表示するもの267)

本調査以前の検索システム比較評価の例

Gull[7] Thorne[8] などがあげられるが，Cleverdon は特に Gull 報告を意識して，というよりも，他山の石として自分の研究調査を計画していったとみるべきであろう。Cleverdon は1962年の本報告，第2章の“主テストプログラム”で次の如くに言っている。

1953年に行なわれた ASTIA—ユニターム 比較試験は完全な報告になっていない。15000 点の資料を索引にしているが，ASTIA の方はその職員が abc 順目録を作成してゆき，ドキュメンテーション研究会の方はユニタームで行っているので，比較するのは困難である。時間も正確にはコントロールしてなくて，ただ材料とした資料が共通だというだけである。質問の解釈，検索が成功したか否かの判定も各グループがしたのでは客観性がすくない。ASTIA の方は abc 順件名標目表による目録の方がよいと結論し，他の方がユニタームはより効果的であるとするようになったのも無理からぬことである。(注5参照)

実験の計画

Cleverdonは自分の比較試験の結果が，ある程度信頼おけるものにするためには，最少限1600の質問を準備する必要があると考えた。これは大変な大事業であるが，次の如き方法で作成していった。

質問には必らず一つは正解が出てくるように，資料自体から作成することにして，この大作業を簡単にしようとはかった。その配分は，最後の第3グループ6000の中から75%を作成し，他の2グループ120小グループの中からは，平等にばらまかれるようにして残りの，25%の質問を作成していった。質問作成に関しては，英・米・加・和の50機関の代表者達に協力を求めた。同一資料に基づく，同種のまたきわめて類似した質問はとりのぞくことにしたので，約1500の質問が成文化されることになった。

実験の当初における最大目標は，4つのシステムの効率比較にあったので，関係者3人全部が慣れた後の第3グループの資料6000点を対象としたテストでの調査に重点をおき，質問作成の75%をこれに求めたというのは，理解出来る。このグループのみについて言えば文書総数に対し約20%の1200質問が用意されたわけである。

第一次の試験は400質問で，(その中300は最初の予備試験から) 4色の色紙で各システムの区別をさせ，回答用紙1600枚を発行した。一つの検索実施で記憶しているために，その後で実施した検索の結果が影響されることがあっては，不公平な比較になるとの懸念から同一の問題を別のシステムで検索するまでには，1ケ月以上経過してからということにしている。

この400質問に対する1600検索を，3人のスタッフで行い，その経験を話しあい「どの程度迄検索を行うべきか？」について一応の線を出し，第二回の1200質問テストをする際の指針としている[9]。

II. 問題点と実際の判定例

検索効果評価実験直後の討議

これについては，特に，Farradane が座長になって，諸家の意見を徴したことがあるが，その時にも強調されたことであるが，この研究調査は，基礎調査であるということである。Thompson, Sharp, Miss Kyle などがそれぞれ，ファセット分類のようにあまり普及していないシステムを加えて，評価していったことに対する質問をしている。Cleverdon はそれらの困難を見逃していたのではなく，承知の上で敢てこの比較調査を行ったものであることは，彼の回答からも窺がわれる。

今後もなお追加のテストをしてゆく必要は認める。資料が出てこなかった場合をごく大ざっぱにではあるが分析・検討してみると，UDC とユニタームの場合は索引作業の段階でのミスによることが多く，ファセット分類の場合は検索のミスによるものが多いようだ。[4]

Liebesney がこの調査では，検索に要した時間のことは何も問題にしなかったのか？ と質問しているが，これも計画としては National Science Foundation の補助があっても，出来なかったことなのであろう。このことはまた Lenel が，報告執筆者がどういう風にして題名を選んでいったか，また索引係はどういう具合に標目を作成し，検索者はどのような手がかりで検索していったかの，心理研究が必要なのではなかろうか？ といっているのにも通ずることであろう。この種の質問がいつでもおこることに対する筆者の見解は既にいくつかの，図書館人論で述べたことであるので，繰りかえさないが，Cleverdon 自身の回答を解説しておく。

この種の比較研究は，検索プログラムの樹立，検索作業の実施，適合・非適合資料がどんな風にとりだされてゆくかの測定をしてゆくために行なわれるのであるが，今回の実験では，検索実施に関するデータは断念せざるを得なかった。それはそれぞれのシステムで，またそれの補助用具，設備等の有無，実施法などにも関係してくるので，検索実施の段階での所要時間調査までは計画にいれなかった。

また心理学的研究は一つの大きな課題であり，興味あることかもしれない。初期の段階に，短期間ではあったが，索引者の気分（気嫌）についての個人的日誌(メモ)をつけたが，実際に役立つようなことは何も発見出来なかった。もしも10点なら10点の同一セット資料を100人に索引させて，それを他の100人に検索させてみて検討してみたら，あるいは利用価値のあるデータが得られるかもしれない。(注4参照)

なお適合・非適合資料の判断については後に Rees 等の論文によって触れる予定であるが，ここでは，検索が成功したか，不成功であったかはどうして判定してゆくかについて述べておこう。

a. 不成功：質問に該当する文書，質問者の質問に役

立つような情報を含む文書は探しだせなかった。

b. 部分的成功または半成功：文書がとりだせたが，それは質問を部分的に満足させるものしか含んでいなかった。

c. 成功：質問者の要求にピタリの文書を1またはそれ以上（その数は質問のタイプによって差異あり）探しだせた。

ここで，a, b の結果の場合には，第2の検索[10]を行うわけであるが，それを行うにあたり，質問者の方で待ちきれなくなる程時間がかかってしまうこともあれば，検索計画が樹立出来ないこともあり，また第2の検索で，成功するということもある。

検索成功判定の実施例

これには後日も種々の批判が集中されることになろうが，ここには，基礎研究調査として，これだけの配慮のもとに，即ち問題点はいろいろとおこることは承知の上でなされたものであるということを解して頂くために，一例を紹介しておく。（注5参照）

アルミ合金鈑のうち腐蝕を防ぐために純アルミで被覆をした所謂 Alclad 鈑とそうでない鈑をリベットつけした場合の接目の疲労負荷についての情報を求めるといった質問を例にとる。

UDC ではアルミ合金 669.715 と疲労試験 620.178.3 を組合せて12文献をとりだしたが，その中には質問作成のもとになった文書は発見されなかった。逆順列で探すと 31 の文献がとりだされたが，その中にも発見されない。結局は，669.715：620.178.3：621.884.057.2 のところに求める文書はあったのである。第三番目の要素「リベット」を考えなかったので失敗したわけであるが，最初の2要素は一致しているので，この件は第一検索で成功したものとして判定。

件名目録の場合，最初は，“薄鈑，アルミ合金—疲労”で探して不成功，二回目“アルミ薄鈑”で探しても不成功，・・・六回目には“接目リベット，アルミ鈑金—疲労”という探し方をしてようやく，原文献にとりついたのである。それまでに結局34の他の文献がとりだされ，一々精査していかねばならなかったのである。

ファセット分類の場合，最初の探索計画には次の4要素を考えたのである。

“アルミ合金” Peal-a “疲労” Rkm “リベット” Hvd “薄鈑” Fsb

この場合，チェーン索引の一番最後に出てくる要素は“疲労”であるから，そこから手をつけて，4要素を持つものはと探すと，1文献にぶつかるが，その文献は求めるものとは別のものである。第2の探索は“薄鈑”という要素をもとに，探してゆくと，2文献がが出てきて，その中の一つが求める文献であった。

ユニタームの場合，最初は“アルミニューム”であるが，そこには勿論あるが，あまりにも数が多すぎて話にならない。これに“薄鈑”のカードを求めてゆくことにする。しかしこのカードを調べてゆく時に気づいたことであるが，この概念のもとで，鈑がリベット接ぎに関連させられるような工夫はされていない。ただ Alclad（アルミ被覆）をも探すようにとの指示がされていたにすぎない。しかし“Alclad”のもとには求める文献はなく，“疲労”を見るようにと指示してくれたにすぎない。ところがこの“疲労”というカードの中に，この求める資料番号があったのである。しかし実際はこのカードに含まれている資料数はまだあまりにも多すぎたので“強度”というカードで限定していったのである。検索実験を行った人の中には，“鈑金”(plate) で索いた場合そこに“アルミ合金の疲労抗力”という語が参照してあったので，その時成功したと見るべきであると感じた人があったが，その立場はとらずに，最初の2回は無駄な検索を行ったのであるから，3回の独立した検索の結果成功したと判定した。

このような検討を1200なり1600の質問について行ってゆくのであるから，これは相当な規模での研究調査であることが察せられよう。勿論，個々の判断・判定について疑義を指摘してゆけばそれは際限のないことであろう。

検索本実験の結果

4種の索引について，1200質問で検索した結果。参考のため300質問で行った予備試験の際のデータを付す。

第2表 （原報告 3.1 表と予備報告）

索引システム	総検索数	成功	失敗	成功率 本試験 %	成功率 (予備)
UDC	1157	875	282	75.6	(78.6)
abc 順件名	1154	941	213	81.5	(81.5)
ファセット分類	1047	773	274	73.8	(71.4)
ユニターム	1146	940	206	82.0	(78.5)
	標準偏差			2.6	(4.6-5)

ここで標準偏差の算定は次式によった

$$\sigma_p=\sqrt{\frac{PQ}{N}}$$

N: サンプル全数
P: 成功部分
Q: 不成功部分

統計分析の方法についての報告は，J. T. Harris が別にまとめている。[11]

ここではあまり，統計学的操作を加えない生のままの検索成功率表の若干を紹介しておく。

第3表（原報告 3.2 表と予備報告）
システム別，索引作成所要時間と成功率表
（検索実施は，研究計画スタッフの場合）

システム＼一点当り	16分	12分	8分	4分	2分	[計]
U D C	82(76)	80(90)	74(75)	77(78)	72(75)	385
a b c 順	89(90)	85(85)	77(78)	85(77)	73(75)	409
ファセット	76(68)	79(77)	71(65)	71(67)	71(73)	368
ユニターム	89(85)	85(82)	83(75)	88(89)	75(63)	420

本実験はいずれのセクションも 173-180 検索を行なっての結果。（ ）内の数字は予備テストの結果。

第4表（原報告 3.2 表と 3.7 表）
システム別，索引作成者別の成功率表（同上）

システム＼索引作成者	Hadlow	War-burton	Sharp	[計]
U D C	73.8(77)	81.0(76)	76.9(76)	231.7
a b c 順	79.6(80)	85.4(84)	82.7(77)	247.7
ファセット	71.4(74)	77.9(71)	71.1(70)	220.4
ユニターム	84.0(83)	83.0(79)	86.0(82)	253.0
[計]	308.8	327.3	316.7	

（ ）内の数字は最初のグループ 12000 点を対象に，但し質問数は約3分の1にし研究計画スタッフ以外の技術者に検索させた場合のデータ

第5表（原報告 3.4 表，3.11 表，と予備報告）
システム別，主題別と成功率表（同上）

システム＼主題	航空力学	その他一般
U D C	72.7 [78] (77)	78.6 [77] (82)
a b c 順	78.9 [79] (72)	84.2 [84] (74)
ファセット	70.3 [69] (62)	77.3 [75] (72)
ユニターム	81.6 [80] (81)	82.2 [77] (81)

（ ）内の数字は前表に同じ，[ ] 内の数字は予備報告よりのデータ

第6表（原報告 3.8 表と予備報告）
検索を，研究計画に参与させなかった技術スタッフに実施させたときの検索成功率

検索システム	成功	不成功	成功率(%)	予備試験
U D C	520	132	79.6	(81)
a b c 順	501	183	73.3	(75)
ファセット	335	167	66.7	(62)
ユニターム	297	69*	81.1	(76)

* 原報告 619 はミスプリント

第7表（原報告 3.9 表）
システム別，索引作成所要時間と成功率表

システム＼一点当り	16分	12分	8分	4分	2分	[計]
U D C	84	86	78	78	76	402
a b c 順	81	78	74	76	63	372
ファセット	62	73	66	55	70	326
ユニターム	85	83	73	87	85	413

第8表（原報告 3.10 表）
システム別，索引作成者別の成功率表

検索システム	Hadlow	War-burton	Sharp	[計]
U D C	81	77	83	251
a b c 順	76	70	78	224
ファセット	63	69	64	196
ユニターム	78	85	82	245
[計]	298	301	307	

但し［計］欄は縦横ともに筆者加算[12]

この表と，第4表とからみて，誰の作成した索引（検索システム）が使いやすいか，といったことを推論することは出来ない。

Cleverdon の分析で，興味の持てるのは，質問と論文題名との一致度を11段階にわけ，あまりにも検出に容易なものまで加えて，成功率を算定するのは妥当であるかを反省・考慮している点といえよう。このスカラ（物指し）を作ることは決して容易ではないが，相関度0から10迄を設定して兎に角この問題にタックルしてゆこうとしている態度は買ってよい。[13]

600 探索について，検索成功率を調べてゆくと次のようになる。

第9表 (原報告 4.2 表)

関連度段階	U D C	abc 順	ファセット	ユニターム
0	48 (48)	45 (45)	52 (52)	44 (44)
1	60 (53)	74 (53)	65 (58)	70 (52)
2	72 (60)	70 (58)	63 (59)	85 (63)
3	69 (64)	73 (63)	74 (62)	79 (65)
4	66 (65)	84 (69)	67 (63)	89 (72)
5	77 (68)	74 (70)	71 (65)	83 (74)
6	72 (67)	79 (72)	68 (65)	91 (77)
7	88 (72)	85 (74)	75 (67)	92 (80)
8	84 (74)	90 (77)	80 (70)	87 (82)
9	81 (76)	89 (79)	78 (72)	90 (84)
10	82 (77)	92 (82)	77 (73)	89 (86)

( ) 内の数字はシステム別で，その段階までの累積検索成功率

関連度の高いものに対しては，機械的に作成されるKWIC 索引でも高い成功率をみるのは当然であるから，あまり容易な質問は消去して計算しなおしている。この消去法に対して Cleverdon の採用したのはしかしながら，関連度段階の高いもの，といった規準ではなくて，システムのいずれでも検索された(成功)質問を無視するという，実際的の規準を採ったのである。このことは，恣意的なやり方であると批判されることにもなるが，基礎実験としてはやむを得ないステップでもあったろう。

第10表 (原報告 4.3 表，4.4 表)
容易な質問を除いた上での検索成功率表
(システム別，全体と索引所要時間別)

検索システム	成功率%	索引時間による区分				
		16分	12分	8分	4分	2分
U D C	54	49	60	54	58	54
a b c 順	58	66	67	58	67	52
ファセット	43	38	52	39	44	46
ユニターム	63	70	69	62	76	45

検索実施は研究計画スタッフ，300 検索。このデータに，筆者の加算を行うと，順位は不変，また索引所要時間が成功率に及ぼす影響では，ファセットでのデータが解釈を困難にしているが，他のシステムを通観すると，2分の場合は最低であるが，1件あたり4分という時間は良いデータを出している。

それからまた，索引に用いた用語(記号)と検索に用いた用語との比較も興味が持てる。用語そのものの比較ではなくて，同種の人であっても索引作業の場合と，検索の場合とではこの程度の共通度，特徴をもつものであるということを示唆してくれるデータである。

第11表 (原報告 4.6 表) 但し [ ] は筆者の追加

略号	システム / 摘要	UDC	abc 順		ファセット	ユニターム
			主標目	副標目		
A	[使用総数]	537	390	176	499	693
B	共通使用	252	141	86	240	347
C	索引のみに	129	131	69	156	235
D	検索のみに	156	118	21	103	111
E	[C-D]	− 27	13	8	53	124
F	[B/A]%	47	36	49	48	50
G	[D/(B+D)]%	38	46	20	30	24
H	[C/(B+C)]%	34	48	45	39	40

この結果について，何を発見出来るかは，まだ言えないが，すくなくともUDC の場合のみ E 項が負数になるということだけは理解出来る。再現性は充分期待し得るテストであると言えよう。

Cleverdon テストの成果の中で，筆者に最も強く訴えるものは，この検索実験した際に，不成功と出てきた場合の分析をよくしている点である。Warburton と Aitchison 夫人とで，本試験 (第3グループ) の中からの 329 文書についておこった，495 例の不成功を，1．質問の段階，2．索引作業，3．検索作業，4．システム自身の欠陥といった具合に区分，細区分している。紹介しておく。[14]

第12表

		UDC	aba 順	ファセット	ユニターム	計
1	a. あまり細部にわたるもの	6	5	5	5	21
	b. あまり一般的すぎる	5	4	8	4	21
	c. 理解しにくい	3	3	3	3	12
	d. 質問の趣旨があいまい	4	6	7	6	23
	e. まちがっている	4	2	3	3	12
	a. 索引が充分でない					
	(i) 索引者個人のあやまり	15	17	15	9	56
	(ii) 索引時間が不足	16	14	19	15	64

2	b. あまりこまかく索引してある	4	4	2	1	11
	c. 間違い					
	(i) 個人的ミス	7	3	5	0	15
	(ii) 時間不足からのミス	5	4	6	1	16
	d. 記入の数が不充分	18	0	1	0	19
	e. 不注意な索引					
	(i) 個人的ミス	22	19	29	18	88
	(ii) 時間不足からのミス	15	5	11	6	37
	f. 引用の標目欠	5	1	1	0	7
	g. 参照欠	1	2	1	1	5
3	a. 理解欠除	6	9	10	7	32
	b. 概念使用上の失敗	3	3	4	0	10
	c. チエーン索引使用失敗	—	—	12	—	12
	d. 体系的に探さなかったため	1	2	5	0	8
	e. まちがった探し方	4	2	3	0	9
	f. 不充分な探し方	3	9	5	3	20
4	a. 同主題にいくつかの場所	2	2	4	0	8
	b. 体系中に場所なし	4	3	5	0	12
	c. 細区分なく，あまり大ザッパすぎる	3	0	0	0	3
	d. 標目用語として不適当	0	1	0	0	1
	e. 同義語	0	0[1]	1[0]	2	3
	f. 特定の概念に組合せが作れない	0	1	0	0	1

この表にあげてあるのは実数であって，各システム毎での母数は，80～160 と変化があるので，比較を容易ならしめるために，各システム毎の百分率で示すと次表の如くになる。( ) 内の数字は参考につけたもので，同様のテストを研究計画に関係しなかった技術スタッフが行ったものから計算した場合のデータである。

第 13 表 (原報告 5.4 表，5.6 表)

失敗の原因	UDC	abc 順	フェアセット	ユニターム	計
質問	13(20)	17(15)	16(16)	25(25)	17(17)
索引	70(53)	56(32)	54(34)	60(53)	60(42)
検索	11(23)	21(49)	24(43)	12(20)	17(37)
システム	6(4)	6(4)	6(7)	3(2)	6(4)

第 14 表 (原報告 5.5 表，5.7 表)
同じテストを別の区分で作表しなおすと

失敗の原因	UDC	abc 順	フェアセット	ユニターム	計
質問	13(20)	17(15)	16(16)	25(25)	17(17)
索引(個人差)	43(40)	36(28)	32(24)	34(33)	36(33)
検索(〃)	11(17)	20(32)	16(33)	12(18)	15(24)
時間不足	23(15)	19(16)	21(14)	26(20)	22(17)
その他	10(8)	8(9)	15(13)	3(4)	10(9)

IV. 本研究計画の反響

この調査以前のもののうち Gull のものに Cleverdon が触れていることは既に述べたが，Thorne[15] のものは実験というよりもむしろ予備調査段階の一考察とみられよう。質問の作成には，各システム毎に，それにマッチしたものを準備しているが，この程度のものは，机上での思索実験と言うべきであろう。

情報検索システムの能率だとか効果についての諸データを得ることは，システム設計の発達，システム操作の改善に必須なものであることの認識は相当早くからあったにもかかわらず，主観的判断に基づく意見のやりとりが多く，袋小路に，あるいはまた迷路に，入りこんでしまった観があった。「システムをどう改善してゆくか」「そのシステムがどんな条件下において，効果を発揮し得るか」，をきめてくれるような測定基準といったものを求めての実験といったものは，残念ながらされていなかったといえよう。勿論諸システムの試験とか評価を論じた文献は沢山あったのであるが，Rees も言っている如く“声高らかにしゃべる人は多いが，労働 (labor) した人はすくない。”[16] といった状況であった。そうした中で，この Cleverdon による AslibCranfield 研究計画は出色のものである。勿論この報告に対する O'Conner, Fairthorne, Kyle, Rees, Richmond, Sharp, Stevens, Swanson, Taube 等による批判もあり，それらに対する Cleverdon による反論もあり，いずれも傾聴に値するものを持っている。しかし個々について詳論してゆく余裕はないので，ここでは Richmond[17] の論文だけを紹介しておく。

Richmond の論文

図書資料の情報内容を識別 (identify) し貯蔵しておき，後日必要に応じ探索・検索してゆくシステムを評価してゆくということは，他の検索システムとの比較

をすることで行われるのが普通である。その場合，充分に慎重な態度をとり，いくつかの因子 (variables) を動かしてゆく操作を行えば，さまざまな欠点も浮び出てきて，その欠点を消去してゆく手がかりも可能な筈である。従来の実験にはそのような態度がなかった。今後はモット慎重に，比較実験を行なって欲しい。

上記の，Richmond の主張に異議をはさむものは誰もいないであろう。問題は，その次にくる，“モット慎重に”の中味である。それによって，Richmond の現状認識，他人の論文を読みとる能力の程度も判断されるということである。いたずらに無いものねだりをすることで，図書館学といった新らしく育てられてゆく実学の進歩に寄与できるものなのか，厳しい批判こそ新らしい研究計画に刺激を与えるものなり，といった消極的，積極的両方の見解も出てくる。

比較試験を行うに際しては，事前にそれらのシステムが比較の対象になり得るものかどうかをハッキリ見きわめておくこと。比較は，いかなる条件や判断規準のもとで行なわれてゆくかも明確にしておくこと。[中略] abc 順件名目録，正規の分類目録，簡易略式の分類目録，abc 順分類目録，その他各種頻度統計とか言語学に基づいて作られた検索システム，いずれをとっても，それらに共通した公約数は見出されよう。これらはいずれも，情報を内蔵している資料についての記録を組織しており，情報検索に役だてられるものであり，大ていの場合，ある種の分類的性格[18]を持っているということである。時によると，主題範囲も共通であるということもおころう。しかしそこまでで類似点はおわり，システムの使い方が異っているとか，目標がさまざまであるといった場合に，同じようなテストを実施してその結果に基づいて評価をしてよいものであろうか？

普遍的な分類体系といったものは，記録された知識全分野にわたるもので，その体系の一部分はお互に他の部分にも依存するというようになっている。どの部分にせよそれを特別にとりだして，展開してゆくなり，順序をなおしてみたりしても，必らずその間にいろいろと分岐が出てきてしまい，どうしても全体 [の一部] としてみてゆかなければならないことが感じとられるのである。[19]

ある等質な主題分野にあっては，個々にはあまり厳重に定義されていない用語でも，それを約束できめた順序に組合せてゆきながら，主題内容の核心に近づき，指示してゆくような，同位索引法のようなシステムが良いこともあろうし，主題分野によって，そこで用いられる用語に，同形異義要素が強い場合には，個々に正確な定義が必要なこともあろう。[20]

これらの相異を無視して，比較“研究”といった実験を行うと，どんな結果になるかを若干くわしく考慮してみよう。

Richmond はその典型的な例として，最初の Aslib-Cranfield 研究報告 (1960, 1962) の不備をついている。要点を紹介すると，はじめにも述べた如く一般的体系の一部分を用いた UDC といったもの（稀釈されたアプローチと呼んでいる），特殊主題(この場合は航空力学といった主題) に，テイラーメードで考えられたファセット分類だとか，この分野の文書から直接にとりだした，用語因子(ユニターム)などによる方法（濃縮されたアプローチと呼んでいる）を同じように扱ってテストするのは不当であると言っている。Don Swanson[21] の裏づけをするように，このような比較テストでの，回収の比較では，濃縮タイプに有利になるのは当然であると論じている。

更にまた質問の作成方法自体も，回答を統計的に扱っていることに対しても，前提に無理があるとして，“稀釈された方法も濃縮された方法同様に効果がある，”という結論に疑問をいだいている。即ち，Cleverdon がいろいろと苦労して，主題分析のタイプとか，索引する人の作業を交替させて実施したのは，何とか，この状況下での不規則な回答を平均化してゆくことが出来ないか，と努力したのが，見逃がされ無視されているのである。

・・・その結果は，各文書について4つのシステムを用いて検索した場合の比較とはいえないものになってしまった。Cleverdon がしたのは，文書の四半分について，一つのシステムで分析したことを，他の3つのシステムに次ぎつぎと変換を行ったというテストに過ぎないのである。

結局このテストは，それぞれのシステムがそれぞれ目標とした枠の中で，どのような効果をあげているかを見るためには何の役にもたたないと言える。そしてそのことは，特にファセット分類の場合に言えること

である。ファセット分類による検索システムは，即時参考調査を目標にする他のシステムとは一線を画すべき性質のものである。そのような要素をも含めた質問をも含めて，ストップワッチ式にテストしてゆくのは無理で，悪い成績となってしまうのは当然のことなのである。

Richmondが最後の節で言っていることは筆者には一応納得はいくが，1200乃至1600の質問の中には即時参考的なものもあったことは確かである。しかしそれがどれだけでそんなに大きく響くものであるか，実測データがない限り判断が出来ない。この種の統計的調査というものに，それ程の鋭敏性があるかもテストされねばならない。いわんや次にRichmondが言っている“専門家が用いるように設計された検索システムが，初心者には使いづらいからとて非難される”というのは，ファセット分類を弁護しているつもりかもしれないが，大げさな比喩にすぎると思う。第5表，第10表などからみると，Cleverdonも気づいていたことと思うが，ファセット分類に不慣れということで，説明するのが自然のようである。本論文では紹介しなかったが，原報告3.5表，は大学職員が検索実験した時の成功率を示しているが，それは，最初のグループ，第2グループ，第3グループとシステム別にどの位の慣れ，向上傾向を示しているかの比較からもほぼ裏づけることが出来る。

勿論Richmondが主張している，ceteris paribus(他の条件が等しければ)の重要性は認めるし，更にはまた，Cranfield-Western Reserve University テスト[22]の場合には比較される二つのシステムが理論的にはあまり相違していないので，そこでの比較実験は，従来の比較実験よりは正確・有効であろう，と言っているのは，よい認識であると思う。

また次に述べていることも，図書館人として当然賛成してよいことであるといえよう。

・・・我々は何も唯一のシステムであらゆる人々に，総てのものを与えようという不可能を達成しようとするのではなく，目的によってそれぞれのシステムを使いわけしてゆき，学歴・興味にマッチした回答を準備してゆかねばならない。ある場合には，直接にそのものずばりの回答とか，時には体系を追って順次核心に触れてゆくといった方法がとられよう。しかし大体にいって，散発弾方式の方が，ライフル銃式の照準よりもよく用いられている。

二つの検索システムを，正確な比較規準も与えずに比較してゆくのは，論点を巧みに避けてゆく形である。優劣を競わせるのではなく，こういう点，条件といったことを明確に表示しておいた上で能率のテストをしてゆき，両者をお互に補いあうようにさせてゆくことが必要である。・・・

特に留意しておかねばならぬ点は，利用者はさまざまな要求を持つということ，しかも彼等が究極的には，その検索システムが成功なのか失敗なのか判断をしてくれる当事者なのであるということである。

この点についてはCleverdonも同じ気持で彼の最初の研究計画を設計し，それをふまえて更に次の研究調査を開始したものと筆者は思う。

V. Cleverdon 報告の意義

Cleverdonが最初の実験で狙ったものは第一には，4つの検索システムの性能測定による優劣判定ということであったかもしれない。多くの図書館人は既に予期していたことであるかもしれないが，4つのシステム間には大差は認められないという結論になった。

しかしこの実験を契機として，回収率 (Recall) だとか，適合率 (Relevance) といわれる概念が普及してきて，検索効果の調査に用いられるようになったことは確かである。その時以来数多くの比較評価に関する報告が発表され，その中にはとんでもない，見当違いの引用だとか，説明もみられることは，Swanson[23]も指摘しているところである。

筆者が最も高く買っているのは，どのような理由で失敗がおこったかの調査分析である。第12-14表に明らかであるように，索引語，即ち検索システム自体の責任になると断定される失敗はあまり多くなく，大ていの失敗は，索引の段階での，また検索時のミスからくるという事実である。このことすらも，伝統的図書館人は“予期していた結論である”と称して，この調査の意義を低くみるかもしれない。しかし筆者はこのことがこのように，ハッキリと数字で出されたことをこの上もなく重要と思う。

Sayersの分類規則で，名辞(用語)に関しての個所を，加藤宗厚氏は“分類に使用する名辞は全体を通じて唯一つの意味[に終始一貫して，consistent]に使用されなければならぬ”と訳しておられる。そしてこの言葉は，

筆者が図書館人一年生になった際に教えられたことである。決して容易なことではないがけだし至言ではある。

索引(目録)を作ってゆくこと，その索引(目録)を検索してゆくことは，きわまりのない難かしい技術である。それにひきかえ個々の記入を作成したり，特定の資料を検索するということでいうと，誰でも出来る容易なこともある。また我々図書館人の努力目標も，なるべく容易に出来る場合を多くするということにある。全体のシステム作りとしてみた場合，全利用者に対する奉仕としてみた場合，実にむつかしいことなのである。我々図書館人は，ややもすると，自分等の作業からくる失敗を直視したり，数えあげてみるという実験をしないで，システムそのものに失敗の責任を帰してしまうようなことがある。その意味で Cleverdon の示してくれたこのデータは千鈞の重みがあるといえよう。

測定データの客観性という問題について

既に Rees[24] も指摘したことであるが，Cleverdon をはじめ多数の人々が行っている，実験のその成果は果して，再現性があるのであろうか？ いろいろのサンプルを使って試験している結果は，実際の場合にあてはめてゆけるのであろうか？ 適合 (Relevance) というのは，個人の主観的回答の物指しなのであるが，それぞれのシステムの固有性として考えてよいのであろうか？ etc. etc. これらの問題に対して断定することは現在では不可能であるが，いくつかの関連問題について，諸家の見解に，筆者の意見を加えて紹介しておこう。

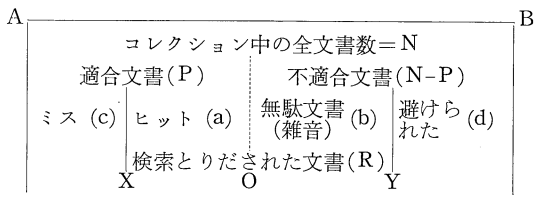
Cleverdon 以来，非常に多くの人々に用いられるようになった概念，回収率 (Recall ratio) というものを知る為には，そのコレクション，蓄積全体に含まれている適合文書の総数というものを知らなければならない。これについては，Zator Company の報告書 (1959) の中で C. N. Mooers が次のように触れている。“質問者自身が，語学能力も時間も充分に持っていて，その図書館の資料全体についてあたることが出来た場合，彼の要求，彼の質問に適合するとピックアップしておく全文書のグループ”ということになっている。この仮定の上にたって推定された，文書の総体，というものは観念的には確かに理解出来ることである。しかしながら，この概念は変転してゆくものであることを承知していなければならない。質問者自身も昨日と今日では違ったセットのグループを選び出すであろう。その彼の選択は彼の同僚の判断とも異なるであろうし，また資料が提供される順序によっても，場所が変っても違うことになろう。

Rees の言っていることを要約紹介しておこう。

> 適合とは，個人の主観的反応を反映するもので，非常に不安定，変化する (dynamic) 現象である。その適合の判定(評価)は，環境だとか心理的要因の複雑な組合せの結果，まちまちに出てくるものである。評価の統一性といったことはあまり高くは望めないのである。システムの効果を判断する規準として，この適合を使用することは充分な検討を要する。[25]

筆者は Rees の見解は，Taube のように“このやり方は利用者の反応としての主観的判断，適合といったものを，システム自体の固有性を特徴づける数値として，すりかえてゆくもので非合理である，”[26] と厳しく批判するのとは一線を画しているものと思う。Cleverdon の実験意図に理解を示して，統計的取りあつかいをすることにより，システム評価の一手段として用いることを承認しているものと思う。すくなくとも，“現在のところ，検索システムを評価してゆくには，この適合にかわるものは存在していない，”[27] と言っているのである。

回収率とか適合率に関しては，Vickery[28] の説明を引用しておこう。



この図で，A から B までの全文書＝N の中で X–Y までの間が，ある質問に対応して検索されたと考えているのである。中央の点線 O の位置はセットされたものとして考えてゆく。先に述べた，Mooers の“理想的”状況は X が左端 A に一致し，Y が O に一致した場合をいうのである。

回収度とは a/P＝a/(a＋c) を

適合度とは a/R＝a/(a＋b) を

意味し，回収率 (Recall ratio)，適合率 (Relevance ratio, Precision ratio) と呼ぶときには百分率で呼ぶことが多い。この他にも ノイズ度 b/R＝b/(a＋b) だとか，細密度 (Specificity) d/(b＋d) を計算する人もいる。また回収度と細密度の和をもって効果度 (Effectiveness) と呼ぶ人もいる。

研究計画と実験の設計

何を比較し，いかなる条件下でテストするかを当初からきめておかねば，報告の段階で混乱がおこり，不要に論争をおこすことは当然であるが，このことに関しては，Rees の論文を引用しておこう。

情報検索システムとは，文書受入れ，索引作業，索引言語，質問分析，検索法，資料配布といった，いくつかの関連しあう要素（これらをサブシステムと呼ぶ人もいる）があつまっている複雑な共同体系であることを考えると，実験の設計が重要となることは，いうまでもない。ここにいう一つ一つの要素をテストする際にもいろいろと，からみあいが出てくるので決して容易でない。たとえば索引作業だけをテストしようにも，質問分析とか検索法との関連を考えずには，試験しようもないのである。もしも各索引者の作業の安定度というものが試験の対象になるのであったなら，検索にあたる人を，ある時は主題専門家に，ある時は司書にしたりして測定してゆくのは賢明でない。その意味で実験計画設計が厳しく検討されていなければならない。

説明を容易にするために，記述子の一組と分類表による二種の索引言語の効果をテストする場合を例として，どのような要素を考慮におくべきか例示してみよう。この際，簡単にする為，利用者層も，質問も，測定規準もすべて存在していると仮定して，両システムの間のそれらの要素の差異が情報検索実施上果して大きな変化を生ずるかどうかを実験的に確かめてゆくことが意図であるとする。

ここで要素間の差異とは次の如きものである。

第15表

要素（variables 変数）	記述子	分類
リンク	有	有
ロール	有	なし
語彙の大きさ（両者で共通の分）	2000用語（1500）	3500用語（1500）
語義上での関連	上下関係と連合	主として上位下位
形式	シソーラス	分類
記号法	不要（国語）	有
範囲の注記	有	なし
展開可能性	開放的［追加可能］	閉鎖的
どこから用語をとってくるか	文献語より	先験的に
索引言語の構成者	主題専門家	司書
事前調整	若干	多量
言外の意味	あまりない	事前調整で多量に
普遍性	限定	一般に通用

ここにあげた，各要素はそれぞれ分離して例えば“シソーラス方式は分類方式よりも有利か，それとも不利か”といった場合に測定·評価してゆくことが出来る。

上述の如くに要素をそれぞれ定めた上で，今度は更に，索引用語はそれぞれの索引システムの中の一要素であるという見地からみての諸関連事項と分離しては測定してゆくわけにゆかないということに気づく。

第16表

	記述子	分類
索引者	主題専門家（博士）	司書
索引深度	20記入	5記入
索引の手がかり	標題	標題＋全文
索引所要時間	30分	5分
索引概念の選出	A, B, C, N, O, P等	A, D, E, H, P
質的調整	有	なし
索引法の指示	印刷された手引	口頭による
訓練期間	3ヶ月	4週間
索引［蓄積］媒体	電算機テープ	カード
索引様式（雛型）	使用	使用せず

上記の表に指摘した事項を一定にするなり換算して行った比較·評価でなければ，その成果は偏向したものになってしまうということは自明である。二つのシステムの効果を比較したつもりの実験でも結局は，索引者のタイプの能力をテストすることになったという例も珍しくない。

我々が慎重に考慮してゆく場合，問題はこれでおわるのではない。個々の索引システムの中での諸要素だとか，検索体系全体での種々の関連事項を明確にしておき，しかも一定に保ち，実験計画を樹立しなければならない。一例を質問分析関係の変数（要素）にとって列挙してみると，分析所要時間，質問者に照会可能なりや否，分析者，探索にかける概念数の標準，用語を拡張してゆく可能性，分析指示，訓練期間，探索にかける概念の追加調節程度等がある。

Recs が結論として言っていることは，我々のこの主題，検索システムの効果測定のための，風調実験室はまだ出来ていないということである。だからといって，実験をしようという意欲に水をさそうというのではなく，丁度，医学の分野を例にとって言えば，壊血病の治療に，ヴィタミン C が特効薬的効果をあげるということは，20世紀になってはじめて知られたことではあるが，治療自体は18世紀から行なわれていたという事実を指摘している。

適正な評価をしてゆくのにはそう便利な近道があるものではない。何か特別によい方法だとか，最上の学派が現在あるのだ，と一辺倒的に信じ込むのは，進歩のブレーキとなるものである。

先程の Vickery の図について卑近な例を示しておこう。X の位置を左寄りにさせるような対策をとった場合，Y は無影響に原位置にとどまっているであろうか？ Y は X 同様左へ移動するのであろうか，それとも右へ移動するのであろうか？

最近の Cranfield 試験では，WRU 索引と，ファセット分類に基づいてのテストを実施する際に，各文書あたりの記入数を平均 12.5，8，5，3 としていった場合の，回収率と適合率とを比較している。

第 17 表

記入数（　）	回収率	適合率
12.5	64.1	34.6
8	60.6	35.5
5	50.7	38.0
3	42.1	41.3

記入の数，索引の深度を増加することにより適合率は 34.6 から 41.3 に増していったが，回収率は 64.1 から 42.1 に減じていったことがわかる。他の要素，（事項・変数）を変えていった場合の影響はまた別であろう。

おわりに

筆者はさきに，日本における KWIC 索引と KWOC 索引の比較調査について指摘し，[29] また *American documentation* 誌 (vol. 15, no. 2, April 1964) 所載の M. J. Ruhl, の "Chemical documents and their titles: Human concept indexing vs. KWIC-mashine indexing" が日本ではどう紹介されているかを指摘した。[30]

他人の労作を批判するばかりでなく，自分でも日本製の索引の比較なり，効果測定を行いたいものと念願している。たとえ被験物はどんなにミニであっても，実験計画設計が宜しきを得れば，相当の知見が得られるものと信じて，Cleverdon の研究計画を中心に，諸家の意見を読んでみた。

Cleverdon の最初の研究計画が，エポックメーキングなものであったことは，諸氏の言を借りて何度か述べたところであるが，それだけにまだ，検討を要する面をのこしてもいる。それらはしかし，後の実験[31-34] で次第に，きめのこまかい推論におきかえられている。検討の余地をのこしていることは，しかしながら，最初から覚悟していたことであって，質問と文書の間にいくつかの不自然な関係のあったことなど承知しつつも，敢てこの実験にふみきったものであることを，注 34 の報告で明かにしている。この論文は，Swanson の批判の一部にこたえているものであるが，回収率と適合率の算定の重要性は今後も続くであろうと強調している。検索システムの蓄積媒体はテープであろうがカードであろうがその物理的形状にはあまり関係しなくて，専ら，概念の索引をするという知的段階が，検索の性能に関係するということを確認している。索引言語が精密 (specific) になればなる程適合率は高くなり，包括的の意味になれば回収率が高くなるのは当然予期されることではあったもしれないが，統計的に確認していっている。それからまた，最近の計画では，適合そのものの判断を 2 分法で割りきるのでなくて，4 段階につけて回答を求めることも実施しだしている。

彼自身は自分の研究報告に対して，実験データを以て批判してくれる人を待ちもうけているという。私自身のミニ被験物での実験もさることながら，比較評価に関心を持たれる日本の図書館員に参考になれば，望外の幸である。

1) Brownson, Helen L. "Evaluation of document searching systems and procedures," *Journal of documentation*, vol. 21, Dec. 1965, p. 261.
2) Metcalfe, John. *Alphabetical subject indication in information.* New Brunswick, N. J. Rutgers University, 1965. 148 p.
3) Richmond, Phyllis A. "Systems evaluation by comparison testing," *College and research libraries*, vol. 27, Jan. 1966, p. 23-30+
4) Cleverdon, Cyril W. "The Aslib-Cranfield research project of the comparative efficiency of indexing systems," *Aslib proceedings*, vol. 12, Dec. 1960, p. 421-31.

5) *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems.* Cranfield, England, The College of Aeronautics, 1962. 306 p.

6) *ibid.*, p. 3.

7) Gull, D. "Seven years of work on the organization of materials in the special library," *American documentation*, vol. 7, Oct. 1956, p. 320-29.

8) Thorne, R. G. "The officiency of subject catalogues and the cost of information searches," *Journal of documentction*, vol. 11, Sept. 1955, p. 130-48.

9) 一般的探索だとか，別に探索するというのは，いかなることかといった問題，原文では general search separate search という具合に，探索という語が用いられているが，ここではこれら4つのシステムに投入された記録から，再びとり出す実験をしているのであるから，検索としておく．別にとは新規に独立の意である．

参考のために指針を紹介しておく．

1. 個々の検索は，それぞれのシステムで考えられる，可能なあらゆる要素の組合せについて実施する．
2. 質問にもりこまれた要素の中から一つまたはそれ以上の要素をとり去ったり，追加した場合は，新規の検索と考える．
3. 完全な検索プログラムというのは，最初に決定した要素の番号よりも一桁すくない番号で行なってもよい．

10) 第二以下の検索はどうして行ってゆくかについて説明しておく必要があろう．

第一の検索が，いくつかの要素をならべたものとすれば，それらを何種かの順列にして検索したのはあくまでも第一の検索として考えるのである．但しその要素を変えたりした場合には第二以下の検索として考えてゆくのである．実例をあげておいた方が理解しやすいかもしれない．

ファセット分類の場合についていえば，

Cd(Ij)　Nr　Oss

で不成功の場合それに更に別の要素をつけ加えてゆくことによって部分的成功，成功にもなり得るのである．

Cd(Ij)　Nbk　Nr　Oss　　部分的成功
Cd(Ij)　Nbj　Nr　Oss　　成功

それぞれの記号の持つ意味を知らないものにはピンとこないかもしれないので，abc順件名目録の場合での例をあげておく．

Wings, Delta—Drag 不成功
Wings, Delta, Supersonic—Drag 部分的成功
Wings, Delta, Transoning—Drag 成功

特定の翼の揚力か何かを求める質問に対しそれだけでは対応する文書が出てこなかった場合に，その翼を「超音速」とか「音速に近い」といった具合に細分化した標目のもとで，揚力関係の文書を求めると，部分的成功なり成功に導びかれるのである．

このような評価法でやると，UDC のように一般的な，(後日 Richmond などが指摘している所謂稀釈的アプローチをとる検索システム) カテゴリー分類などに対し，甘い判定が下されるのではないかといったことは，Cleverdon は気づいたのであろう．検索をどこで打ちきるべきか，あまり沢山の文書を指示しておいて，その中の一つが該当するからといって，成功と認めてよいのかという問題についても触れている．

最初の検索が，ABC という3要素を含むものであるとした場合，それで文書が見あたらない時に，1要素をとおして，AB, BC, AC などについて探してゆくのは可であるが，2要素をとおしてしまい，A, B, C のもとで探すのは不可としている．

またどの程度まで，不適合の文献がまざっていても許容すべきかという点では，文書数 21—40迄は 3，—100 は4，—500 は5，—2000 は 6，—10000 は 7，—20000 は 8 といった一応の標準数をあげている．

11) Cleverdon, *Report on... op. cit.*, p. 119-129.

Harris はシステム別，索引作成所要時間，検索実施者別(研究計画スタッフか他の技術スタッフ)での成功率表（原報告 3.2 表と 3.9 表），システム別，索引作成者別，検索実施者別(研究計画スタッフか他の技術スタッフ)での成功率表（原報告 3.3 表と 3.10 表），システム別，主題別（航空力学とそれ以外），検索実施者別での成功率表（原報告 3.4 表と 3.11 表）を3因子の分散分析にかけている．

以上はい大れも本実験，即ち第三ラウンド 6000 点についての 1200 質問での実験結果に基づいての算定成功率表であるが，又別に，大グループ毎の比較，などもしている成功率表(原報告 3.5 表，3.6 表，3.7表）についたは2因子の分散分析をしている．

Harris は更に Snedecor's F テストなるものを行い，システムの優劣を評価する試みをしている．研究調査スタッフが検索した場合について検討すると，かなり明瞭に，ユニターム，abc 順件名目録がすぐれ，若干水をあけて，UDC，ファセット分類による検索システムという順になっている．しかしこの計画に関与しなかった技術スタッフが検索者になった場合のデータからは，abc 順件名目録と UDC の順位が逆になっている．このデータは一般的にいって，変動が多く，再現性も乏しいのではないかと推察されるので，どうしても研究調査スタッフが検索したときのデータの順位が強調されてしまうことになる．

しかしファセット分類による検索が最下位であるということについては，なお検討してゆかなければならないと，Harris 自身も言っている．このことは，B. Kyle 女史も気づいていたことと関連があるものと思われる．

12) この加算が厳密に意味を持つものとは考えられない．しかし，前注での順と一致するのは興味がある．第3表，第4表の研究計画スタッフの検索実施の際の加算の順も同様である．但し索引作成者毎の加算数に意味を求めるのは無理であろう．

13) ここには，その実例サンプルをあげておく．

質問：Enthalpy boundary layer profiles on a cone at large yaud.

標題：Laminar heat transfer on three dimensional blunt nosed bodies in hypersonic flow.

質問：Transformation theory of the partial differential equation of gas dynamics.

標題：On some solutions of the bodograph equation which yield transonic flows through a Laval nozzle.

質問：What is the average decrease in gas consumption per horse power hours for reciprocating aircraft engines in the course of their history.

標題：The economics of large aircraft.

以上いずれも関連度 0

（4システム 共に検索不成功）

質問：Relation of surface-cooling to boundary layer transition.

標題：Effects of extreme surface cooling on boundary layer transition.

関連度 10

（4システム いずれにても検索成功）

14) Cleverdon, *Report on the testing... op. cit.*, p. 38–50.

15) Thorne, *op. cit.*, p. 130–48.

Royal Aircraft Establishment の図書館にある UDC 分類目録と，ユニターム同位索引法との効率比較をめざしたものである．National Aeronautical Research Institute, Amsterdam が実施したものであるか，そこでの効率算定は，一回一回探索要求があった場合に，各文書一つずつあたってゆき，一つでも該当するものが発見できればそれでよい，即ちあとの探索はしないでよいという仮定のもとでの総経費と，システムを利用しておいた場合の推定費用との比から出そうとしている．

その為には

C 全所蔵資料数

S 単位当り探索費用（一冊一冊手にとり精査してゆく際の平均を推定）

Q 年間探索依頼（照会）数

N 年間資料増加数

の他に，それぞれの検索システムについて，一資料当りに必要な記録費用のP，検索当り費用（目録で探せずという場合も求めての平均単価）Rなどと設定している．この論文の結論として言っていることは，“検索成功率Aを大にすることは重要ではあるが，P やRを無制限に増加させることは出来ない．他の手段とも併用して効果を発揮させることは考えられるが，個々のシステムの実際に即した比較まではまだ出来ない．”といった趣旨である．Thorne が考察（speculation に用いた標準データは，$\eta=\frac{A}{100}-(PN+RQ)/SCQ$ といったものである．

16) Rees, Alan M. “The evaluation of retrieval systems ＜Second Conference on Technical Information Retrieval Administration. 1965 ed. by Arthur W. Elias＞ p. 130.

17) Richmond, P. A. *op. cit.*

18) ある種の分類的性格とは何かに関連したことは，中村初雄，海外における分類法の動向，図書館学会年報（14巻1号，1967, p. 37）に触れている．

19) Richmond のこの見解は，米国議会図書館十進分類課の Henshaw 夫人に訓練された結果で，図書館での作業を身をもって体験した人は誰でもが持つ見解である．一般的分類表の一部分を基礎にして独立した分類体系を作成する為に，他の部分や新規の概念を追加して使用しだしてみても，おそかれ早かれ，他の部分との関連において考えてゆかねばならぬことに気づくのである．

20) Richmond は，Mooers が最初に用いた時の“記述子”といった用語などは，限定された意味で用いられていたものであるが，今日では，類用語“索引用語”といった意味に使われるようになったことを指摘している．また狭隘分野に於ける検索システムでの基本条件ともいうべき簡素さということが，維持し続けられなくなってきている事実を指摘しているが，日本の図書館員には次の論点を理解してゆくのに，そのような道具だては不要であろう．

21) Swanson, Don R. “The evidence underlying the Cranfield results,” *Library quarterly*, vol. 35, Jan. 1965. p. 13.

22) Aitchison, Jean B. and Cleverdon, Cyril W. *A report on a test of the Index of Metallurgical Literature of Western Reserve University*. Cranfield, Eng., College of Aeronautics, 1963. 207 p. このテストは，金属工学という特殊分野のために特別に設計されたファセット分類と，意味因子化法（Semantic factoring method）での解答効率の比較をしている．質問は両方のシステムで主題分析された文書に基づいて作成されたものである．

23) Swanson, *op. cit.*, p. 1–20.

24) Rees, *op. cit.*, p. 130 f.

25) *ibid.*, p. 131 f.
26) Taube, Mortimer "The pseudo-mathematics of relevance," *American documentation*, vol. 16, April 1965, p. 69–72.
27) Rees, *op. cit.*, p. 132.
28) Vickery, B. C. *On retrieval system theory*. 2 ed. London, Butterworths, 1965. p. 174 f.
29) 中村初雄, "Panizzi, Jewett, Cutter の目録規則と要語索引の概念," *Library Science*, no. 5, 1967, p. 102, 106.
30) 中村初雄, ソフィスティフィケーションの功罪<間宮不二雄先生喜寿記念図書館学論文集 1968.> p. 204 f.
31) Cleverdon, Cyril W. "The Cleverdon-WRU experiment: Conclusions," <*Information retrieval in action*, Cleveland, 1963> p. 101–107.
32) Cleverdon, Cyril W. and Mills, J. "The testing of indexing language devices." *Aslib proceedings*, vol. 15, April 1963, p. 106–130.
33) Cleverdon, Cyril W. "The Cranfield hypotheses," *Library quarterly*, vol. 35, April 1965, p. 121–124.
34) Cleverdon, Cyril W. *Factors determining the performance of indexing systems*. Vol. 1, Design. Part 1, Text; Part 2, Appendices. 378 p.