SURFING THE DIGITAL WAVE?
LESSONS FROM THE IT WORLD

Karen Spärck Jones

Computer Laboratory
University of Cambridge

2/06
talk structure

1. BA Policy Review ( brief plug )

2. BA survey - what users do ( brief example )

3. Searching for ( text ) information :
   history and experience

4. Implications for NCSE
1. E-resources for Research in the Humanities and Social Sciences [Web]

ICT issues & opportunities
from researcher point of view

interested party consultation
institutions and individuals

factor analysis and review
technological eg resource forms
organisational eg repositories
*many* recommendations -

providers respect users, coordinate
encourage secondary resources
address licensing and fair use
attack long-term preservation

...
2. BA user survey

‘Using electronic resources has made all sorts of things possible that didn’t used to be possible. EVERYTHING is different from how it used to be.’

‘The most important research tool for me is Google, probably.’

‘It is maddening that copyright constraints prevent the Web dissemination of resources of no commercial value.’
‘A lot of what I want is in Baghdad.’
3. Searching for text information (IR/DR/TR)

received wisdom vs actual practice

lessons from digital text data files
received wisdom:

quality control vital -
    indexing languages & classifications
    overcome linguistic variation
    identify important notions

assumptions:
    content and its representation univocal
    know in advance what’s wanted

both completely wrong, hence damaging
brief history of automated document retrieval

growth of technical literature (50s+)
arrival of mechanisation

aim in automating:

replicate conventional library indexing
  specialised vocabularies
  few keys
  limited term relations

alternative strategies proposed - what best?
  tests - document, request, relevance sets
results totally unexpected

emulating human indexing not useful

strict languages and indexing not effective
lessons ( on matching, importance ):

authors know what they’re talking about and their words are good

though many language variations over time
if topic same, language connection
if topic matters, language repetition

language is profuse, not parsimonious
redundancy anti ambiguity, pro point
strategy:

matching - use natural language

the more (any) words the better
recall - get a if not b if try both
precision - both a and b if can get

importance - use word frequency

the more words unexpected the better
in a document, in the file

[ also use word cooccurrence frequency ]
strategy features:

meaning indirectly by statistics
minimal prescriptiveness presumption
maximal descriptive response

robust theoretical underpinnings

excellent experimental support

good for any ‘text’ especially ‘full’ text

* Web engines use these ideas
4. Implications for eg NCSE

simplicity is good:

sound basic indexing, searching

robust over time

tolerant over system change

accommodating over user change

[ do underpinnings decently -
character codes, formats, languages etc -
but don’t go ott ]
keep description, access simple

easy to implement
easy to change
	sound failsafe, always offers some handles

treat classification schemes etc as extras :

	ie as support tools, not basic tools
a resource example:

BA PORTAL Web site -
    simple search, lightweight classification

some user comments -

'I ... like the "utilitarian" feel of the site'

'no time wasting with aesthetically questionable graphics'

'It was certainly very easy to use'