Learning the Lessons of the Past

Robert Fugmann

Presented before the Second Conference on the History and Heritage of Scientific and Technical Information Systems

Philadelphia - 2002 November 16

Introduction

- The re-use and processing of reported infor-mation is ubiquitous in almost every field of activity because it is crucial to any progress. However, information processing is observed to be widely executed in a more or less dilettantish manner. Many newcomers enter the field of information science with a feeling of competence but without acquiring a suffi-cient background because the field appears "deceptively simple" to them (Bates 1998).

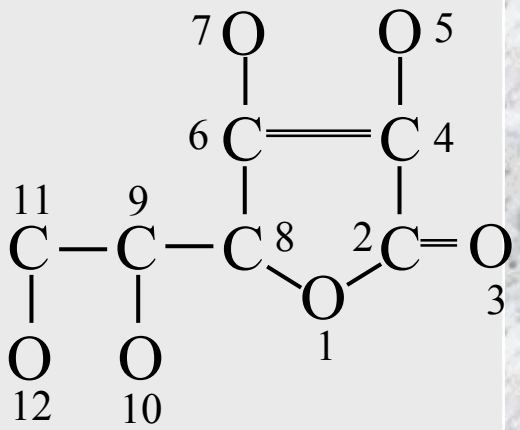
- Hence, much futile research is going on here (as criticized, e.g., by Bates 1999, Goldmann 1987).
- Several lessons to the better have been taught in the past, and through their remembrance a tremendous waste of time and manpower could be avoided.
- In this paper, some of these lessons are recalled.

Lesson 1: The Analytico-Synthetic Approach

- Realize the features of the uniquely successful topological approach to chemical molecular documentation, with its ratios both of precision and recall of 100 percent, as a rule, in a file of presently more than 20 million items.

L-Ascorbic acid
(+)-Ascorbic acid
3-keto-Gulofuranolactone
3-Oxo-L-gluofuranolactone
Adenex
Allercorb
Antiscorbic vitamin
Ascoltin
Ascorbajen
Ascorbic acid
…
Xyloascorbic acid, L-

$C_6H_8O_6$



Atom No.	1	2	3	4	5	6	7	8	9	10	11	12
Conn.	-	1	2	2	4	4	6	1	8	9	9	11
Element	O	C	O	C	O	C	O	C	C	O	C	O
Bond		1	2	1	1	2	1	1	1	1	1	1
RC	6	1	8									

- This topological approach can be looked upon as a variation of the Analytico-Synthetic Approach (Ranganathan 1957).
- It can serve as a model for subject analysis and indexing in fields different from chemistry.

- The Analytico-Synthetic Approach suggests a mode of subject analysis by way
  - both of a predetermined, sufficiently perspicuous vocabulary and
  - of a grammar

- If a heavily used, large and continually growing information system of an enduring usefulness and effectivity is the goal, such an approach is recommendable, in spite of the considerable intellectual effort in the input stage here.
- The return is excellence in retrieval and the system's survival power under the strong and steadily increasing strain of practice.

Lesson 2: Verbal vs. Concept Plane

- Distinguish the *verbal plane* from the *concept plane* (Ranganathan's "idea plane") and
- remember the distinction between *concept* and *expression*.

- A word can represent many concepts and a concept can be represented through an infinity of words from which authors and questioners make their choice, which is often an unpredictable one.
- The omission of the distinction between words and concepts constitutes an obstacle to concept indexing and results in mere textword "indexing".

- Clinging to the verbal plane does not lead to an index but to a concordance,
- i.e., a list of locators for textwords, at most with the inclusion of their morphological variations.

Lesson 3: Recall vs. Discovery

- Distinguish questions of recall from questions of discovery (Bernier 1960).
- They are quite different with respect to the conceptual tools for their adequate execution.

Lesson 4: Individual vs. General Concepts

- Distinguish general concepts from individual concepts (cf. for example, v. Freytag-Loeringhoff).
- Individual concepts are easy to store and to retrieve, quite in contrast to the general ones.

Lesson 5: Natural Language Limitations

- Realize the
  - ambiguity,
  - ellipticalness,
  - unpredictability, and
  - indeterminacy (e.g., Blair 1990) that are inherent in natural language expressions.
- Hence, the usefulness of uncontrolled natural language for the purpose of retrieval is limited.

Lesson 6: Programming Indeterminate Processes

- Be sceptical of all reports claiming "success" in programming indeterminate processes
- Recognize the impossibility of adequately formalizing and computerizing indeterminate processes, especially that of natural language text interpretation and processing (e.g. Bar-Hillel 1964), selected display examples notwithstanding.

Lesson 7a: Necessity of an Index Language

- Recognize the necessity of using an index language in input for attaining those degrees of predictability of concept representation that are necessary in large and/or fast growing mechanized information systems.
- This language is necessarily an artificial one because natural language does not meet several requirements for adequate retrieval (see lesson 5).

Lesson 7b: Obey Cutter's Rule

- Obey Cutter's Rule for attaining an adequate indexing and retrieval quality
- In the common practice of merely "controlled indexing" the vocabulary terms are only the permitted ones, and the indexer is not obliged to use only the most appropriate ones.

- Such an input policy cannot result in advanced ratios of precision and recall. It is in opposition to what Cutter has phrased more than a century ago as the rule of the usage of the best-fitting index language terms.
- This rule has been obeyed by generations of librarians and professional indexers.

- Cutter's Rule requires the indexer to trace and to use only the most appropriate terms from an index language vocabulary. Hence, any input according to Cutter is preceded by a search in the vocabulary for those terms which most appropriately represent the *concepts to be entered*.

- Searching a data base for the documents of interest is also preceded by a vocabulary search, and in fact by a search for those terms which represent the *search topic* most appropriately.
- This makes high demands on the perspicuity of the vocabulary. This can be attained and maintained through a complementary index language grammar.

Lesson 8: Limitations of Index Languages

- Recognize the desirability of complementing the vocabulary-based document representation through uncontrolled natural language text input.
- This improves (or even makes possible) an adequate capability of the system
  - for the execution of questions of recall (Bernier 1964),
  - for searches for individual concepts, and
  - for searches for those concepts which are external to the field of the system.

Lesson 9: Retrieval as an Order-Creating Process

- Recognize the nature of the retrieval process as an order creating process (Landry and Rush 1970), the demands on the effectivity of which increase with
  - the increase of the search file,
  - the frequency of searches to be executed, and
  - the progress of specialization in the community of questioners and authors.

Lesson 10: The Small-System Syndrome

- Distrust all small-scale storage and retrieval experiments which expressly or tacitly
  - claim to scale up and
  - claim to be able to meet the requirements of everyday-practice (cf., for example, the warnings by Soergel 1985, Blair 1990b; Blair and Maron 1990c; Brown 1990).

Lesson 11: Don't Adopt Inap-propriate Concepts

- Don't uncritically adopt concepts and their definitions from fields external to yours. This has often been done merely for the sake of measurability (criticized , e.g., by Bar-Hillel 1964). An example is the peculiar "information" concept adopted from message transmission technology and "consistency" from the natural sciences.

Lesson 12: Avoid Inadequate Information Philosophies

- Don't fall victim to the philosophies of instrumentalism and positivism (cf., for example, the warnings by Budd 1995).
- They lead to inappropriate mechanization and to the neglect of what is not easily visible, for example, a low recall ratio.

Summary: Document Interpretation Yields

- Meaning recognition (in the interest of good precision)
- Essence recognition (in the interest of good precision
- Ellipses filling (in the interest of good recall)
- Paraphrase lexicalization (in the interest of good recall)

Summary: Document Interpretation

- Interpretation Omitted
  - Restricted to verbal plane
  - Access to natural language text words via full text files or concordances
- Interpretation Executed
  - Access to concept plane
  - Access to concepts optimally through analysis and synthesis and indexes, i.e., concepts represented with sufficient predictability and fidelity according to Cutter's rule

- Interpretation omitted
  - Suitable for questions of recall (known item searches)
  - Natural language search terms are
    - Remembered
    - Looked up
    - Pre-given
    - Of uncertain meaning
    - Constitute descriptor candidates
- Interpretation executed
  - Suitable for questions of discovery (the objects of interest are only partly known)
  - Search parameters from index language of system, especially for
    - General concepts
    - Syntactical concept connectivities

- Interpretation omitted
  - Input is fast and cheap
  - Output quality suffers from the omission of the four interpretation steps
    - Meaning recognition
    - Essence recognition
    - Ellipses filling
    - Paraphrase lexicalization
- Interpretation executed
  - Input is slow and costly
  - Output quality is high but depends on
    - indexing vocabulary comprehensiveness, specificity, and transparency
    - Degree of grammar employment
    - Indexer's expertise and care

Conclusion 1

- So-called "modern" natural language processing is brilliant in information technology but it is stone age with respect to information philosophy. Much of what librarians and professional indexers have known for decades is being rediscovered in a slow and expensive process. An example is "metadata".

- The detrimental result of this neglect and of the over-emphasis on technology can be viewed world-wide when searches of discovery or searches for general concepts are executed in the Internet.

Conclusion 2

- Full-text storage, in recognition of all its specific strengths, dispenses with proper document interpretation, which includes meaning clarification, essence recognition, ellipses filling, and paraphrase lexicalization. These omissions can only to a very limited extent and only for a selection of examples be counterbalanced by search algorithms or text processing algorithms (e.g., Wellisch 1992).

- This is due to the indeterminacy which is inherent in any natural language text phrasing, and hence in any text interpretation, too. The consequences are inherently low ratios of precision and recall in the searches in full-text files.

Conclusion 3

- Concealing or denying these weaknesses of mechanization constitutes an obstacle to the establishment of those information systems which—and for good reasons—are more or less based on human intervention.

- This attitude also endangers the future of the more traditional, practice-proven information systems, without being able to provide a workable alternative. This is much to the detriment of the communities to be served.