effort to place at a user's fingertips the communications equipment needed to interrogate a large store of information under the control of a computer while numerous other users are simultaneously using it.

The machine research-and-development activities pertinent to the field of information storage and retrieval several desiderata emerge: to find automatic means of converting printed data to machine language; to achieve more compact storage of source material; to enhance intellectual access to information; and to display or provide information rapidly in a form suitable for individual use. Microphotography, the evolution of unconventional subject-classification systems, the application of computers, and the use of communications techniques, among other things, represent various stages in the historical development of information storage and retrieval.

BASIC MODEL OF INFORMATION RETRIEVAL SYSTEMS

Models of information retrieval systems are commonly found in information retrieval texts and papers. Such models are generally in the form, with varying amounts of additional descriptive detail depending of the purpose of the description.

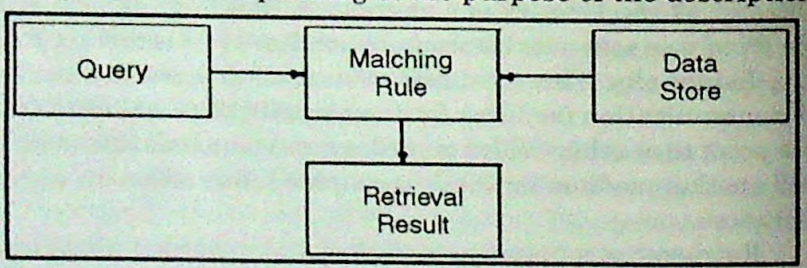

Fig. General Model of Information Retrieval Systems.

A "complete formal model for information retrieval systems" using production grammars and hypergraphs to represent text structure, indexing, and access. However this is really a procedural model of text retrieval techniques. Descriptions of the operation of individual retrieval systems are likely to have detailed flow diagrams of that particular system's components. Here, however, we are interested in

developing a complete, generalized functional analysis of information selection systems. To develop a complete and general model of the functional components of bibliographic information storage and retrieval systems we proceed by outlining a descriptive model of information storage and retrieval procedures. This illustrative model is intended to be minimally complete in that it includes all the *different types* of functional components found in all retrieval systems. The hope is that the components identified in this basic, illustrative model could be used to construct a functional representation of any information storage and retrieval systems of any complexity, including extended retrieval architectures. As a check on the adequacy of the analysis three examples of information storage and retrieval systems will be examined later.

INPUT

System boundaries are arbitrary. Where they are drawn determines which flows of data are regarded as flowing in to and out of the system in whatever form. The inputs, queries and records, may be retained or discarded (perhaps by being relegated to other storage).

There can also be feedback concerning any process:

- One kind of input supplies the stored and potentially retrievable data: documents, bibliographic records, images, etc., and/or representations of them.
- Queries constitute another kind of input in one of several forms: free-text, boolean keywords, formal query language statements, etc.
- External knowledge may also be drawn upon in the form of controlled vocabularies, syndetic structures, subject headings, descriptions, etc.

There can be multiple outputs. The most obvious output is the expression of the retrieval results, in whatever form. More generally there can be feedback reporting the effects of any procedure. With this in mind, we now outline the functional components that appear to be necessary and sufficient to represent information storage and retrieval systems. Not all components are present in all systems. Some components could

be present more than once. Components may be implemented in more than one way, *i.e.* using different techniques. Note also that, as is usually the case in systems analysis, the granularity of the components is somewhat arbitrary. We propose that the following components are necessary and sufficient, between them, to represent the functionality of all operational information storage and retrieval systems. The intention is that the analysis will be technologically independent, one that would be as valid for paper-based as for computer-based retrieval systems.

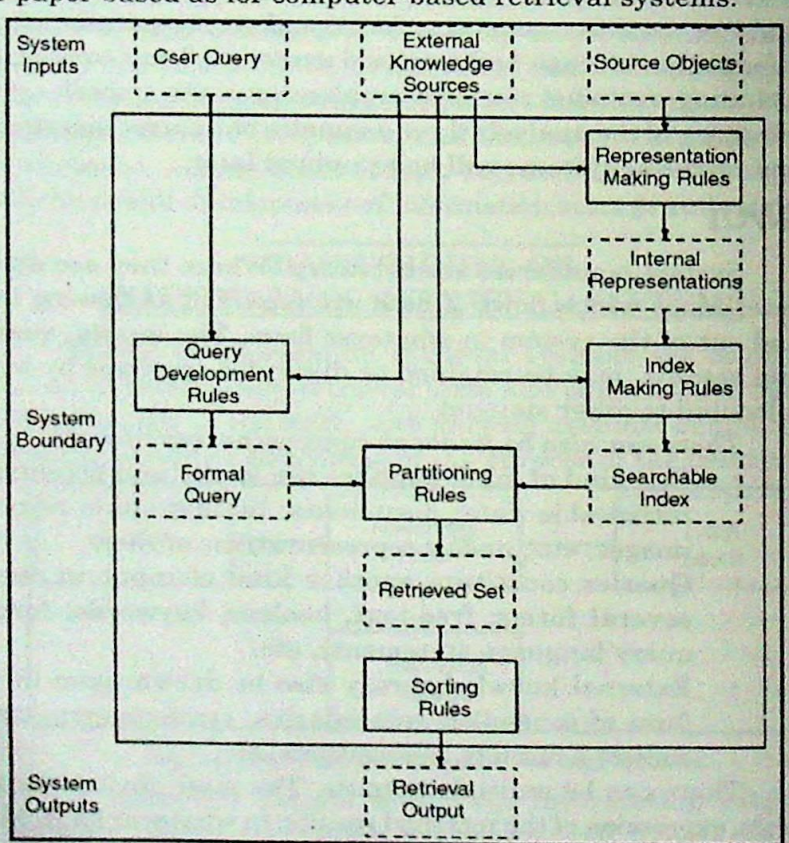

Fig. A Minimally Complete Model of Bibliographic Oriented Information Retrieval (Selection) Systems. Solid Boxes Indicate Processes ("Transformers" or "Partitioners"). Dashed Boxes Data Objects. *Italics* Indicate Optional Components. Arrows Show Flows (or Streams) of Information Objects. Note that the only Required "Process" is the Central Partitioning Rule, and that Subcomponents are Formed by Patterns of Objects ⇒ Process ⇒ Objects.

INPUT STREAMS

User Query

One form of input from the environment is the User's query, an expression of the user's information need, more or less compromised by the user's expectation of and experience with the information retrieval system. The "user" is ordinarily thought of as a human being, but the query could well be generated by a machine and only indirectly by a human being, such as in the case of relevance feedback or multi-stage retrieval.

Source Objects

A set of Source Objects of interest: documents, records, artifacts, images, signals, etc. These arise in the environment outside of the information selection system. A general theory of information storage and retrieval should be able to include bibliographic systems (searching records representing documents), "full-text" searching, copies or representations of museum artifacts, and, indeed, of any kind of definable phenomena, including imaginary ones. The set of source objects may well be a carefully selected set, as in a library or museum collection.

These objects and/or copies and/or transformations of them become "resource input" to the retrieval system. Through a variety of possible processes, they become the Representations in the system.

External Knowledge Sources

External Knowledge Sources are used in information selection Representation Making, Index Making and Query Development. In Representation Making, the external knowledge may be in the form of what people know or has been recorded concerning the Source Objects, their contexts (*e.g.* domain knowledge), their possible representations (*e.g.* linguistic knowledge, thesauri, etc.), or their internal structures and interrelationships, are another resource which can be drawn on as a supplement to or as a substitute for the source

objects. Such knowledge can also be used in Query Development and Index Making in the form of thesauri, controlled vocabularies, subject heading lists, classification schemes, syndetic structures, dictionaries, search intermediaries, etc.

One of the major research areas in this field is to see how far this external knowledge can be formalized and moved *inside* the system and used in this way.

In an ideal world these three processes (Representation Making, Index Making, and Query Development) would all draw on the same, identical knowledge sources, but this is unlikely in practice. With the rise of client/server architectures, we can expect separation of the Query Knowledge Sources from the Representation Knowledge Sources and the Searchable Index Knowledge Sources.

Knowledge Sources need to be continuously revised and updated and there is no assurance that the updating will be identical and synchronized. Further, the use of an External Knowledge Source in creating Representations is chronologically prior to the use of the External Knowledge Source for Query Development and may be several years prior, creating possible vocabulary problems even if the same External Knowledge Sources were used.

INTERNAL COMPONENTS

Representations

Representations of the source objects are composed from the resource inputs in some combination of a copy or transformation according to the Representation Making Rules of part (or all) of the Source Objects and/or any (external) representations or descriptions (from External Knowledge Sources) of those resource objects.

Representations can be derived from:

1. Source Objects
   - Part or all of the object itself, possibly copied and/or transformed. For textual objects these could include the text, title, original abstract, etc. For images, these could be scanned copies.

- Descriptive features implicitly in or algorithmically derivable from the object: *e.g.* word occurrence, frequency, and co-occurrence; automatic abstractions from (or patterns recognized in) images; etc.

2. External Knowledge Sources
   - Features derivable from other objects inside (*e.g.* relative word frequency in relation to a corpus) or outside (*e.g.* synonyms of topical terms) the retrieval system that are related to this object.
   - Description or documentation of the object: description of the physical object and/or statements about the origins of the object and/or what the object signifies, *e.g.* subject headings, subject classification.

66167

Depending on the nature and extent of the Representation Making Rules, the Representations, then, might be a more or less transformed copies of the Source Objects: in a collection of unedited full-texts, each text (or copy of it) would constitute its own Representation. It might be a more of less transformed description of the object: in museum registration the representation might include an image of the object, but none of the original object itself (unless, presumably, it is a museum of electronic objects). In other cases the representation could be derived in part from the source object and in part from a description: in bibliographic systems, such as a library catalog, fragments derived from the object (*e.g.* title, publisher's name) would be combined with pieces of description (*e.g.* subject headings).

Searchable Index

Since the Representation is what is stored, the Representation is also that which could, in principle, be searched and, following selection, produced as output for display or other purposes. But this is not necessarily supported in practice. Current online library catalogs, for example, typically restrict searching to a few fields (notably author, title, and subject headings) within Representations that contain

several other fields in which searching is not supported. This is sufficient reason why it is necessary to make a distinction between the Representation and the *Searchable Index*. The Searchable Index, in this technical sense, is the searchable part of the Representation.

We use "Searchable Index Rules" to denote whatever determines what is to be searchable. Retrieval systems commonly have in addition, a syndetic structure for mapping permissible searches, which we also treat as a second component of the Searchable Index. Again, in the case of unedited full-text, the Searchable Index will be co-extensive with the Representation and, therefore, with the Source Object (the original text). But, as noted, in other cases, such as library catalogs, the Index Making Rules can restrict which parts of the Representation are available in the Searchable Index. The Searchable Index (like the Representation and the Source Object) might be partitioned into separate (sub)indexes, to allow more precise, targeted searching.

Procedurally, the Index Making process can be implemented in different ways: the Searchable Index might be derived by literally making parts of the Representations available for searching; it might be derived by copying parts of Representations; it may even be that part or all of the Representations exist physically only as fragments distributed via the Index Making Rules to the Searchable Index to be reassembled if and when needed. But we regard these alternatives as functionally equivalent and are not interested here in the technical details of implementation (storage costs, search effort, delay, etc.) that will make one technique preferable to another.

Query Development Rules and Formal Queries

Query development is a function that mediates between the User Query and the Formal Query. It transforms the user's query in order to harmonize it with the system's vocabulary of retrieval commands, index specification, and index vocabulary prior to retrieval. This role has traditionally been seen as an important function for skilled human intermediaries.

Computer-based query development that can match queries with the vocabulary in (or expected to be in) the system's Searchable Index is commonly called an "entry vocabulary" module. Examples include CITE, PaperChase, and Grateful Med. Automation of this function is promising and offers scope for expert and probabilistic techniques. "Entry vocabulary" modules parallel the syndetic structure, thesaurus and controlled vocabulary aspects of External Knowledge Sources used to create the Representation. It might ideally draw on the *same* thesaurus or other knowledge representation scheme, but it cannot be assumed that the same external sources will be used for these different components.

A query development system may be absent, present, or multiply present in any given retrieval system.

The Formal Query is the query as it is seen by the Matching Rule, after it has been transformed by the Query Development Rules. Examples of such formal transformations include truncation, weighting, substitution, normalization, vectorization, etc., many of which are conversions of "external" representations to "internal" representations. Such transformations apply both to computer and human based retrieval systems.

Retrieved Sets

A Retrieved set is logically a subset of the Representations as partitioned off by the outcome of the Matching Rule applied to the Formal Query and the Searchable Index. When displayed (or delivered as output) the retrieved set may be complete copies or very incomplete, transformed versions of members of the set of Representations. Note that this is not necessarily a simple binary outcome: Retrieve and Not Retrieved.

Sorting Rules

Commonly, but not necessarily, there is a separate process of sorting the retrieved set. Online library catalogs typically reorder retrieved sets alphabetically by author (strictly, by "main entry") prior to display. In card catalogs the order of the retrieved set is predetermined by the order in which the cards