Gems from our Digitization Project

Reprinted from *International Classification: Journal on Theory and Practice of Universal and Special Classification Systems and Thesauri = Zeitschrift zur Theorie und Praxis universaler und spezialler Klassifikationssyteme und Thesauri.* Vol. 5 (1978) No. 3. The masthead identifies the "editors" as: Ingetraut Dahlberg, Alwin Diemer, Arashanipalal Neelameghan and Jean M. Perreault. The article is reprinted without editorial interpolation—*Ed.-in-chief.*

Elaine Svenonius
University of Western Ontario,
School of Library and Information Science, London, Canada

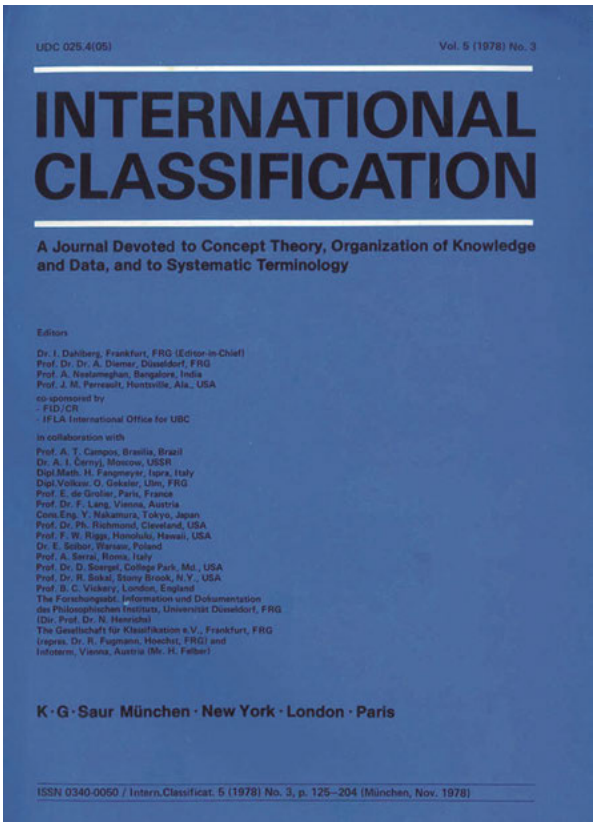
Facet Definition: A Case Study

Abstract: Historical account of the sophisticated method of indexing developed by J. 0. Kaiser (1896/97), a librarian at the Philadelphia Commercial Museum who established his index on cards (a novelty then) and distinguished his items according to the categories 'concrete', 'process', and 'country'. He also introduced "statement indexing" and rules to this end in order to permit the supply of "complete information" on a subject in a document. In summarizing these findings, the author stresses the necessity of establishing well-defined categories if an organization of terms is to serve e.g. information retrieval. (I. C.)

Svenonius, E.: Facet definition: a case study. In: *Intern. Classificat.* 5 (1978) No.3, p. 134-141.

1.0 Introduction

Julius Otto Kaiser developed a method of indexing called "Systematic Indexing." The publication of the first draft of this scheme of indexing in Philadelphia in 1896-97 is an important milestone in the history of indexing theory. Olding credits Kaiser's work as the greatest single advance in indexing theory since Cutter (1). Metcalfe, even more eulogistic, says that "in sheer capacity for really scientific and logical thinking, Kaiser's was probably the best mind that has ever applied itself to subject indexing" (2). Kaiser seems to have been the first to recognize indexing language *qua* a language with grammatical categories and rules of syntax. He may thus be regarded as the originator of faceted indexing. The purpose



of the present paper is to examine Kaiser's indexing theory in some detail and then to relate briefly this theory to modern work in the developing of string index languages and in the structuring of natural language text for automatic information retrieval. But first some words of background.

2.0 Background

Kaiser was librarian at the Philadelphia Commercial Museum from 1896 until 1899. It is perhaps significant, considering his linguistic approach to indexing theory, that before this he earned his living as a teacher of languages and music. [1] Turning to a new field in 1896 ushered in a period of creativity for Kaiser. The first draft of his indexing scheme was completed within a year. This scheme underwent a period of testing for several years with an index consisting of some 50.000 cards. Then it was rewritten and used in making three different cards indexes of a technical nature (3).

At the turn of the century the making of an index on cards was a novelty. In 1908 Kaiser described how one went about this in a book called *The Card System at the Of-*

fice (4). Dealing with questions of managing and filing materials, this book was published as Volume 1 of "The Card System Series." It seems to have been enthusiastically received by the press. It was seen as the application of system … to business, [2] "well worthy of the attention of any one who has to bring into an order convenient for quick and easy reference any large array of miscellaneous facts or points.[3] Of Kaiser's system, the *Modern Business* of November 1908 wrote:

> The card-index system (sic) of filing letters, papers, etc., has undoubtedly come to stay, and the old "letter-book" method is becoming more and more a thing of the past. For the last few years a revolution has slowly but surely been taking place in the office methods of modern business houses. Makers of filing cabinets and the accessories thereto have all their work cut out in order to meet the rapidly increasing demand for these articles.[4]

The Card System at the Office serves as an introduction to the second volume in "The Card System Series," the more theoretical work *Systematic Indexing.*[5] *Systematic Indexing* was published in London in 1911. At this time Kaiser was working in London as Librarian of the Tariff Commission. The suggestion has been made that Kaiser's indexing system, even though invented as early as 1896, was particularly tailored to deal with commercial information.[6] The Tariff Commission records contained information of a varied sort relating to commerce and industry, including "in addition to correspondence, evidence of witnesses, extracts from official reports and newspapers, estimates of costs, details of competition in innumerable articles in all leading countries of the world, and so on.[7] It is true that most of the examples in *Systematic Indexing* are taken from commerce and industry. As will be seen Kaiser focussed particularly on commodities, their properties and the countries from which they came. It is thus plausible that the theoretical expression his system took was in part determined by the fact that its primary application was in a business library. One might speculate as well on the plausibility that Kaiser's training in languages and music was a determining influence, as was the circumstance that his system was developed to be used on cards.

3.0 Systematic indexing

Kaiser understood indexing to be that "by which we make our information accessible"[8] (45). He is modern in his emphasis that it is information and not books, the containers of information, that is to be made accessible.

> But for business purposes we must try to dissociate *information* from literature, we do not want books, we want information and although this information is contained in books, it should be looked upon as quite a different material and it must be treated differently from books (83).

Indexing as viewed by Kaiser has both a negative and a positive function, throwing out what is not required and concentrating on that which is required (45).

> By the process of indexing therefore we boil down, we reduce our materials to that which is essential for our purpose, we create a nucleus of effective information, information which will be of real use to us in the pursuit of our business (46).

It has been suggested that Kaiser never read Cutter's Rules.[9] However, like Cutter, he held that the purpose of indexing was to bring like subjects together.

> Our purpose in analysing literature is: to discover those elements by means of which we may bring together knowledge or information of a like kind (297).

and

> The statement (index expression) as will be seen gives the elements which we require to collect information on like subjects … It does not give us the complete information (303).

Kaiser was an admirer of "system." He argued that systematic effort must in the long run effect economies, since, by system, duplication is eliminated and control concentrated (18). By systematic indexing he meant indicating information not with natural language expres- sions, as Cutter was advocating, but by expressions constructed artificially according to formula.

> We shall take literature to pieces and re-arrange the pieces systematically so as to answer best our object in view. We shall see that by this method almost mathematical exactness can be reached in the manipulation and coordination of our information (16).

Kaiser used the expression "literature" almost synonymously with "text." There are various ways in which a text can be analyzed or "taken to pieces." There is grammatical analysis which "has for its basis words and for its purpose the correct use and combination of these words" (296). There is logical analysis which has for its basis reason and

its purpose the demonstration of correct ways of reasoning (296). And then there is a third kind of analysis recognized by Kaiser: one which is based on knowledge and which has for its purpose bring- ing together knowledge or information of a like kind (297). This sort of analysis is the first step in systematic indexing.

The second step in systematic indexing is synthesis. By re-arranging pieces of literature systematically Kaiser meant combining them according to prescribed rules. As was mentioned Kaiser was not a proponent of natural language indexing. Like others before him, Leibniz for instance, he grudged natural language its approximateness:

> Language as a means of expression is not a systematic effort. There is no machinery for regularizing or standardizing language (67).

It was to provide just such a "machinery for regularizing or standardizing language" that Kaiser developed his Systematic Indexing language. This language is an artificial language, but not a language in which to reason, like Leibniz' *characteristica universalis;* rather it is a language to be used for the special purpose of indexing, that is, for bringing together knowledge or information of a like kind (297).

4.0 Epistemological foundations

Kaiser recognized three kinds of index terms:

> (1) terms of concretes, representing things, real or imaginary (e.g. money, machines); (2) terms of processes, representing either conditions attaching to things or their actions (trade, manufacture); and (3) terms of localities, representing, for the most part, countries (France, South Africa). The division of terms into those naming concretes and processes has some grounding in epistemological theory. Knowledge begins with observation, and, according to Kaiser, observations are limited to concretes and their conditions … "there is nothing else to observe" (56).

Kaiser did not consciously borrow or himself construct a sound epistemological theory. The slightest probing reveals paradox, for instance in the matter of concretes being both knowable and unknowable. In a simple sense, anything that can be pointed to is knowable.

> Even in their most complex forms - for instance a battleship specifically pointed out - we know of what they are composed, there is no margin for doubt as to what is included and what is excluded. Each concrete represents something definite to handle and there is a fair chance therefore of bringing a number of concretes into a reasonably ordered sequence (108).

We can perceive the outlines of a concrete object like a battleship. We can touch it. in this sense it is knowable. Abstract things, like subject disciplines, e.g. chemistry and physics, are not so knowable. One reason is that their boundaries are not defined in space. Even abstract boundary conditions seem difficult to formulate. Kaiser believed that the classification of things as amorphous as subject disciplines was impossible (43). He thus preferred to ground his classification in tangibles, *viz.* concretes that occupy space and have form.

But there is a sense in which even these very tangible concretes are unknowable. Kaiser at times writes in a somewhat Kantian vein. We cannot really observe concretes, that is, we cannot observe them in themselves. All that we can observe, only, are concretes in action or concretes under certain conditions.

> Concretes are only known to us superficially. We perceive their likenesses and differences by comparing them. We are unable to give a complete description of any concrete, no matter how many attempt a description (54).

and:

> Since we cannot tell what concretes are, we are obliged to give increased attention to their processes, to what they do or what we can do with them. We observe their behavior under given conditions, we compare results. Electricity for instance is a concrete, but it is only known to us by its actions, and it is by observing its actions that we arrive at any appreciation at all as to what its probable nature is (55).

Kaiser thus gives the impression of believing something like a battleship to be knowable, while acknowledging that it is not. Some resolution of the paradox might be achieved by distinguishing between knowing in the sense of knowing boundary conditions and knowing in the sense of knowing the true nature of a thing, as opposed to its phenomenal nature. Still the fact cannot be glossed that Kaiser will both have his cake and eat it; on the one hand we are "unable to give a complete description of any concrete" (54) and, on the other hand, "we know of what they are composed, there is no margin for doubt as to what is included and what is excluded" (108).

Indexing languages that purport to have a semantics, in the sense of real-world mappings, are only as systematic as the epistemology on which they are grounded.

Kaiser's wavering over the knowability or unknowability of concretes had some effect on the systematics of his Systematic Indexing. As will be seen this is particularly evident when he comes to deal with abstract objects, such as the notions of mathematics.

5.0 Semantic theory

A theory of meaning undergirds Kaiser's indexing language. Sometimes called the naming theory of meaning it is one of the oldest views existent, being introduced first by Plato in his *Craty/us.* It is called the "naming theory" of meaning because in it words are regarded as referring to things and hence as the names or labels for things. Kaiser writes:

> The subjects of our observing and reasoning are things in general, real or imaginary, and the conditions attaching to them. We shall call them *concretes* and *processes* respectively. The concretes are given names to distinguish them, the various conditions attaching to them are also named separately. Names are rendered by means of signs or symbols - letters; letters are grouped into words; names may consist of one or more words. Words are brought into relation according to recognized rules and thus give language (52, 53).

It is not clear whether Kaiser thought that all words had a naming function (the above passage suggests this) or only those that were to be used for the special purpose of indexing. It would be nice actually if there were evidence for the latter, more sophisticated view. Some evidence is provided by the following (the italicized portion): *"for the purpose of indexing* we shall divide our stock of names or terms into those on concretes, processes, and countries" (73). But one must allow that the mention of a special purpose may be casual here; certainly it is not conclusive.

One of the usual criticisms levelled against the naming theory of meaning is that many words lack real-world referents; for instance it is difficult to imagine what is named by words such as *love, truth* and *beauty,* since these correspond to no physical entities in the real world. Words whose main function is syntactic also present problems; for instance, prepositions and articles lack ontological grounding. To meet this criticism those who endeavor to maintain a consistent naming theory of meaning are obliged to invent perceptual or conceptual constructs to serve as referents for abstract words. However, inventions of this sort are open to the ghost-in-the- machine objection, *viz.* concepts are invented to account for meaning the way ghosts may be posited to account for the working of a machine. The difficulty is that explanations, like definitions, are supposed to account for what is unknown in terms of knowns and not other unknowns.

Kaiser was certainly aware of the problems with names. He worried about the extension of things referred to by names.

> Names certainly represent concretes and processes, but it would be rash to say that there is a general agreement as to what is exactly covered by a particular name. The difficulty of definition is aggravated when we come to collective names. Names have come about in a haphazard way … (112).

It would have suited Kaiser's system better if each concrete and each process to which it was subject were represented by a unique name. Homonyms he found awkward. In particular he did not like those which seemingly could name either a concrete or a process:

> Naturally one should have thought that there would be distinct names at any rate for concretes and for processes, but that is not always the case. Thus the word organisation may be either the name of a concrete or a process. In the concrete sense we may speak of the army as an organisation, in the process sense we may speak of the work connected with bringing an army into being as organisation (111).

Homonyms are always a problem in index languages because in indexes words stand alone and there is no context to resolve which of two or more meanings is intended by a given homonym. Kaiser was quite conscious of this and he designed his index language so that a distinction could be made between homonyms which named processes and those which named concretes.

> Besides organisation there are many other names with both meanings, and to keep these two kinds of names sharply apart is one of the main features of the method of indexing proposed in this book (111).

There were primarily two means by which Kaiser kept apart the two kinds of names. The first was to insist that where ambiguity was possible a process term should be stated in the gerundive, i.e. *organizing* rather than organization. The second was to indicate syntactically, by means of position, whether a homonym named a concrete or a process. This was possible because in an expression in Kaiser's language the name of a process is normally preceded by the name of a concrete. What could not be resolved, however, were homonyms that named two different concretes or two different processes.

Equally worrisome to Kaiser was the fact that some words seemed, simultaneously, to name both a concrete and a process:

> … our names are of a very mixed character. Leaving aside the question of relatively specific and collective terms, they may be divided into:
> names of concretes …… coin, copper, etc.
> names of processes ..... minting, insurance, etc.
> and combinations of both concrete and process, for example the following:
> bibliography …… book description
> agriculture …… land cultivation (184).

This was an anathema to Kaiser, that one word could name both a concrete and a process, for above all what characterized his indexing as systematic was that these two kinds of names could be kept separate. Not only were the two categories of terms, concrete and processes, to be mutually exclusive, but any term even when seen out of context could be recognized as belonging to one or the other category. To deal with problematic words which could not be so recognized Kaiser resorted to a measure that at first sight seems extraordinary. At least it seems extraordinary in light of the fact that most indexing theorists from Cutter onward have opted for "natural language indexing." Not Kaiser, however. He was ready to remold natural language to suit his ontological commitments. In particular, he felt that single words, such as *bibliography,* which implicitly refer to both a concrete and a process, should be replaced by two separate words which explicitly referred to the concrete and process as distinct from each other:

> However, our language is a very heterogeneous mixture of terms; it happens that it actually comprises terms made up of a concrete term and a process term. In the list you will find AGRiculture and BACTERiology belonging to this class. How are we to deal with these? If they were admitted into the index like concretes it would upset the entire arrangement; we should be forced to fall back on a mixture of terms as used in book classification, from which I have been trying to escape at all cost. The only way open is to cut these terms in two, separating them into concrete and process, although I dislike interfering with terms as given. Thus Agriculture etymologically means "LAND … cultivation"; for Bacteriology we may use "BACTERIUM … study," etc. (Aslib Report, p. 154)[10].

A first principle in Kaiser's systematic indexing is that all information is to be filed under the concrete it is "about." Kaiser could not therefore deal with terms in which were embedded both a concrete and a process. By his own admission, however, he seems to have opened a Pandora's box with the suggestion that language might be redesigned to suit the purposes of his systematic indexing. A case in point is the logic of concretes expressing money:

> All terms of money, as credit, dividend, capital, debenture, export duty, bounty, surcharge, etc. are concretes and should be treated as such, even price may be treated as a concrete, if the exigencies of the business warrant it... The price of coal implies the exchange of coal and the exchange of money and logically we should have to index the two concretes. But this would be going too far … (325).

The logic of concretes becomes even more fuzzy when it comes to terms that express energy of some kind, e.g. *Labour, Power, Light.* In the 1926 Aslib Report Kaiser writes:

> Terms of commodities and terms of energies may therefore be put into one class; I have called them CONCRETES, in the sense of concrete existences ... (Inclusion of energy is forced, because commodities comprise latent energy.) (Aslib Report, p. 149.)

If Kaiser had his way he might have banished all words whose referents were problematical. He admitted, for instance, that it was a weak point in his system that it could not handle mathematical terms:

> there still remain certain terms which are neither concrete nor process. These are mainly mathematical terms such as Coefficient, Constant, Factor, Ratio, etc. Of course, I might say: "Exceptions prove the rule," and content myself with that; but in systematic work this way of reasoning would be fatal. To my mind one single exception proves that the rule is *no rule.* Here then is a weakness in my scheme. (Aslib Report, p. 155.)

Had Kaiser been born slightly later he might have made use of the set theoretic definitions of mathematical terms, definitions which during his own lifetime were being developed by Russell and Whitehead. As well he might have found the distinction of logical types (first order entities, second order entities, etc.) useful in a classification of concretes.

In any case it seems clear that Kaiser was well aware of the difficulties inherent in the view that all words function as names. In one place he suggests that notation (call numbers) provide better "names" than nomenclature, since there is not the difficulty of definition (133). He seems to

have been especially wary of prepositions (words particularly unname-like) for the reason that they create confusion in filing (324). Yet despite these difficulties of category definition, Kaiser could not relinquish his view that when we look at the world all we observe are concretes and processes and there are the things that words of language name. In the Aslib Report he writes:

> I am still hoping that some way may be found to incorporate the few mathematical terms and at the same time make the definitions of concrete and process more precise. (Aslib Report, p. 155.)

It has been suggested that Kaiser regarded Country or Locality as a special variety of Concrete.[11] The suggestion is warranted by some places in the text, but there are also enough contrary indications to make for doubt. In (299) Kaiser classifies concretes ("concrete articles" or "commodities") into three types: movable (silk, hardware …), immovable (land, rivers …) and abstract (labour, mental and manual …). Immovable commodities he saw as including countries; yet he also saw countries as representing a distinct class.

> Immovable commodities include one kind of special importance - *countries* in the political sense. Their peculiarity is to be sought not so much in their territories, but more especially in the authority exercised within each territory as expressed in their laws etc. In addition there are the peculiarities of the inhabitants as expressed in their language, customs and habits. For these reasons we are obliged to treat the political divisions called countries as a distinct class. (300)

The passage seems to be internally inconsistent, stating on the one hand that countries form a subclass of concretes and, on the other hand, that they form a distinct, nonoverlapping class. Under these circumstances, it is difficult to say what Kaiser really thought. Given his ontological commitment to two kinds of entities (all that we observe are "things in general, real or imaginary, and conditions attaching to them" (52), it seems reasonable to suppose that he wanted to recognize only two categories of terms. However, in numerous places he makes reference to three distinct categories of terms. The evidence seems weighted in favor of the tripartite division. A country is not a concrete in the sense of being something "definite to handle." More telling perhaps is that the grammar of his index language quite obviously assumes that concretes and countries are separate syntactic categories. In summary, one might say that Kaiser, while he recognized that the category country was required, from a practical point of view, was nevertheless not going to allow it to intrude upon his theory. It is significant that in the Aslib Report he does not even consider the question of countries, except to say that they have not been mentioned because they do not lead to any difficulties (p. 151).

6.0 Syntactic rules

An expression in Kaiser's index language is called a *Statement*. It consists of a sequence of names or terms. Permissible sequences of terms are prescribed by a set of rules which make reference to term categories. That is, the order of terms in a Statement is determined by the categories to which these terms have been assigned. Only three citation orders are permitted: (1) a term in the concrete category followed by one in the process category, (e.g. wool-Scouring); (2) a country term followed by a process term (e.g. Brazil-Education); and (3) a concrete term followed by a country term, followed by a process term (e.g. Nitrate-Chile-Trade). Strictly only the last formula is "complete." In (303) Kaiser writes that "A statement strictly speaking must always consist of concrete, country and process." He implies thus that it is both necessary and sufficient to name three aspects (facets) of a piece of information in order to bring all information on like subjects together. "The statement as will be seen gives us the elements which we require to collect together information on like subjects" (303). Kaiser justifies his first two "incomplete" formulas on the grounds that sometimes the country or concrete facet is very general or is well understood:

> but experience will show that often no country is given, and sometimes there is apparently no concrete. A moment's reflection will make it clear however that the country is only omitted where the action is not necessarily confined to a particular country, the action may hold good for all or most countries, and similarly where the concrete is missing, its character is so general or unmistakable that in ordinary language the process indicates sufficiently the concrete (303).

A canonical Statement then is a concrete-process-country combination. These three terms are sufficient to collocate information on like subjects; however, they may not suffice "fully" to describe an article or piece of literature. Kaiser allowed for fuller descriptions by allowing that a Statement could be extended by appending to it an Amplification. As an index term, for Kaiser, corresponds to a Statement, so an abstract corresponds to an Amplification. The purpose of an Amplification is to "complete the information" on a given concrete:

> In the statement we have reduced the information to a skeleton, divested of all qualifying verbiage, the amplification serves to supply whatever is required to complete the information, and in the form in which it is desired (349).

It is interesting that Kaiser had some notion of what constitutes "complete information" on a given concrete (when is a concrete completely described?) In (350) he specifies the various data elements which might appear in an Amplification as follows: date of information; extension of Statement (i.e. a further elaboration of the subject); authors; name of publication; place and date; pagination, edition, etc.; and call numbers. Though concerned about the possible data elements to be included in an Amplification, Kaiser was not very particular about its structuring:

> While the statement must be constructed on very definite rules because it is also used for the filing or classing of the information, more latitude may be allowed in the amplification … Again while the statement is obligatory, the amplification is more or less optional (349).

Together a Statement and an Amplification constitute the complete information on any given concrete and is called by Kaiser an "index item" (305) or a "unit piece of knowledge" (Aslib Report, p. 149). Thus Kaiser handles the question of aboutness. A Statement, its Amplification, and the two taken together as an "index item" are about a concrete. Aboutness applies only to concretes, and all Statements are about concretes, whether they explicitly include a concrete term or not.

> As has already been pointed out an information, an article, a paragraph or a chapter contains as many items for indexing as it contains separate statements, in other words, there will be at least as many items as there are concretes, for some- times it happens that the same concrete must be taken more than once because the description includes widely different processes (308).

Kaiser regarded the Statement as the main feature of his indexing method (306) and, indeed, the inventing of it represents a giant step forward in indexing theory. There is no doubt that Kaiser wished to break with the past. Existing library classifications he saw as wasteful, because of their excessive duplication. What he disliked especially was that different aspects of the same concrete were scattered all over a classification. An example he cites in the Aslib Report is the handling of coal in the Dewey classification:

> certain information may be filed under Coal, but with equal reason it may also be filed under Combustion, Analysis, etc., or under their respective call numbers. When information is wanted on Coal, every one of such likely headings would have to be searched each time in addition to Coal, which not only involves a good deal of extra time, but also considerable uncertainty as to what headings should be searched or disregarded. Maximum duplication occurred in the index with just such terms of commodities as Coal and terms implying an action or verb, like Combustion, etc. (Aslib Report, p. 147.)

As we have seen Kaiser's means of eliminating such duplication is to restrict headings to terms of concretes, subdivided by terms of process.

It has been suggested that Kaiser's systematic indexing was a development of Cutter's alphabetic subject heading language[12]. But it is doubtful that he ever intended to construct a consistent grammar for subject-heading language. Indeed there was no need for such a grammar, since in an index language such as the Library of Congress Subject Headings (LCSH), which makes use of an authority list, the allowable expressions of the language are specified by enumeration (at least for the most part). Where a language can be described by a complete or near complete enumeration of its allowable expressions it is redundant to also provide a structural, i.e. grammatical description of the language (Possibly an abstract, structural, description would be of use in demonstrating whether a language is to a degree systematic. J. Harris' work with the LCSH might be looked upon as an attempt to reach such a description.) (5) In any case, the LCSH language is predominantly an enumerative language, one which by specification in an authority list, lays down the expressions an indexer must use as headings. Kaiser's Systematic Indexing, on the other hand, is predominantly a synthetic language. It is synthetic in the sense that it provides rules whereby indexers can create new expressions by combining terms. The difference between Cutter's subject heading language, as it developed into LCSH, and Kaiser's Systematic Indexing is huge. It is as huge as the difference existing between describing a language by enumerating all allowable expressions in it and describing this same language by constructing for it a generative grammar, i.e., by postulating a set of formulas or sentence-types (e.g. concrete-Process) which completely specify all possible sentences of the language.

Ranganathan, with his Colon Classification, ushered in the era of synthetic indexing languages. Kaiser is rightly his precursor. He was the first to recognize the usefulness of facets in the construction of expressions in a synthetic index language. While indexing and classification theo-

rists prior to Kaiser busied themselves with classifying terms, the classes they constructed were not properly facets in that they had no syntactic function. As has been shown, Kaiser viewed Systematic Indexing as a two-step procedure, the first step, analysis, being the partitioning of the vocabulary of terms of a given subject into categories or facets - Concrete, Country, Process; and the second step being the combining of these terms, once faceted, into expressions of the language. Just as the syntax of English grammar may be defined with reference to grammatical categories, such as adverbs, verbs, nouns, etc. (or NP, VP etc.) so in index languages which incorporate faceting the syntax is defined in terms of the facet categories. Thus the order of terms in an expression in Kaiser's index language, i.e. in a Statement, is determined by the facet categories, Concrete, Country, and Process, to which the terms are assigned.

7.0 Category definition

We come then to the definition of Kaiser's categories. A convenient way to approach this is to ask two questions: (1) Is there a correspondence between Kaiser's categories and the parts-of-speech categories, noun and verb, used in the classification of natural language words; and (2) to what extent do Kaiser's categories correspond to the grammatical categories of subject and predicate used in the analysis of natural language sentences. We will begin by looking at the parts-of-speech categories.

A *noun* may be defined as:

> any member of a class of words distinguished chiefly by having plural and possessive endings, by functioning as subject or object in a construction, and by designating persons, places, things, states, or qualities:

and a *verb:*

> any member of a class of words that function as the main elements of predicates, typically express action or state, may be inflected for tense, aspect, voice and mood, and show agreement with subject or object.

These definitions are cited by Lyons (in *Semantics)* as being "taken from a particularly good and authoritative dictionary of English (Urdang, 1968) (6). One of Lyon's purposes in citing these definitions is to demonstrate how unfortunately complicated definitions of parts of speech are. Most unfortunate is that they seem to comprise morphological, grammatical and semantic criteria that are potentially noncoincident. As will be seen this is a problem also with Kaiser's categories.

The morphological criteria in the above definitions are "having plural and possessive endings" and "may be inflected for tense, aspect and mood." On a formal level these criteria are not helpful in distinguishing between a concrete and a process, or, for that matter, between the facets in any indexing language. For the most part indexing languages function independent of context. Their vocabularies consist largely of nouns (or nominals) and consequently there is no need for verb markers indicating tense, aspect and mood. Kaiser's process terms, while they "contain verbs" do so in the form of verbal nouns which are indeterminate with respect to inflection.

> The process expresses the action which the concrete is under going or has undergone … Although the process contains the verb it need not necessarily be expressed in the form of a verb so long as it expresses the action … (344).

Kaiser does introduce some morphological conventions. He, for instance, states that "the term of the concrete should always be expressed in the singular, excepting in the case of collections which have no singular, as ironworks, cotton goods, etc. (319). Another use of a morphological convention is in the case of words like *organization* which can name either a concrete or a process. Kaiser suggests the referential ambiguity be resolved by reserving the "tion" ending for the concrete and using the "ing" ending to denote the process (Aslib Report, p. 149).

Generally, however, it would seem that morphological criteria are neither sufficient nor necessary for determining whether a given term is to be classified as a concrete or a process.

Traditionally a simple declarative sentence has been viewed as consisting of two obligatory constituents, a subject and a predicate. Since the time of Plato the subject-predicate distinction has been closely associated with the parts-of-speech distinction between nouns and verbs. Referring back to the grammatical parts of the noun and verb definitions given above, we see that the noun "functions as a subject or object in a construction" and the verb functions as "the main element of a predicate." Kaiser seems to recognize these functions when he says that literature "names things" and that these things are "spoken of":

> From the standpoint of knowledge literature is confined to the description of concretes and of the conditions attaching to them, and for our purposes literature may be analysed into terms of concretes and terms of processes. They are the constant elements with which we have to deal. To put it into the simplest language we may say that literature names things and that these things are spoken of or

> described. The knowledge conveyed by literature all has reference either to things or to spoken or, i.e. concretes and processes. (298)

In this passage Kaiser is clearly distinguishing between the referencing and predicating functions of language. Interestingly enough, in another passage Kaiser wishes to observe that the referencing-predicating distinction is not always identical with the subject-predicate distinction:

> Care should be taken not to confound the two elements concrete and process with subject and predicate. In the sentences "Synthetic indigo is in great demand," "There is a great demand for synthetic indigo," "India suffers a great deal through the manufacture of synthetic indigo" the concrete is *synthetic indigo* whatever its position. (301)

Kaiser is illuminating in his observation that surface structures can be misleading. The point is that the concrete-process distinction is not a surface structure distinction, though often it may, in fact, coincide with the grammatical subject-predicate distinction[13]. It follows from this that a term cannot simply by inspection of its grammatical function be identified as a concrete or a process. What must also be taken into account are contextual clues that indicate whether the term is operating in a referencing or a predicating mode. Context is important is determining whether a term belongs to the concrete or process category.

The semantic part of the definition of verb given above is that a verb expresses an "action or state." In the passage just cited (301), immediately after identifying processes, functionally, with "what is spoken of" Kaiser makes a semantic leap:

> The second term *spoken of* implies an action, i.e. what things do or what is done to them. It must in all cases contain the verb. (301)

Specifying that a process must contain a verb denoting an action results in a fairly narrow definition of Process. Understandably Kaiser does not stay with this narrow definition. In the Aslib Report he interprets processes more generally, allowing states as well as actions to be denoted.

> Similarly the terms of actions or verbs may be supplemented very conveniently by adding those implying a state or condition generally, which terms can also be used for divisions of concretes. Such terms are: Condition, State, Property, Qualification, Industry, Science, Service, Yield, Demand, etc. The two classes of terms i.e., those of actions and those of states, I have called collectively PROCESSES in the sense of dynamic or static conditions of concretes. (p. 149)

Processes, then, are terms which express "dynamic or static conditions of concretes"[14]. Two questions may be asked here. The first is whether this semantic definition can be operationalized to permit the unambiguous identification of terms as process terms. The second is whether the functional and semantic definitions of process terms are coincident - i.e., is everything that may be spoken of x be categorized as a static or dynamic condition of x?

The semantic part of the definition of noun is that a noun is used to designate "persons, places, things, states." We have seen already, in the earlier discussion on problems of reference, that Kaiser tended to limit concretes to a certain subclass of nouns. The world of business, in which his system had its primary application, was a simplified world where concretes, for the most part, could be limited to commodities.

In addition to movable and immovable commodities, Kaiser recognized abstract commodities such as *Labour, mental* and *manual*. The introduction of abstract commodities opens the door to semantic difficulties. Referents begin to lose their grounding. The referent of labour is not a concrete in the sense of being a physical object that can be pointed to, like a battleship. Kaiser justified including energy terms, such as *labour,* as concretes on the grounds that concretes represent latent energy (Aslib Report, p. 149) but surely this was something of a compromise considering his original premise that only those things "definite to handle," viz. concretes, were capable of being classified (108). As was mentioned earlier, terms as abstract as mathematical terms could not be dealt with at all by his system.

Insofar as the concrete and process categories are functionally defined, "what is spoken of" and "what is spoken about," Kaiser is able to maintain a fair distinction. With the categories so defined, the assignment of a term to one or other category cannot be done in isolation but depends rather on the use of contextual information. The trouble comes when Kaiser assumes that the functional distinction is also, neatly, a semantic distinction, a distinction between terms naming concrete objects and those naming conditions attaching to them. With the semantic distinction, there seems to slip in as well the assumption that terms can be categorized independently of context. Indeed in many of his examples Kaiser considers the categories of terms without reference to a context. But then he allows that at times he can not be sure whether a term (one gathers "viewed in isolation") should be assigned to the concrete or the process category: What for instance is the referent of an abstract term like memory? Is it a concrete or a condition attaching to a concrete?

> There may be sometimes doubts or difficulties in deciding whether a given term should be treated as a concrete or as a process … but this does not detract from the obvious advantage of separating sharply these two kinds of terms. In case of doubt we must decide one way or the other and abide by our decision. Thus memory may be taken either as a concrete or as a process according to what standpoint we take. But these cases do not arise generally on the main subjects of a business[15].

Nevertheless Kaiser did worry about category definition. In the Aslib Report he restates his original problem. "Given a vast number of terms; the problem is to divide them into a very small number of classes so that there shall be no overlapping between the classes and yet so that all the terms are completely covered and if any relation can be established between the classes, so much the better." He continues in the same paragraph by saying he hopes still to incorporate mathematical terms in his scheme and "at the same time to make the definitions of concrete and process more precise" (Aslib Report, p. 155).

8.0 Implications

Facet definition as discussed in this paper is of historical interest as it relates to Kaiser. But it bears as well on issues of current interest. Faceting, or the categorization of terms used as subject indicators, is a feature of analytic-synthetic classificatory languages, such as the Colon Classification, and also of modern string indexing languages, such as PRECIS. In the PRECIS indexing language terms are assigned role operators and are thus categorized according to their semantic/syntactic roles, for instance as an agent of a transitive action or the object of such an action. The categories used for faceting in the Colon Classification are the well-known Personality, Matter, Energy, Space and Time. Quite a number of other categories of terms are recognized in special purpose faceted classifications, for instance Substance (product), Organ or Part, Constituent, Structure, Shape, Property, Raw Material, Action, Operator, Process and Agent. Facets used in classificatory languages have associated with them notational indicators as well as natural language indicators of subjects. The use of a notation in fact represents an obvious and perhaps the chief difference between classificatory languages and indexing languages based on synthetic principles and employing a categorization of terms.

What is the purpose of faceting? Why is it worth discussing? The categorization of terms used as subject indicators in a classificatory or indexing language serves a function quite similar to that performed by the parts-of-speech or grammatical categorization of words in a natural language. As was earlier mentioned, words in a natural langnage such as English are viewed as belonging to categories such as noun, verb, adverb, etc. The syntax of English grammar may then be defined with respect to these categories. In an analogous manner the syntax of expressions in a string indexing language may be defined with respect to an initial categorization of terms into facets. For instance, the order of terms in a PRECIS expression follows the ordinal value assigned to each of the role indicators. The "context" of the Preserved Context Indexing System is operationally defined by a citation order, for instance by the following formula: (3) agent of a transitive action; (2) action; (1) object of a transitive action; (0) location. Similarly, the order of elements in a Colon Classification number follows the PMEST formula. Possibly this order represents an Absolute Syntax underlying the order prescribed by other citation principles, such as the "general-before-special" and the "wall-picture" principles.

A less obvious purpose of faceting or categorizing terms used as subject indicators is exemplified in its use in constructing standardized or canonical representations of what a given document is about. "Aboutness" is a matter of concern among indexing theorists dealing with document representation. But this concern is not limited only to indexing theorists. In the area of Artificial Intelligence a basic issue is that of knowledge representation. How is knowledge to be represented in a computer program? The argument is made that standardized representations, which in some way avoid the anomalies of natural language, are required for the various purposes of AI, including the effective retrieval of information. An example of the type of work that is done in this area is that of Ross Quillian. His information retrieval program is based on an analysis of natural language text into two categories or facets, one which has as elements: objects, events, ideals, assertions … the type of thing "which can be represented in English by a single word, noun phrase, or sentence" and the other which has as elements: properties which express predication, "such as might be stated in English by a verb phrase, a relative clause or … a modifier" (8).

While advances are continually being made in computer understanding programs, problems of ambiguity seems so formidable, that one is led to assume that natural language text will have to be normalized in some manner for the purpose of sophisticated information retrieval. Such a normalization would undoubtedly entail the assigning of natural language words to categories or facets, since these would be needed to form the basis of a systematic grammar. What we are talking about here is an artificial language which with less vagary than natural language, can represent the knowledge or information content of documents.

Given that an artificial language is needed for some indexing purposes and for sophisticated methods of infor-

mation retrieval and that such a language must incorporate the faceting of terms, then how these facets are defined becomes a matter of great importance. Unless facets or term categories are defined with some precision, that is, stating explicitly conditions of membership, then the assigning of terms to these categories will depend on intuition, with resulting disagreement, inconsistency and "fudging." In the literature there is some recognition of problems of category definition. For instance, Gopinath writes suggestively about category definition in the Colon Classification:

> Until the publication of CC edition 6, the matter isolates were few. This was because at that time, matter was said to consist usually of materials used for construction, consumption, etc. … However, during the period 1960 to 1966, the developments in the general theory of classification led to the recognition of property isolates as manifestations of matter. A systematic examination of the CC edition 6 schedules for recognizing property isolates led to the realization that a majority of what were enumerated as "energy *cum* personality isolates" - such as "anatomy," "physiology," "disease"- were really property isolates (9).

Another hint of definitional problems is given in the following passage from the *PRECIS Manual* where consideration is given to how names of phenomena should be classified.

> The names of phenomena, more than any other category of terms, establish an indexing language as something which is recognisably different from a natural language. Terms such as "Football," "Diseases" and "Foreign relations" would probably be considered as actions (or, in Ranganathan's terms, as foci belonging to the "Energy" facet) in almost all index languages, yet none of them strictly resembles a verb in the traditional sense … we can be reasonably sure that we are dealing with a phenomenon term if (i) it appears to represent things engaged in an action rather than an action *per se* and (ii) it cannot be reduced to an infinitive (10).

For the most part, however there seems to be a lack of concern about precise category definition among indexing and classification theorists, among those working in Artificial Intelligence and also among linguists (for instance in the definition of case roles). It seems not a little surprising that Kaiser, living and writing at the turn of the century devoted more attention to systematic category definition than writers today who have at easy disposal the tools of modern logic.

The purpose of the present paper has been twofold: to present an historical account of the little-known but quite sophisticated method of indexing developed by Julius Otto Kaiser; and, to focus particularly on Kaiser's attempts at facet definition, with a view to explicating the problems, epistemological as well as definitional, that are involved. The point the paper wishes to make, and of which the historical account is illustrative, is the following: if the categorization or classification of terminology is introduced for a systematic purpose, such as information retrieval, care must be devoted to definitions. Categories must be well defined in the sense that conditions for membership are explicitly stated.

Notes:

1. See Metcalfe (2), p. 297.
2. See J. Kaiser in (3), fifth unnumbered page in the final section of the book entitled "Some opinions of the press."
3. See (3), first unnumbered page.
4. See (3), third unnumbered page.
5. See J. Kaiser (3). A third volume in the Series was intended, "The card system at the factory," but apparently never realized, See the fifth unnumbered page in "Some opinions of the press."
6. See J. Metcalfe (2), p. 298.
7. "Some opinions of the press," in (3), first unnumbered page.
8. Sec (3), Paragraph 45. In the remainder of this paper citations to "Systematic indexing" will be referenced by paragraph number enclosed in parentheses, e.g. (45). (Citations to sources run only until (10), I.C.).
9. See (1), p. 141.
10. Aslib: Report of proceedings of the third conference held at Balliol College, Oxford, Sept. 24-27, 1926, p. 20-33. Reprinted in (1). In the remainder of this paper citations to this report will be referenced by the report name and the reprint page number enclosed in parentheses, e.g. (Aslib Report, p. 154).
11. See (1), p. 141.
12. Ibid.
13. Kaiser's concrete-process distinction would seem to be closer to Hockett's topic-comment distinction than to the predicate-subject distinction - at least insofar as the topic-comment distinction purports to operate at the level of deep structure. See also (7), p. 335.
14. Lyons suggests that the term "situation" be used to cover states on the one hand and events, processes and actions on the other. He further suggests a distinction be made between dynamic and static situations. Semantically, this is close to Kaiser's processes

defined as "dynamic or static conditions of concretes." See (6), p. 483.

15. See (3), paragraph 663 under the heading Concrete and Process.

References

(1) Olding, R. K. (Ed.): Readings in library cataloguing. London: Crosby Lockwood 1966. p. 141.

(2) Metcalfe, J.: Subject classifying and indexing of libraries and literature. New York: Scarecrow 1959. p. 298.

(3) Kaiser, J.: Systematic indexing. (The Card System Series, Vol. II). London: J. Gibson 1911. paragraph 20.

(4) Kaiser, J.: The card system at the office. (The Card System Series, Vol. I). London: Vacher 1908.

(5) Harris, J. L.: Subject analysis: computer implications of rigorous definition. Metuchen, N.J.: Scarecrow 1970.

(6) Lyons, J.: Semantics: 2. Cambridge: Cambridge University Press 1977. p.425.

(7) Lyons, J.: Introduction to theoretical linguistics. Cambridge: Cambridge University Press 1968. p. 334.

(8) Quillian, R.: The teachable language comprehender: a simulation program and theory of language. In: Comm. ACM 12 (1969) p. 459-476.

(9) Gopinath, M. *A.:* Colon Classification. In: Classification in the 1970's. A. Maltby, Ed. London: C. Bingly 1972, p. 71.

(10) Austin, D.: PRECIS: a manual of concept analysis and subject indexing. London: Council of the British National Bibliography 1974. p. 135.