Appendix 1 : List of reference works consulted

British Library Research and Development Department, Inventory of Bibliographic Data Bases Produced in the U.K., BLR&DD Report No. 5256, British Library, London, 1976.

Hall, J.E. On-Line Information Retrieval 1965-1976: A Bibliography with a Guide to On-Line Data Bases and Systems, Aslib Bibliography No. 8, Aslib, London, 1977.

Leigh, J.A. Guide to Computer-Based Literature Searching Services in Science and Technology available in the U.K., Science Reference Library, British Library, London, 1976.

Thomas, A. (Ed.) London University Central Information Services (LUCIS) Guide to Computer-Based Information Services, 2nd Ed., Central Information Services, University of London, 1977.

Tomberg, A. (Ed.) Data Bases in Europe: A Directory to Machine-Readable Data Bases and Data Banks in Europe, 2nd Ed. Aslib & Eusidic, London, 1976.

Williams, M.E. and Rouse, S.H. Computer Readable Bibliographic Data Bases: A Directory and Data Source, ASIS, Washington D.C., 1976.

Appendix 2 : Sample entry from Williams and Rouse' Data Base Directory

STI

1. BASIC INFORMATION

NAME OF DATA BASE
ACRONYM/SHORT NAME: STI
FULL NAME: Specialized Textile Information Service

FREQUENCY OF UPDATE: bimonthly

NUMBER OF TAPES ISSUED PER YEAR: 24

TIME SPAN COVERED BY DATA BASE: 01/70 to present

CORRESPONDENCE WITH PRINTED SOURCE:
1: World Textile Abstracts

FEWER REFERENCES ON TAPE THAN PRINTED SOURCE: yes

2. PRODUCER/DISTRIBUTOR/GENERATOR INFORMATION (See Introduction section 4.2)

PRODUCER OF DATA BASE

NAME: Shirley Institute
Manchester M20 8RX England
PERSON TO CONTACT RE. INFORMATION ABOUT TAPES: Mr. R. J. E. Cumberbirch
(NOTE: Four research institutions collaborate in covering the literature for STI: British Launderer's Research Association (covers all aspects of laundering and dry cleaning); Hatra (covers all aspects of knitting and making-up); Shirley Institute (covers all fibres other than wool and hair, and their properties and processing other than knitting, including lacemaking, knotting and braiding, and bonding, needling and tufting); Wira (covers wool and hair and their properties and processing other than knitting))

DISTRIBUTOR OF DATA BASE

NAME: Shirley Institute
Manchester M20 8RX England
PERSON TO CONTACT RE. DISTRIBUTION OF TAPES: Mr. R. J. E. Cumberbirch

GENERATOR OF (PHYSICAL) DATA BASE

NAME: Shirley Institute
Manchester M20 8RX England
PERSON TO CONTACT RE. TAPE FORMAT, SOFTWARE DATA: Dr. K. C. Ellis

3. AVAILABILITY AND CHARGES FOR DATA BASE TAPES

CURRENT FILES: 1975, 24 bimonthly issues
RESTRICTIONS: ownership of the data base remains vested in the STI Service at the Shirley Institute
LEASE: $1100.00 base fee plus $260.00 for cost of tapes and air mail postage.

BACK FILES: 1970-1974, annual issues
RESTRICTIONS: ownership of the data base remains vested in the STI Service at the Shirley Institute
LEASE: $1100.00 base fee per annual issue, plus $250.00 for cost of tapes and air mail postage

SAMPLE TAPES: no charge to bonafide potential subscriber

4. SUBJECT MATTER AND SCOPE OF DATA ON TAPE

SUBJECT MATTER AND SCOPE: Covers the literature of permanent technical value on the science and technology of textiles plus all relevant UK and US patent literature.

SUBJECT CATEGORY: Chemistry and Chemical Engineering; Patents; Textiles;

STI (cont'd.)

TARGET USER COMMUNITY: Research and industry

ANTICIPATED GROWTH RATE (AVG. NO. OF SOURCE ITEMS ADDED PER YEAR): 8,000

BIBLIOGRAPHIC DATA BASE SOURCE ITEMS CAN BE APPROXIMATED AS:

40% Journal articles Of these, 50% are published in English
No. journals from which selected articles are entered: 500
0% Government reports/documents
40% Patents Of these, 50% are U.S.A. patents
0% Monographs, published proceedings, theses, etc.
0% Preprints, papers presented at conferences
0% Manufacturers' catalogs
0% News items from releases, press reports, broadcasts, etc.
20% Other: Manufacturer's technical publications, government reports/documents; preprints; monographs, published proceedings, theses, etc.
100% Total

5. SUBJECT ANALYSIS/INDEXING DATA

Controlled keywords (from thesaurus). Avg. no. terms/document: 10

Chemical identifiers (nomenclature codes, notations, fragmentation schemes):
Trade name(s)

6. BIBLIOGRAPHIC DATA BASE ELEMENTS PRESENT ON TAPE

Author(s)
Author address
Editor(s)
Editor address
Corporate author(s)
Corporate author address
Title of item(original lang., translat., translit.)
Title of source item(journal, conf. proc.)
Bibliographic reference (volume,issue)
Page(s), inclusive or total
Date(publication date of item, dates for patents)
Publisher
Place of publication
Cited references by source item: total no.
Patent information
(NOTE: The reference given for patents consists of (1) the patent number, (2) the publication date, and (3) the application date and number in the country issuing the patent, or if a prior date of application (the convention date) and the name of the country and the number.)
Language (of item)
Indication of type of item(e.g. jnl. art., mono., govt. doc., etc.)
Treatment code or level of approach(e.g. review, app'n., theory, etc.)
Item accession number, unique id

7. TAPE SPECIFICATIONS

CODE: ECD
CHARACTER SET: upper and lower case
DENSITY (BPI): 556
NUMBER OF TRACKS: 7; 9
LABELS: not present
RECORD FORMAT: fixed RECORD FORMAT: blocked
NUMBER BYTES/BLOCK: 4,096 or 16,384 NUMBER BITS/BYTE: 6

8. SEARCH PROGRAMS

9. DATA BASE SERVICES OFFERED (Brokers not listed. See Introduction section 4.9)

DOCUMENT DELIVERY, REPROGRAPHIC SERVICES AVAILABLE FROM: producer;

TRANSLATION SERVICES AVAILABLE FROM: producer;

10. USER AIDS OFFERED BY DATA BASE PRODUCER

VOCABULARY/TERM LIST, THESAURUS:

STI Keyterm List. An approved list of keyterms that shows the relationship of each term to other keyterms; AVAILABLE IN: hardcopy; PRICE: available free of charge to data base subscribers; non-subscribers $17.00 for both keyterm lists.

Advisory Lists of Related Keyterms. AVAILABLE IN: hardcopy; PRICE: available free of charge to data base subscribers; non-subscribers $17.00 for both keyterm lists

DATA BASE TAPE DOCUMENTATION:

World Textile Abstracts Service and Specialized Textile Information Service, Manual for Abstracts, January 1975. Describes the coverage, subject indexing production of tapes and data base format and data elements; AVAILABLE IN: hardcopy; PRICE: available on request

Appendix 3 : Example of data base questionnaire as sent out

Williams Code	SECTION 1 NATURE OF DATA BASE	
	BASIC INFORMATION	
010.0	Name of data base	Materials
030.0	Frequency of update	Biweekly
040.0	Time span covered	Jan '75 to present
045.0	First available in machine-readable form	Jan '75
060.1	If subset data base, name of parent	
075.0	Related machine-readable files	None
080.0, 085.1	Corresponding printed compilation	
090.1-090.3	Same/fewer/more references on tape than compilation	
	PRODUCER ETC. INFORMATION	
110.0	Producer organisation	Chemical Abstracts Service
110.1-110.5	Producer address	The Ohio State University, Columbus OH 43210
110.6	Person to contact	Marketing Department
130.0	Distributor organisation, in U.K.	United Kingdom Chemical Information Service
130.1-130.5	Distributor address	The University, University Park, Nottingham.
130.6	Person to contact	Dr. A. Kabi
150.0	Generator (of physical data base) organisation	United Kingdom Chemical Information Service
150.1-150.5	Generator address	The University, University Park, Nottingham.
150.6	Person to contact	Dr. A. Kabi
	SUBJECT, SCOPE INFORMATION	
310.0	Subject matter and scope	Chemical and chemical engineering aspects of the production, properties and applications of industrially important materials.
320.0	Subject category	Chemistry & Chemical Engineering; Mining; Metallurgy.
340.0	Approx. number source items by December 1976	
350.0	Average number items added per year	

360.1	Percent journal articles	55
360.11	Percent of these in English	57
360.12	Number of journals from which all articles taken	
360.13	Number of journals from which some articles taken	
360.14	Approx. number of journal titles reviewed for input	14,000
360.2	Percent government reports, documents	2
360.3	Percent patents	35
360.31	Percent of these which are U.K.	
360.4	Percent of monographs, theses, conference proceedings. etc.	8
360.5	Percent preprints, conference papers, etc	0
-	Percent non-government reports, documents	0
360.6	Percent manufacturers catalogues	0
360.7	Percent news items, etc.	0
360.8	Percent other	0
360.81	Description of other	
360.9	Percent total (100%)	100
-	Percent material not in English	
-	How much per item translated to English	
	INDEXING INFORMATION	
410.0	No special indexing	
415.0	Enriched titles	Patents only,
415.1	Average number added terms per title	
420.0	Uncontrolled (natural language) keywords	Yes
420.1	Average number of keywords per document	2 phrases
-	May these be word strings or only single words	Phrases of approx. 4 words

425.0	Controlled (thesaurus) terms	
	Thesaurus name	
425.1	Average number of terms per document	
430.0,	Subject headings	Yes
-	Subject heading system name	
430.1	Average number of headings per document	
435.0	Subject codes	Yes
435.1	Subject code system name	
435.2	Average number of codes per document	
-	Descriptive phrase or sentence	
-	Any other indexing	
-	Indexing source	
(461.0-480.0)	Are chemical identifiers used	Yes
	Are these in a specified record field	
	Average number per document	
	Percent data base having them	
	BIBLIOGRAPHIC INFORMATION	
505.0	No bibliographic information	
510.0	Author(s)	Yes
511.0	Author address	Yes
512.0	Editor(s)	Yes
513.0	Editor address	Yes
514.0	Corporate author(s)	Yes
515.0	Corporate author address	Yes
520.0	Title of item (indicated as original, translation, transliteration)	Original, translation, transliteration
525.0	Title of source	Yes
530.0	Bibliographic reference (volume, issue)	Yes

531.0	Pages, specified or total	
532.0	Publication date	Yes
535.0	Publisher	Yes
536.0	Place of publication	Yes
540.0, 541.0	References cited by source, in total or details	
545.0-548.1	Standard bibliographic codes, CODEN, ISSN/ISBN, other	CODEN
550.0	Abstract	Yes
-	Short digest	
555.0	Patent information	Yes
560.0	Report number	
565.0	Language	Yes
570.0	Indication of type of item (e.g. article, monograph, etc.)	Yes
575.0	Treatment code or level of approach	
580.0	Item accession or other unique identifying number	XXX Yes
585.0	Price	Yes

continued

SECTION 2 USE OF DATA BASE

If you run a search service on your data base, please complete Section 2. If you only supply the data to search services run by others, please complete Section 3. (If you both run your own service and supply others, please complete both sections.)

KSJ Code		
1010.0	Data base only searchable via abstract journal, printed index, etc.	
1020.0	Retrospective off-line searching available	
1025.0	SDI searching available	
1025.1	Time period for SDI	
1030.0	On-line searching available	
1030.1	All or part of data base available on-line	
1040.0	Approx. number of searches per month, altogether	
1040.1	Approx. number off-line searches	
1040.2	Approx. number SDI searches	
1040.3	Approx. number on-line searches	
1050.0	Approx. number subscribers represented, altogther	
1050.1	Approx. number individual users represented, altogether	
1050.2	Approx. number off-line users	
1050.3	Approx. number SDI users	
1050.4	Approx. number on-line users	
1060.0	Indexing fields available for searching	
1070.0	Bibliographic fields available for searching	
1080.0	Searching by Boolean logic	
1080.1	Searching by simple coordination	
1080.2	Searching with term weights	
1080.3	Arbitrary term truncation	
1080.4	Other search methods	
1080.5	Is search formulation and searching by user or intermediary	
1090.0	Person to contact about search service	

SECTION 3 SUPPLY OF DATA BASE

1110.0 UK search services to whom data base supplied (name, address, person to contact)

1120.0 Is data base available on Lockheed's DIALOG system

Signed

Date

Appendix 4 : List of CA and CAB subbases

CA subbases

United Kingdom Chemical Information Service

CACon : CA CONDENSATES

CBAC : CHEMICAL-BIOLOGICAL ACTIVITIES

CIN : CHEMICAL INDUSTRY NOTES

CT : CHEMICAL TITLES

ECOLOGY AND ENVIRONMENT

ENERGY

FOOD AND AGRICULTURAL CHEMISTRY

MATERIALS

POST : POLYMER SCIENCE AND TECHNOLOGY

CAB subbases

Commonwealth Agricultural Bureaux

Animal Breeding Abstracts

Apicultural Abstracts

Dairy Science Abstracts

Field Crop Abstracts

Forestry Abstracts

Helminthological Abstracts

Herbage Abstracts

Horticultural Abstracts

Index Veterinarius

Nutrition Abstracts and Reviews

Plant Breeding Abstracts

Review of Applied Entomology

Review of Medical and Veterinary Mycology

Review of Plant Pathology

Soils and Fertilisers

Veterinary Bulletin

Weed Abstracts

World Agricultural Economics and Rural Sociology Abstracts

Appendix 5 : Tabulated data base questionnaire replies

Williams code	SECTION 1 NATURE OF DATA BASE	CA CAConn	CA CBAC	CIN	CT	ECOL. & ENV.	ENERGY	FOOD & AGR.	MATERIALS	POST
0.0	BASIC INFORMATION Name of data base									
0.0	Frequency of update	week	biweek	week	biweek	biweek	biweek	biweek	biweek	biweek
0.0	Time span covered	68-	65-	74-	62-	75-	75-	75-	75-	67-
5.0	First available in machine-readable form	68	65	74	62	75	75	75	75	67
0.1	If subset data base, name of parent									
5.0	Related machine-readable files	–		–	–					–
0.0, 5.1	Corresponding printed compilation	CA		CIN	CT					CAS 3546
0.1-0.3	Same/fewer/more references on tape than compilation	same		same	same					
	PRODUCER ETC. INFORMATION									
0.0	Producer organisation	CAS	CAS	CAS	CAS	CAS	CAS	CAS	CAS	CAS
0.1-0.5	Producer address									
0.6	Person to contact									
0.0	Distributor organisation, in U.K.	UKCIS	UKCIS	UKCIS	UKCIS	UKCIS	UKCIS	UKCIS	UKCIS	UKCIS
0.1-0.5	Distributor address									
0.6	Person to contact									
0.0	Generator (of physical data base) organisation									
0.1-0.5	Generator address									
0.6	Person to contact									
	SUBJECT, SCOPE INFORMATION									
0.0	Subject matter and scope	chemistry & chem. engineering	chem. & biol. interactions	chem. trade & industry	chem. & chem. engineering	chem. re ecology & envir.	chem. re energy	chem. re food, agric. chem	chem. & chem. eng re materials	macromolecular chem
0.0	Subject category	chemistry, environment, life sciences	chem., envir., life sciences	chem. engineering, business	..	envir., chem. & chem. eng.	chem. & chem. eng., energy	chem., agric., food.	chem., eng., mining, metall.	chem. & chem. eng
0.0	Approx. number source items by December 1976	2.6M	275K		1.8M					300k
0.0	Average number items added per year	470K	47K	50K	150K					45k

		CA Con	CBAC	CIN	CT	ELEC & ENV	ENERGY	FUND & MGT	MATERIAL	POST
360.1	Percent journal articles	72	85	(100)	100	75	73	74	55	48
360.11	Percent of these in English	57	57		63	57	57	57	57	57
360.12	Number of journals from which all articles taken	260			260					
360.13	Number of journals from which some articles taken	7.2K		82	440					
360.14	Approx. number of journal titles reviewed for input	14K	14K	82	700	14K	14K	14K	14K	14K
360.2	Percent government reports, documents	2	3	(0)	0	4	4	1	2	1
360.3	Percent patents	16	4	0	0	10	13	15	35	50
360.31	Percent of these which are U.K.			10						
360.4	Percent of monographs, theses, conference proceedings. etc.	10	8	0	0	11	10	10	8	1
360.5	Percent preprints, conference papers, etc	0		0	0	0	0	0	0	0
-	Percent non-government reports, documents	0		(0)	0	0	0	0	0	0
360.6	Percent manufacturers catalogues	0		0	0	0	0	0	0	0
360.7	Percent news items, etc.	0		(0)	0	0	0	0	0	0
360.8	Percent other	0		0	0	0	0	0	0	0
360.81	Description of other									
360.9	Percent total (100%)		100	100		100	100	100	100	100
-	Percent material not in English									
-	How much per item translated to English									
	INDEXING INFORMATION									
410.0	No special indexing									
415.0	Enriched titles								Patents	patents
415.1	Average number added terms per title									
420.0	Uncontrolled (natural language) keywords	yes	yes	yes		yes	yes	yes	yes	yes
420.1	Average number of keywords per document	8		3		2.3	2.3	2.3	2	2.3
-	May these be word strings or only single words	strings	strings	strings		strings	strings	strings	strings	strings

		CA Calcon	CBAC	CIN	CT	ECOL & ENV.	ENERGY	FOOD & AGR.	MATERIALS	POST
25.0	Controlled (thesaurus) terms							yes		
-	Thesaurus name									
25.1	Average number of terms per document									
50.0,	Subject headings	yes					yes		yes	
-	Subject heading system name									
50.1	Average number of headings per document									
55.0	Subject codes	yes	yes			yes		yes	yes	yes
55.1	Subject code system name									
55.2	Average number of codes per document					1	1	1		
-	Descriptive phrase or sentence		yes							
-	Any other indexing									
-	Indexing source									
61.0-80.0)	Are chemical identifiers used		yes			yes	yes	yes	yes	yes
	Are these in a specified record field									
	Average number per document									
	Percent data base having them									
	BIBLIOGRAPHIC INFORMATION									
05.0	No bibliographic information									
10.0	Author(s)	yes	yes		yes	yes	yes	yes	yes	yes
11.0	Author address	yes	yes			yes	yes	yes	yes	yes
12.0	Editor(s)	yes	yes			yes	yes	yes	yes	yes
13.0	Editor address	yes	yes			yes	yes	yes	yes	yes
14.0	Corporate author(s)	yes	yes			yes	yes	yes	yes	yes
15.0	Corporate author address	yes	yes			yes	yes	yes	yes	yes
20.0	Title of item (indicated as original, translation, transliteration)	orig. transl. transliter.	orig. transl. transliter.		orig. transl. transliter.	orig. transl. transliter.	orig. transl. transliter.	orig. transl. transliter.	orig. transl. transliter.	orig. transl. transliter.
25.0	Title of source	yes	yes	yes		yes	yes	yes	yes	yes
30.0	Bibliographic reference (volume, issue)	yes	yes		yes	yes	yes	yes	yes	yes

		CACon	CA CBAC	CIN	CT	ENERGY	ENVIRONE	FOOD & AGR	MATERIAL	POST
531.0	Pages, specified or total									
532.0	Publication date	yes	yes	yes	yes	yes	yes	yes	yes	yes
535.0	Publisher	yes	yes			yes	yes	yes	yes	yes
536.0	Place of publication	yes	yes			yes	yes	yes	yes	yes
540.0, 541.0	References cited by source, in total or details									
545.0-548.1	Standard bibliographic codes, CODEN, ISSN/ISBN, other	CODEN	CODEN	CODEN	CODEN	CODEN	CODEN	CODEN	CODEN	CODEN
550.0	Abstract		yes	yes		yes	yes	yes	yes	yes
-	Short digest									
555.0	Patent information		yes			yes	yes	yes	yes	yes
560.0	Report number	yes								yes
565.0	Language	yes	yes			yes	yes	yes	yes	yes
570.0	Indication of type of item (e.g. article, monograph, etc.)	yes	yes			yes	yes	yes	yes	yes
575.0	Treatment code or level of approach									
580.0	Item accession or other unique identifying number	yes	yes	yes	yes	yes	yes	yes	yes	yes
585.0	Price	yes	yes			yes	yes	yes	yes	yes

continued

SECTION 2 USE OF DATA BASE

If you run a search service on your data base, please complete Section 2. If you only supply the data to search services run by others, please complete Section 3. (If you both run your own service and supply others, please complete both sections.)

J Code		CA CACon	CBAC	CIN	CT	ECOL. & ENV.	ENERGY	FOOD & AGR.	MATERIALS	POST
10.0	Data base only searchable via abstract journal, printed index, etc.									
20.0	Retrospective off-line searching available									
25.0	SDI searching available	yes								
25.1	Time period for SDI	2 weeks								
30.0	On-line searching available	yes								
30.1	All or part of data base available on-line	all								
40.0	Approx. number of searches per month, altogether	350								
40.1	Approx. number off-line searches									
40.2	Approx. number SDI searches	350								
40.3	Approx. number on-line searches	10								
50.0	Approx. number subscribers represented, altogther	200								
50.1	Approx. number individual users represented, altogether									
50.2	Approx. number off-line users									
50.3	Approx. number SDI users									
50.4	Approx. number on-line users	10 month								
60.0	Indexing fields available for searching	keyword, section								
70.0	Bibliographic fields available for searching	title, author etc.								
80.0	Searching by Boolean logic	yes								
80.1	Searching by simple coordination									
80.2	Searching with term weights	display								
80.3	Arbitrary term truncation	yes								
80.4	Other search methods									
80.5	Is search formulation and searching by user or intermediary	mostly inter.								
90.0	Person to contact about search service									

SECTION 3 SUPPLY OF DATA BASE

		CA-Con	CA	CIN	CT	Ecol. Env.	Energy	Food & Agr.	Materials	Post
1110.0	UK search services to whom data base supplied (name, address, person to contact)	UKCIS								
1120.0	Is data base available on Lockheed's DIALOG system	yes								

Signed

Date

CAB subbases

Williams Code	SECTION 1 NATURE OF DATA BASE	CAB System	ANIMAL BREEDING	DAIRY SCIENCE	HORTICULTURE	INDEX VET.	PLANT BREEDING	APICULTURAL
	BASIC INFORMATION							
1.0	Name of data base							
1.0	Frequency of update	month	month	month	month	month	month	quarter
1.0	Time span covered	73-	30-	38-	31-	71-	30-	
5.0	First available in machine-readable form	74	73	73	73	72	73	73
1.1	If subset data base, name of parent		CAB System	CAB System	CAB System	CAB System	CAB System	CAB System
5.0	Related machine-readable files							
1.0, 5.1	Corresponding printed compilation	abstract Journals	AB Abs.	DS Abs.	H Abs	Index Vet.	PB Abs.	Apic. Abs
1.1-1.3	Same/fewer/more references on tape than compilation	more	Same	same	same	Same	same	Same
	PRODUCER ETC. INFORMATION							
1.0	Producer organisation	CAB	CB Animal Breeding	CB Dairy Science	CB Horti-culture	CB Animal Health	CB Plant Breeding	Inter. Bee Research Assoc.
1.1-1.5	Producer address							
1.6	Person to contact							
1.0	Distributor organisation, in U.K.	CAB	CAB	CAB	CAB	CAB	CAB	CAB
1.1-1.5	Distributor address							
1.6	Person to contact							
1.0	Generator (of physical data base) organisation							
1.1-1.5	Generator address							
1.6	Person to contact							
	SUBJECT, SCOPE INFORMATION							
2.0	Subject matter and scope	agri-cultural Sciences	animal breeding	dairying, milk	horti-culture, plantation crops	veteri-nary	genetics, cytology, breeding of crops	bees, honey
2.0	Subject category	agriculture, life sciences, medicine	agric., life sciences	agric., food science life sciences	horti-culture, plants, crops	medicine, life sciences, agriculture	life sciences, agriculture	agriculture, food science, life sciences
2.0	Approx. number source items by December 1976	400K	100K	100K		90K	45K	4K
2.0	Average number items added per year	130K	6.5K	8K	10K	18K	12K	1.25K

		CAB System	CAB subbases: Animal Breed.	Dairy Sci.	Hort.	Index Vet.	Plant Sc.	Apic.
360.1	Percent journal articles		85	80	90	85	80	80
360.11	Percent of these in English	60	50	50	46	75	60	
360.12	Number of journals from which all articles taken		0	0	1	400	0	4
360.13	Number of journals from which some articles taken		2.5K	3K	1.2K	760	2K	1K
360.14	Approx. number of journal titles reviewed for input	8K	1.2K	1.3K		1.2K	2K	120
360.2	Percent government reports, documents		2	1			5	
360.3	Percent patents		0	6	0	0		
360.31	Percent of these which are U.K.			25				
360.4	Percent of monographs, theses, conference proceedings. etc.		5	4	2	some	5	9
360.5	Percent preprints, conference papers, etc		5	5	2	some	5	0
-	Percent non-government reports, documents	40	2	3		some	5	
360.6	Percent manufacturers catalogues	0	0	0		0		0
360.7	Percent news items, etc.	0	1	0		0		0
360.8	Percent other			1	2			10
360.81	Description of other	books		standards	books			misc
360.9	Percent total (100%)		100					
-	Percent material not in English		50	48	54		40	
-	How much per item translated to English	title			title		title etc	title
	INDEXING INFORMATION							
410.0	No special indexing							
415.0	Enriched titles	some				some		
415.1	Average number added terms per title							
420.0	Uncontrolled (natural language) keywords	yes	partly					
420.1	Average number of keywords per document	4.5	12					
-	May these be word strings or only single words	strings	strings					

		CAB System	CAB subbases: ANIMAL BREED.	DAIRY SCI.	HORTI.	INDEX VET.	PLANT BR.	APIC.
5.0	Controlled (thesaurus) terms	Some				77-	yes	yes
	Thesaurus name		in prep			Vet thes.		EASI
5.1	Average number of terms per document	3-4					5	4
0.0,	Subject headings	yes	yes	yes	yes	-77		
-	Subject heading system name	CAB chapter headings		DSA subj. heads		Ver subj. head		
0.1	Average number of headings per document	1-2	1.5	4.5	4			
5.0	Subject codes	yes	yes	yes			yes	
5.1	Subject code system name	CAB subject codes		DSA codes				UDC
5.2	Average number of codes per document	3-4		1.1			1	4
-	Descriptive phrase or sentence							
-	Any other indexing							
-	Indexing source							
51.0-30.0)	Are chemical identifiers used	no	no	no				no
	Are these in a specified record field	yes						
	Average number per document	1-2						
	Percent data base having them	10%						
05.0	BIBLIOGRAPHIC INFORMATION No bibliographic information							
10.0	Author(s)	yes	yes	yes	yes	yes	yes	yes
11.0	Author address	yes	yes	yes	yes	yes	yes	yes
12.0	Editor(s)	yes	yes	yes	yes	yes	yes	yes
13.0	Editor address	yes	yes		yes		yes	yes
14.0	Corporate author(s)	now	yes	yes	yes	yes	yes	yes
15.0	Corporate author address	yes	yes		yes	yes	yes	yes
20.0	Title of item (indicated as original, translation, transliteration)	orig., trl or t	yes	yes	yes	yes	yes	orig. transl.
25.0	Title of source	yes	yes	yes	yes		yes	yes
30.0	Bibliographic reference (volume, issue)	yes	yes	yes	yes	yes	yes	yes

		CABS System	CABS subfiles: Animal Breed.	Dairy Sci.	Horti.	Index Vet.	Plant Bl.		APIC
531.0	Pages, specified or total	yes	yes	yes	yes	yes	yes		yes
532.0	Publication date	yes	yes	yes	yes	yes	yes		yes
535.0	Publisher	yes	yes	yes	yes	yes	yes		yes
536.0	Place of publication	yes	yes	yes	yes	yes	yes		yes
540.0, 541.0	References cited by source, in total or details	total		total?	total		total?		
545.0-548.1	Standard bibliographic codes, CODEN, ISSN/ISBN, other	I/I	I/I	I/I	I/I		yes		I/I
550.0	Abstract	yes	yes	yes	yes	yes	yes		yes
-	Short digest								
555.0	Patent information	yes	yes	yes					yes
560.0	Report number	yes	yes	yes	yes	yes	yes		yes
565.0	Language	yes	yes	yes	yes	yes	yes		yes
570.0	Indication of type of item (e.g. article, monograph, etc.)	yes	yes			yes	yes		some
575.0	Treatment code or level of approach								
580.0	Item accession or other unique identifying number	yes	yes	yes		yes			
585.0	Price	yes		yes	yes	yes	yes		yes

continued

CATS System

CATS subbases

ANIMAL BREED.

DAIRY SCI.

HORTI.

NON VET.

PLANT SC.

APIC

SECTION 2 USE OF DATA BASE

If you run a search service on your data base, please complete Section 2. If you only supply the data to search services run by others, please complete Section 3. (If you both run your own service and supply others, please complete both sections.)

UJ Code								
010.0	Data base only searchable via abstract journal, printed index, etc.			yes				
020.0	Retrospective off-line searching available	yes			yes		yes	
025.0	SDI searching available	yes	yes		yes			
025.1	Time period for SDI	month	month					
030.0	On-line searching available	yes	yes		yes	yes	yes	
030.1	All or part of data base available on-line	all	all			all	all	
040.0	Approx. number of searches per month, altogether							
040.1	Approx. number off-line searches							
040.2	Approx. number SDI searches				3-4 month			
040.3	Approx. number on-line searches				7-4 month			
050.0	Approx. number subscribers represented, altogther							
050.1	Approx. number individual users represented, altogether							
050.2	Approx. number off-line users							
050.3	Approx. number SDI users							
050.4	Approx. number on-line users							
060.0	Indexing fields available for searching	yes	yes		yes	yes	yes	
070.0	Bibliographic fields available for searching	yes	yes		yes	yes	yes	
080.0	Searching by Boolean logic	yes	yes				yes	
080.1	Searching by simple coordination	yes					yes	
080.2	Searching with term weights							
080.3	Arbitrary term truncation	yes	yes				yes	
080.4	Other search methods	yes						
080.5	Is search formulation and searching by user or intermediary	intermed.	both				both	
090.0	Person to contact about search service							

		CABS System	CABS subsystem: ANIMAL NUTRITION	DAIRY SCI.	HORTIC.	INDEX VET.	PLANT BN.	APIC.
	SECTION 3 SUPPLY OF DATA BASE							
1110.0	UK search services to whom data base supplied (name, address, person to contact)	CABS Farnham	CABS Farnham		CABS Farnham			CABS Farnham
1120.0	Is data base available on Lockheed's DIALOG system	yes	yes	yes	yes	yes	yes	yes

Signed

Date

Williams Code	SECTION 1 NATURE OF DATA BASE	INSPEC		STN		SCI
.0	BASIC INFORMATION Name of data base					
.0	Frequency of update	biweekl/ month		bimonth		week
.0	Time span covered	69-		70-		61-
.0	First available in machine-readable form	70				64
.1	If subset data base, name of parent					
.0	Related machine-readable files	INSPEC Phys				
.0, .1	Corresponding printed compilation	Phys Abs. Elec. Abs. Comp Abs.		World Textile Abs.		SCI
.1-.3	Same/fewer/more references on tape than compilation	same		fewer		more
.0	PRODUCER ETC. INFORMATION Producer organisation	IEE		Shirley Institute		ISI
.1-.5	Producer address					
.6	Person to contact					
.0	Distributor organisation, in U.K.	IEE		Shirley Institute		ISI, Uxbridge
.1-.5	Distributor address					
.6	Person to contact					
.0	Generator (of physical data base) organisation					
.1-.5	Generator address					
.6	Person to contact					
.0	SUBJECT, SCOPE INFORMATION Subject matter and scope	physics, mechanical & electro. engineering, computing		science & economics of textiles		science & technology
.0	Subject category	physics, space, energy, engineering, electronics		textiles, chemistry, chem. eng.		science & technology
.0	Approx. number source items by December 1976	1M		60K		6.4M
.0	Average number items added per year	150K		8K		530K

		INSPEC		STI		SCI
360.1	Percent journal articles	80		40		100
360.11	Percent of these in English	70		50		
360.12	Number of journals from which all articles taken	350		10		3.8K
360.13	Number of journals from which some articles taken	2K		500		0
360.14	Approx. number of journal titles reviewed for input	2.3K		350		3.8K
360.2	Percent government reports, documents	7.5		0		0
360.3	Percent patents	6.2		40		0
360.31	Percent of these which are U.K.			50		
360.4	Percent of monographs, theses, conference proceedings, etc.	6.2		5		0
360.5	Percent preprints, conference papers, etc	0		0		0
-	Percent non-government reports, documents	0		.5		0
360.6	Percent manufacturers catalogues	0		0		0
360.7	Percent news items, etc.	0		0		0
360.8	Percent other	0		15		0
360.81	Description of other					
360.9	Percent total (100%)	100		100		100
-	Percent material not in English	30		40		
-	How much per item translated to English					titles
410.0	INDEXING INFORMATION No special indexing					
415.0	Enriched titles	Some				
415.1	Average number added terms per title	1-2				
420.0	Uncontrolled (natural language) keywords	yes		yes		
420.1	Average number of keywords per document	7-8		3		
-	May these be word strings or only single words	strings		strings		

		INSPEC		STI		SCI
5.0	Controlled (thesaurus) terms	yes		yes		
	Thesaurus name	INSPEC thes.		STI keyterms		
5.1	Average number of terms per document	3		10		
0.0,	Subject headings	yes				
-	Subject heading system name	INSPEC ther.				
0.1	Average number of headings per document	3				
5.0	Subject codes	yes				
5.1	Subject code system name	INSPEC classn.				
5.2	Average number of codes per document	2				
-	Descriptive phrase or sentence					
-	Any other indexing					
-	Indexing source					
1.0-0.0)	Are chemical identifiers used	yes		yes		
	Are these in a specified record field	no		no		
	Average number per document					
	Percent data base having them					
	BIBLIOGRAPHIC INFORMATION					
5.0	No bibliographic information					
0.0	Author(s)	yes		yes		yes
1.0	Author address	yes		yes		yes
2.0	Editor(s)	yes		yes		yes
3.0	Editor address	yes		yes		
4.0	Corporate author(s)	yes		yes		yes
5.0	Corporate author address	yes		yes		yes
.0	Title of item (indicated as original, translation, transliteration)	orig. transl. translit.		orig. transl. translit.		orig. transl. translit.
5.0	Title of source	yes		yes		yes
.0	Bibliographic reference (volume, issue)	yes		yes		yes

		INSPEC		STI		SCI
531.0	Pages, specified or total	yes		yes		yes
532.0	Publication date	yes		yes		yes
535.0	Publisher	yes		yes		
536.0	Place of publication	yes		yes		
540.0, 541.0	References cited by source, in total or details	total		total		total details
545.0-548.1	Standard bibliographic codes, CODEN, ISSN/ISBN, other	CODEN, IIE				
550.0	Abstract	yes				
-	Short digest					
555.0	Patent information	yes		yes		
560.0	Report number	yes				
565.0	Ianguage	yes		yes		yes
570.0	Indication of type of item (e.g. article, monograph, etc.)					yes
575.0	Treatment code or level of approach	yes		yes		
580.0	Item accession or other unique identifying number	yes		yes		yes
585.0	Price	yes				

continued

SECTION 2 USE OF DATA BASE

If you run a search service on your data base, please complete Section 2. If you only supply the data to search services run by others, please complete Section 3. (If you both run your own service and supply others, please complete both sections.)

J Code		INSPEC		STI		SCI
10.0	Data base only searchable via abstract journal, printed index, etc.					
20.0	Retrospective off-line searching available					yes
25.0	SDI searching available	yes				yes
25.1	Time period for SDI	week				week
30.0	On-line searching available					yes
30.1	All or part of data base available on-line					part
40.0	Approx. number of searches per month, altogether					
40.1	Approx. number off-line searches					
40.2	Approx. number SDI searches	130 indiv. 400 std.				
40.3	Approx. number on-line searches					
50.0	Approx. number subscribers represented, altogther	130 indiv. 300 std.				
50.1	Approx. number individual users represented, altogether	130 indiv. 300 std.				
50.2	Approx. number off-line users					
50.3	Approx. number SDI users	130 indiv. 300 std.				
50.4	Approx. number on-line users					
60.0	Indexing fields available for searching	yes				yes
70.0	Bibliographic fields available for searching	yes				yes
80.0	Searching by Boolean logic	yes				yes
80.1	Searching by simple coordination					yes
80.2	Searching with term weights					
80.3	Arbitrary term truncation	yes				yes
80.4	Other search methods					citation
80.5	Is search formulation and searching by user or intermediary	both				
90.0	Person to contact about search service					

		INSPEC		STN	SCI
	SECTION 3 SUPPLY OF DATA BASE				
1110.0	UK search services to whom data base supplied (name, address, person to contact)				
1120.0	Is data base available on Lockheed's DIALOG system	yes			yes

Signed

Date

Appendix 6 : CAB subbase sizes

	by 1976/7	1977 increase
Animal Breeding Abstracts	23.3 K	6 K
Dairy Science Abstracts	27.4 K	8 K
Field Crop Abstracts (1)	44 K	11 K
Forestry Abstracts	31.3 K	8 K
Helminthological Abstracts	28 K	8.5 K
Herbage Abstracts with(1)		6 K
Horticultural Abstracts	41.3 K	12.5 K
Index Veterinarius (2)	90 K	18 K
Nutrition Abstracts and Reviews	36 K	10 K
Plant Breeding Abstracts	41.1 K	12 K
Review of Applied Entomology	37 K	12 K
Review of Medical and Veterinary Mycology with(3)		2.5 K
Review of Plant Pathology (3)	28 K	6.5 K
Soils and Fertilisers	22.3 K	8 K
Veterinary Bulletin with(2)		7.5 K
Weed Abstracts	12.9 K	4 K
World Agricultural Economics	25.5 K	8 K

Appendix 7 : Analysis of relevance judgement requirements

This appendix provides the argument for the number and nature of relevance assessments for the 'ideal' collection. This is initially presented in a very elementary form. A summary of the assumptions made, and a tabulation of the numbers of assessments required in different circumstances, follow. Some implications of the approach are then discussed. In the last section an alternative presentation in more conventional statistical language is provided.

A. Elementary presentation

The essential object of our calculations is to ensure that adequate relevance information is collected for the evaluation of future experimental results, in the case where exhaustive relevance assessment is impossible. In the past, test data has either been 'globally' exhaustive in the sense that the entire collection is assessed for the test requests, so that the status of any document retrieved by a new strategy, i.e. indexing or searching device or procedure, is known; or 'locally' exhaustive in that some or all of the output of particular strategies being considered is assessed, so that the performance of these strategies can be compared with respect to the combined assessed output for the strategies.

The problem encountered in considering relevance assessment for the 'ideal' collection is that while global exhaustion is not possible, local exhaustion as conventionally defined cannot be used for future strategies since these may produce output not related in a well-defined way to the initial output for which assessments are provided: i.e. the new output is neither included in the assessed output nor overlapped with it in a coherent way; and if an attempt is made to meet this difficulty of local exhaustion by making the initial searches so broad that their output is likely to be exhaustive of future output, this appears to imply that an unacceptably large number of assessments have to be made.

The question is therefore whether the initial output can be obtained and assessed, at the time when the 'ideal' collection is set up, in such a way that future experimental output can be properly evaluated.

Essentially, our argument is that under suitable conditions, this can be achieved by sampling from the initial output: that is, that in the collection building, we conduct searches for the given requests (i.e. based on the given need statements), probably a variety of alternative searches for each request, and establish a pool of retrieved documents for each request. From this pool a sample is drawn for assessment. This sample constitutes the set of documents of known relevance status which is used to characterise, and more importantly to compare, performance for new strategies.

Our argument has two components: it covers, first, the way in which future experiments are to be conducted, i.e. comparative evaluation is to be carried out; and second, the characteristics of the relevance data needed to support this evaluation methodology.

1. evaluation

The object of a retrieval test, at the lowest level, is taken to be a comparison between two strategies, A and B, representing different choices of indexing, searching, or whatever. As indicated in the Report

text, we will for clarity take these to be two strategies not used to generate the 'ideal' collection itself, though either or both can in principle be generating strategies. To compare the two strategies, we consider only that part of each output that has already been assessed; the remainder is discarded. The relative performance of the two strategies is then represented by their relative success in retrieving assessed relevant documents and rejecting assessed non-relevant ones.

More specifically, the following assumptions are made about the way in which such comparative evaluation is to be conducted. We are concerned with recall and precision,* and these are interpreted as probabilities to be estimated by proportions based on samples. That is, recall is the probability of retrieving a document given that it is relevant, and precision is the probability of a document being relevant given that it is retrieved, where these probabilities for a request and a document collection as a whole may be estimated from the proportions of relevant and non-relevant retrieved by a strategy from a proper sample of the collection which is fully characterised for relevance. To establish a significant difference in performance, over a set of requests, between strategy A and strategy B, we apply the sign test. We base it on the assumption that a percentage difference, say of 5%, between the recall or precision performance of the strategies for a single request is represented by $Prob_A - Prob_B = 5\%$; and over all the requests we look for a particular significance level, say 5% or 1%, and want the test to have a particular power, say 95%. That is, an individual measurement for the application of the test is a single request comparison between strategies A and B, so the set of measurements is the set of comparisons over the complete set of requests. We also assume that the sampling distribution for the performance measurement comparisons being considered, i.e. the differences of proportion representing recall or precision, is normal; and for convenience we assume a normal approximation to the binomial distribution for the power of the test. Finally, the overall assumption is made that the probability of strategy A being superior to strategy B is constant over the request set.

2. data

If we are thus to evaluate performance comparatively, this imposes certain requirements on the assessment data needed. The evaluation cannot begin without assessment information, so the requirements concern the amount and properties of the assessment data exploited in the application of the test. The essential requirement is for a certain number of assessments overall; for practical reasons this can be referred to in terms of the number of requests required and the number of assessments per request, but the two are inversely related so the total of assessments is the same.

Clearly, the fundamental requirement for the whole process is that the relevance status of some of the documents retrieved by strategy A and by strategy B should be assessed. Thus it is not useful to provide asessment data in the initial collection creation by assessing a random sample of the entire collection in relation to the requests. For a large collection in particular this is likely to find no relevant documents at all. On the whole, 'real' search strategies do better than random sampling, so an effective way of seeking to ensure that some of the documents retrieved

* or related performance characterisations

by future strategies A and B have been assessed is to provide the assessment data initially by evaluating actual search output. That is, strategy performance is evaluated by reference to assessed initial search output in order to ensure output overlap, rather than by reference to assessed randomly selected documents. It may further be sufficient to assess a sample of the initial output. However, for this use of initial search output assessments to be valid, the same requirements must apply to the search output, or any sample of it, as apply to the entire collection and any sample of this.

Thus we assume, globally, that the initial output as a whole contains all the documents relevant to a request, and all the output of future searches for the request. Further, we assume that any sample drawn from the initial output is a random sample; and that any such sample is also a random sample of the output of a particular strategy.

Taking the proposed evaluation procedure and data requirements together gives specific percentage samples of the initial output which must be assessed to provide adequate evaluation data for different conditions. In particular, we find that as the number of requests considered decreases, the size of sample increases (up to 100%). This data is tabulated below. Since the comprehensiveness requirements of the initital output are only likely to be satisfied in practice by combining the outputs of several alternative searches for a given request, the output is referred to as the pool.

The table covers different sizes of request set. The results for each set are independent of those for others: the results taken together simply show how for different sizes of set the number of assessments to be made as a percentage of the pool varies. For each request set, the assessment data is given for a sign test significance level of 5% or of 1% for any comparison between strategies A and B. The table then shows the critical region of the test: the number of individual measurements, i.e. request comparisons, favouring one of the strategies (say A) needed for a significant result; the probability that the measurements will favour A over the set required for 95% power in the test; and the sample size required to identify a difference between the two strategies that this implies: the sample size is the number of assessments for each of the strategies that must be provided, i.e. the extent to which the strategy output overlaps the assessed pool output.

The actual formulae used in the numerical calculations are not given here: they are of an orthodox statistical nature.

The second section of the table shows the percentage of the pool to be assessed for recall and for precision respectively, for given numbers of relevant documents per request, on average, and for given numbers of retrieved documents. That is, for a reliable recall comparison between two future strategies A and B for 500 requests, say, with an average of 25 relevant documents per request in the total collection, 36% of the pool would have to be assessed for a 5% significance level in the sign test. For precision and say 100 documents retrieved on average, 9% must be assessed. Note that the percentage to be assessed in any given case is always higher for recall than for precision; and also that for very low numbers of requests and relevant documents, a difference at 5% or at 1% cannot be established. Note also that the figures are approximate, i.e. have not been worked to a very high level of accuracy.

B. Summary and tabulation

For reference the assumptions underlying the table can be summarised as follows:

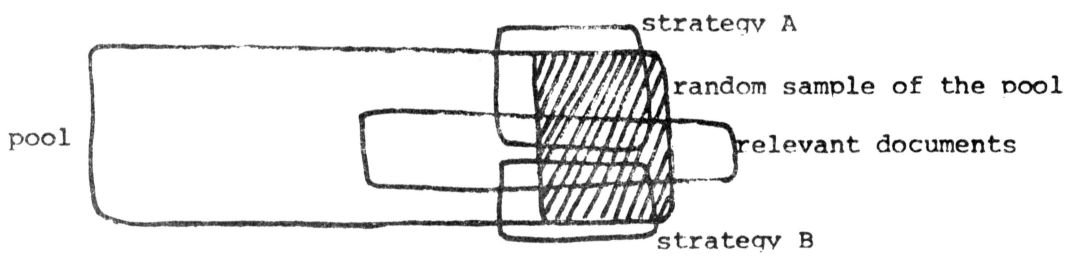
1 for future experiments comparing strategies A and B

1 we evaluate using recall and precision;
2 recall and precision are probabilities estimated by proportions based on samples;
3 we use the sign test for validating performance differences;
4 a percentage difference, say of 5%, between A and B, in recall or precision, is indicated by $Prob_A - Prob_B = 5\%$;
5 a normal sampling distribution for difference of proportions;
6 a normal approximation to the binomial distribution for the power of the sign test;
7 the probability of finding A better than B is constant across requests.

2 for assessment data

1 all relevant documents are contained in the pool;
2 the output of A,and of B, is contained in the pool;
3 a sample from the pool is a random sample;
4 a pool random sample is also a strategy output random sample.

The situation being modelled can be illustrated thus:



no. requests	sig. %	S >	F_A >	for 95% power P >	P_A-P_B =5% N >	RECALL rel = 5 % pool	rel =10 % pool	rel =25 % pool	rel =50 % pool	rel =75 % pool	PRECISION retr 25 % pool	retr 50 % pool	retr 100 % pool	retr 250 % pool	retr 1000 % pool
300	5	2.0	167	.605	15	*	*	60	30	20	60	30	15	6.0	1.5
	1	2.6	173	.624	21	*	*	84	41	28	84	41	21	8.4	2.1
400	5	2.0	220	.592	12	*	*	48	24	16	48	24	12	4.8	1.2
	1	2.6	226	.606	15	*	*	60	30	20	60	30	15	6.0	1.5
500	5	2.0	272	.581	9	*	90	36	18	12	36	18	9	3.6	0.9
	1	2.6	279	.595	13	*	*	52	26	17.3	52	26	13	5.2	1.3
600	5	2.0	324	.574	8	*	80	32	16	10.6	32	16	8	3.2	0.8
	1	2.6	332	.587	10	*	100	40	20	13.3	40	20	10	4.0	1.0
700	5	2.0	376	.569	7	*	70	28	14	9.3	28	14	7	2.8	0.7
	1	2.6	384	.580	9	*	90	36	18	12	36	18	9	3.6	0.9
800	5	2.0	428	.564	6	*	60	24	12	8	24	12	6	2.4	0.6
	1	2.6	437	.576	8	*	80	32	16	10.6	32	16	8	3.2	0.8
900	5	2.0	480	.561	6	*	60	24	12	8	24	12	6	2.4	0.6
	1	2.6	489	.571	7	*	70	28	14	9.3	28	14	7	2.8	0.7
1000	5	2.0	532	.558	5	100	50	20	10	6.6	20	10	5	2.0	0.5
	1	2.6	541	.567	6	*	60	24	12	8	24	12	6	2.4	0.6

Figures all approximate

sig. % = significance level of test
S = critical region of test
F_A = number of measurements favouring strategy A
P = probability that measurement on strategy A will exceed that on B over the request set
N = sample size for hypothesised difference between strategies A and B, i.e. number of documents of known relevance status required in the output of A, and of B.

C. Discussion

There are two obvious limitations in the model:

a) the probabilities of difference are not likely to be constant across requests. However a general form of the central limit theorem might be exploited to modify the model to deal with this;

b) all the relevant documents for a request, and all the retrieved documents for a strategy, are unlikely to be in the pool. But since the pool is only used as a base for comparing two strategies, the uncertainties might be equalisable.

That is, we believe that the type of procedure used to generate the table data could be elaborated to deal with these problems, and hence provide assessment percentages for a greater range of contexts. We emphasise that a short statistical project covering such investigations is desirable. We do not believe it would show the whole approach to be mistaken: it would rather provide fuller information covering more contingencies, and could well also show that satisfactory experiments could be conducted in less stringent conditions than those considered here, without material implications for the cost of provising the assessment data. Indeed a more carefully detailed statistical design could well show that the cost of providing the collection could be reduced.

In this connection one particular practical implication should be noted. Choosing a particular size of request set and assessing for it would apparently imply that in any future experiments all these requests would have to be used: this might well be inconvenient. A question therefore also requiring statistical investigation is the 'tolerance' of given request and assessment data for sampling: i.e. if 700 requests are provided with, say, 28% pool assessment, can this information be used to evaluate performance for a random sample of, say, 300 of the requests? It appears not, since 300 requests in principle require 60% assessment, for the same number of relevant documents per request. It may, however, be the case that a detailed statistical analysis would show that some compromise would be adequate, so that, for instance, the initial data could be provided with 700 requests and (suppose) 45% assessment, which would provide information acceptable for experiments with a sample of (suppose) not less than 300 requests. (A perhaps safer alternative would be to provide, on collection creation, a random sample of the requests with exhaustive pool assessments: but note that the general statistical argument would require that this sample should not be too small.)

Clearly, the practical implications of the most critical assumptions, 2.1 and 2.2, are important, since they affect the search procedure used to generate the pool. In practice, therefore, some idea of sensible pool-generating procedures is needed, which must be buttressed by sampling to see how far 2.1 and 2.2 are met. However, discussions suggest there is no overriding difficulty about providing suitable alternative strategies for this, the only practical consequance being that an 'exhausting' pool is likely to be large, so more assessments are needed. Observation in different investigations in the past suggests that, for example, for 30,000 documents/a pool of size 3000 could be expected to meet 2.1 and 2.2, and for requests with few relevant documents on average, the pool could well be smaller. The practical implications for assessment of this point are discussed in the Report text.

The most important point about the whole argument is that the design is consistently for the worst case. Thus the sign test is a weak test adopted because there may be insufficient knowledge of the collection structure to support the application of a stronger one such as Wilcoxon. However, if the data structure is known, any data to which the sign test applies is in these circumstances also a field for the use of Wilcoxon. A second illustration of this point is that the assumption is that the outputs of strategies A and B are independent: but in practice some relevant documents seem to be more easily retrieved than others, which implies that the outputs are not likely to be independent. However, in this case the power of the test is simply increased, so the proposed design in itself covers this case.

D. Statistical presentation

1. We assume that what we are trying to tablish is that there is a significant difference between two probabilities (or two proportions) based on sample estimates of them. Throughout we use the normal approximation to the binomial, that is

$$N(0,1) \sim \frac{x - np}{\sqrt{np(1 - p)}} \qquad n \rightarrow \infty \qquad (1)$$

where x is the number of successes and p the probability of success.

2. For significance test we choose the sign test because it makes few a priori assumptions about the data. For two strategies A and B we order each request in terms of effectiveness, i.e. effectiveness of Q under A $\gtrless$ effectiveness of Q under B. Effectiveness here is either precision or recall which are assumed to be probabilities. The null hypothesis (H_0) is that there is no difference, i.e. Prob(A $>$ B) = Prob(B $>$ A) = ½ . Since the test is based on the binomial distribution we can use the approximation (1) to find the critical region, that is, that value of the standardised normal variable which needs to be exceeded for H_0 to be rejected at 5% significance level. If k is the number of requests, then under H_0 : p = ½ and we get

$$\frac{x - k/2}{\sqrt{k/4}} = \frac{2x - k}{\sqrt{k}} .$$

Using normal tables (Hoel, 398) we find

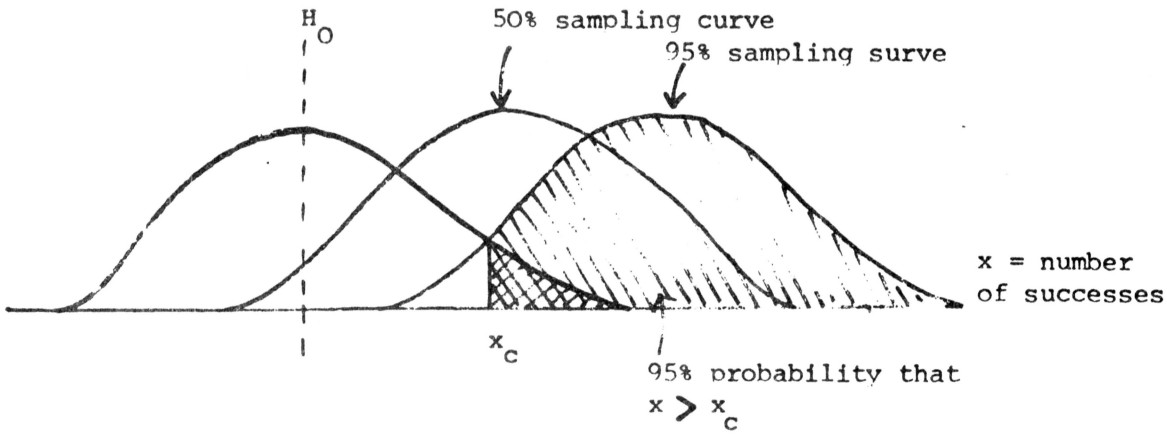
$$\frac{2x - k}{\sqrt{k}} > 2$$

gives 5% significance. This means for k = 100 (requests) we must have at least 60 A's $>$ B's say.

3. The above is all we would need to be concerned with if there were no uncertainties in the probabilities we are comparing, that is, no uncertainty for precision or recall at each request. Unfortunately our decision whether A $>$ B or B $>$ A is based on two samples, one for A and one for B. So that even if there is a real difference between A and B, because we are sampling this difference will fluctuate. Of course were we to take infinite (read, very large) samples we would get the true difference. Assume now that the probabilities we are trying to estimate (recall and precision) are constant across requests; we can then calculate a minimum sample size for each request (it will be the same) necessary for the sign test to show a significant difference. To do this we must assume what the real difference is. Obviously, the bigger the real difference the smaller the sample size necessary to reflect it. There

is a sampling theorem for differences (see Hoel, 149) which again allows us to use the normal approximation to the binomial. The effect of using the theorem is for us the calculation of $P(x_A > x_B)$ for any given n (sample size). Conversely, given the $P(x_A > x_B)$ we can calculate the n necessary to achieve it. Once we have done this the constancy across requests will tell us the expected number of requests with A > B. Conversely, given the number of A's > B's dictated by the sign test and letting it equal the expected number derived above, we can choose a sample size to achieve the expected number. Because we design for an expected number it is reasonable to assume that 50% of the times the number of A's > B's will fall below the critical value and 50% of the times above. But we would like a higher chance of significance, or to put it another way, a higher chance of rejecting the null hypothesis if it is in fact false (i.e. $P_A - P_B = 5\%$ is true). This can only be done by increasing $P(x_A > x_B)$ (or equivalently increasing the sample size). We want to ensure a 95% chance when $P_A - P_B = 5\%$ that the number of A's > B's will exceed the critical value. In other words, for what value $P(x_A > x_B)$ will it be the case that there is a 95% chance of significance This we again get by using the normal approximation to the binomial.

We may illustrate the relationship between the critical region defined by $x > x_c$ and a 50% or 95% power of the test by the following very crude diagram:



If k = 100 $x_c = 60$ $H_0 = (P = \frac{1}{2})$

$H_1' = (P = .60)$ 50% chance of 5% significance

$H_1'' = (P = .68)$ 95% chance of 5% significance .

Comments:

a) Once we have the sample size we can use it to calculate the percent of the pool. The basic idea is that we want a random sample from the future outputs and relevant documents big enough to estimate precision and recall. For this we need assumptions 2.1 and 2.2 of section B above.

b) The table given earlier shows a number of alternatives. One can do with fewer requests by increasing the number of assessments per request.

c) The sign test could be replaced by a stronger test, in which case the design would be somewhat cheaper.

P.G. Hoel, Introduction to Mathematical Statistics, 3rd Ed., Wiley, 1962.

Appendix 8 ; Research project questionnaire

POSSIBLE RESEARCH PROJECT USING THE 'IDEAL' TEST COLLECTION

The 'ideal' retrieval test collection is intended to permit a variety of controlled indexing and retrieval experiments on real material, to encourage inter-project comparisons, and to reduce data preparation effort. It would consist essentially of a large set of basic document descriptions, from which different subsets with particular properties and fuller descriptions could be drawn: of off-line and on-line queries; and of associated relevance judgements. The collection would be set up in a well-organised way, and would be available in machine readable form.

The first specification for the collection is given in K. Sparck Jones and C.J. van Rijsbergen, "Report on the Need for and Provision of an 'Ideal' Information Retrieval Test Collection", 1975; a more detailed one is provided by K. Sparck Jones, "Outline Specification for the 'Ideal' Information Retrieval Test Collection", 1976, both available from K. Sparck Jones.

Project topic

Objective

Methodology

Data requirements

a) content

b) form (machine/manual)

Scale

a) time: 1,2,3, or more years

b) manpower: 1-2, 3-4, 5-6, or more staff

Status

would like to start as soon as material is available (if not, is this because of other commitments, or because project is tentative)

Name

Address

Appendix 9 : Teaching and on-line education questionnaires

INFORMATION RETRIEVAL TEST COLLECTION: USE FOR TEACHING AND RESEARCH IN DEPARTMENTS OF COMPUTING, INFORMATION STUDIES, OR LIBRARIANSHIP

1 a) Topics under the general headings of information or data management, processing or retrieval, of interest to your department:

b) Topics specifically studied in courses:

2 General data requirements, e.g. type and volume of material:

for 1 a) :

for 1 b) :

3 Levels of study, and numbers of students involved, in information processing:

undergraduate, 3 years :

2 years :

1 year :

postgraduate, diploma :

master's degree :

doctor's degree :

Name

Department

Address

THE 'IDEAL' INFORMATION RETRIEVAL TEST COLLECTION :

POSSIBLE USE IN CONNECTION WITH ON-LINE EDUCATION

1 Do you, or are you intending to, teach on-line searching?

2 If so, do you think that such data as that contained in the proposed test collection, if set up on a convenient computer, could be of value for your teaching activities?

3 Have you any special requirements in mind?

4 Would you expect or like to be able to use a local computer, or have to rely on remote access?

5 Number of students likely to be involved:

a) undergraduate

b) postgraduate

Name

Department

Address