When Might Human Indexing Be Strongly Justified

Julian Warner
Queen's University Belfast, j.warner@qub.ac.uk

Introduction

This paper addresses the issue of a strong justification for human indexing, in an era where computational generation of descriptions of source texts, for direct human interrogation of those descriptions, have proliferated and diffused. We understand human indexing in the classic sense of assignment of terms from a controlled vocabulary by human indexers to source documents, with documents broadly understood. An inclusive model of the structure of index languages, as articulated by Gardin (1973), is adopted, with the development that index terms are understood as gathering together signifieds of the described documents, rather than univocally designating concepts (Warner, 2018). An immediately acceptable and strong justification for such indexing would be when its value exceeds, or is at least commensurate with, its costs. The more specific questions for investigation are, then, of the nature of the value obtainable and the major source for costs.

A long historically established value for retrieval systems is selection power, or the enhanced capacity for informed choice for the searcher (Warner, 2010). It is analogous to bibliographic control, or "mastery over written and published records" (UNESCO/Library of Congress 1950, p.1), but can more comfortably extend to media beyond the written and printed documents implied by *bibliographic* and its reminiscence of the printed document, Byblos, and the bible. Selection power can then be adopted as the underlying value for both human and computational indexing. We can then refine our initial question to, the particular form of value or selection power enabled by human indexing.

The costs of human indexing have been found primarily to reside in the human labour or work involved in description of information objects (Hayes, 2000). Costs identified can be discerned as strongly influential on practice. In particular, the distribution of the products of human description work or labor, as catalogue records and the like, effectively reduces the unit costs to participating institutions of document description. Processes of distribution go back to at least the early 20th century but have gained in intensity, pervasiveness, and volume (Warner, 2010, pp.47-48, 62-65). Costs of searching would also primarily reside in the human labour expended (Warner, 2010).

We can then reformulate our initial assertion as, A strong justification for human indexing would be when the selection power provided exceeds or is commensurate with its costs, with costs recognized as primarily residing in human intellectual labor extended over time. We remain interested in the particular form of selection power likely to be provided by human indexing.

The question of the justification for human indexing can be made more analytically tractable, by taking a long-term historical perspective, which distinguishes different information technology modes. We approach human history on an almost paleontological model of different periods used by another paper at the meeting (Montoya, 2019; see also Childe, 1944).

Historical perspective

The current historical moment follows a technological transition, which can be persuasively understood as a revolution in the mechanization of mental labor (Wentsun, 2002). After such transitions, consciousness may lag behind practice – classically, "The tradition of the dead generations weighs like a nightmare on the minds of the living" (Marx, 1852/1973, p.146) – but questioning of received consciousness may also emerge.

Consciousness, as librarians and information scientists, can be read from seminal post-1945 assertions. Without bibliography, the "records of civilization would be an uncharted chaos of miscellaneous contributions to knowledge, unorganized and inapplicable to human needs" (UNESCO/Library of Congress, 1950, p.viii). At that stage of technological development, the creation of indexes for bibliographic organization required direct human intervention or labor. UNESCO regarded itself as born into "appalling post-war bibliographic chaos" (Murra, 1951, p.47), and the distribution of responsibility to the national agencies required to produce national bibliographies and allied works on a shared model was conceived as both a remedy for the chaos and the path toward universal bibliographic control. Classically, then, human description and indexing was implicitly received as an inescapable necessity. Such consciousness remains strong, although the practice has further transformed, through the global distribution of catalogue records, and, with deeper implications for theory, direct interrogation of computationally generated descriptions.

Questioning of received consciousness in the sense of the need for human description can be found in the work of Patrick Wilson. A late review (White, 2019) asks, "What if the intellectual foundations really were built to justify the limits of old technologies?" and argues that it is "time to start afresh" (Wilson, 2001, p.204). The recognition of the need to question such deeply embedded foundations is analogous to the philosopher F.P. Ramsey's "calmness in infanticide" (Braithwaite, 1930, p.ix), or readiness to discard emerging ideas. It can also be differentiated, as the founding assumption of the received tradition is questioned. We can move towards a fresh start by looking at an information system situated at a previous transition in information technologies, written literacy emerging from orality, to consider what elements continued across the transition, and, by contrast, what is specific to written literacy.

The Icelandic law-speaker provides a transitional form that both inherits elements from orality and anticipates characteristics of written literacy. The law-speaker was required to recite the law, with a rock cliff projecting the voice (see Figure 1. Alþing in session), and to answer queries on legal and parliamentary procedures by oral pronouncements based on his memory of the law (Njal, 1280/1960, p.306–308). Law-speakers are characteristic of oral societies, and the Icelandic law-speaker is of relatively late date and well-documented, with some concurrent and developing elements of written literacy. The law-speaker would

Figure 1. *Alþing in session*. **19th century. W.G. Collingwood.** (Collingwood, 2019)**.**

recite the law to the members of the annual assembly, or Alþing, as a linear spoken utterance, reciting one third of the laws at each meeting (Short, 2008). From a modern perspective, the law-speaker could be regarded as an information system embodied in a single, socially designated individual.

Selection power, which we adopted as our primary value, is evidenced in dialogic questioning and response, with evidence for the possibility of questioning separately from the recitation of the law. For instance, in *Njal's Saga*, the law-speaker is consulted for confirmation of an aspect of the law.

> Flosi asked if this were the law, but Eyjolf replied that he did not know for certain and said that the Law-Speaker would have to settle that point. Thorkel Geitisson went on their behalf and told the Law-Speaker the situation, and asked if there were any legal basis for Mord's submission.
>
>   "There are more great lawyers alive today than I thought," replied Skapti. "I can tell you that this is so precisely correct that not a single objection can be raised against it. But I had thought that I was the only person who knew this specialty of the law now that Njal is dead, for to

the best of my knowledge he was the only other man who knew it." (Njal 1280/1960, p.308).

Selection labor can be discovered in the law-speaker's mental work of memorization and recall, the communicative labor of recitation, and in the attention of the auditors. Technology is manifested in a natural object adapted to a human purpose, the rock face used as a sounding board for the voice of the law-speaker.

> All those things which labour merely separates from their immediate connection with their environment are objects of labour spontaneously provided by nature, such as fish caught and separated from their natural element, namely water, timber felled in virgin forests, and ores extracted from their veins. (Marx 1867/1976, p.284)

Selection power, selection labor, and technology are discernible in an information system emerging from primary orality (Ong, 1982). They have then persisted across oral, written literate, and computational modes, indicating their centrality to information retrieval (Warner, 2010, pp.26-28).

Crucially, and by contrast, there is no direct analogue to human description practices in the form of products of description labor or metadata, strongly suggesting that the human assignment of metalanguages is a historically specific development of written literacy.

An extended historical perspective on information technologies indicates, then, that metadata or humanly assigned descriptions are not an ahistorical necessity. In modern practice, searching on words and phrases directly and computationally derived from documents has become the accepted norm, coupled with the ability to impose various computationally possible orders on documents retrieved. Consciousness and theory has, to date, lagged behind practice.

We can bring consciousness into accord with practice, by reformulating the question of the justification for human indexing, reversing, recent in terms of technological transitions, historical order and transforming the question into a more specific and analytically tractable consideration.

What forms of selection power are *not* readily obtainable from established searching facilities, for written text.

Human indexing would then be justified when it yields a value – a particular form of selection power – not readily obtainable from searching on words and phrases. The term *readily* implies an absence of binary opposition between description and searching and the possibility of interchange between them. Our exposition is primarily concerned with retrieval from documents in written language (with examples drawn from English language), but the argument will admit of expansion.

Creativity	Example	Searching	Description
Level of creativity	Oranges (as fruit)	Selection of words for their intended meaning.	Classically only for certain culturally central texts, *The Bible* and *The Koran* .
		Elimination of unintended retrievals: lesser level of creativity.	Generic categorisation (oranges as fruit rather than as telephone company).
	Oranges and lemons Slave and labor	Connection with words related in meaning.	
	"so mechanical or routine"	Selection of phrases for their intended meaning (reduction in unintended recalls).	Classically only for certain culturally central texts.
	"so mechanical or routine" AND "mechanical procedure"	Conjunction of words and phrases with which a connection is desired.	Some, more technical, facilities in description and searching.

Figure 2. Comparison of searching and description

Values obtainable from searching

Searching and description can be considered in parallel. First mention and analysis is given to searching, to accord with its priority in our guiding question. We then ask, for each widely diffused searching facility, how could or were these values be supplied by human description. We will restrict our attention to selection rather than ordering, for focus and the analytic clarity.

Searching facilities can be considered in broadly ascending levels of creativity required in searching (see Figure 2). The term *creativity* points to the possibility of joining things together, a classic sense of creativity. The analysis then

also has potential legal significance for intellectual property in the products of searching and of description, given the globally influential United States decision, *Feist v. Rural* (Feist, 1991), which required a minimal degree of *creativity* for copyrightability. Creativity for the decision has recently been elucidated as noncomputable human activity, or labor, directly motivated by engagement with meaning (Warner, 2013).

A widely diffused form of searching, involving an initial and everyday level of creativity, is the selection of words for their intended meaning, for instance, *orange* as fruit. This form of selection could extend to words similar in expression (for instance, *orange*, *oranges,* and *orangeade*). Historically, under written literacy, description to enable this form of searching involved human (which could be predominantly clerical rather than directly intellectual) labor of extended duration. Costs of such labor in description ensured that direct selection of words in searching was generally only possible for certain culturally central texts, such as *The Bible* or the *Qu'rân.* Concordances were also distinguished from indexes (Gardin, 1973, p.140). The virtually ubiquitous and generally expected provision of such searching facilities for wide ranges of written documents implies that modern readers are in a privileged position historically previously largely reserved to specialized scholars, in their capacity for selection and recall. An enhancement of human capacities therefore emerges.

The subsequent elimination of unintended recalls, after the selection of words and documents by their intended meaning, can be understood as a lower level of creativity. Such elimination is a feature of everyday experience. The generic categorization offered by humanly assigned index languages (for instance, oranges within the category *fruit* rather than as a telephone company) can reduce the number of unintended recalls (it may also have the effect of excluding material of possible relevance, given the uncertainty of labeling decisions). A synthesis of the generic power of human indexing and the possibility of specificity in modern and current searching is possible and daily enacted in practice, where specific searching takes place with the generic scope offered by humanly assigned terms.

A higher level of creativity in searching would be making connections between words related in meaning, such as oranges and lemons or slave and labor, with an explicit or implied Boolean AND or OR connecting the terms. Classically, such connections were made by humanly assigned indexing terms, for instance *fruit* for oranges and lemons. A slightly hidden, but potentially significant, contrast is that for humanly assigned terms in description connections had to be named, whereas in modern practice connections need not be specifically named, but can remain implied or felt. Full exploitation of the modern potential to connect words related in meaning may require deliberate study of the language of recalled documents and could include words not related in expression but with potentially associated meanings (consider *slave* and *labor*). Technically, and more than technically, the association in meaning between such terms is not computable, in

the sense of a generally applicable procedure for determining similarity or connection and may then require direct human mental labor, motivated by meaning.

A further level of creativity, more novel as a searching possibility, is selection of phrases for their intended meaning, with the strong possibility of a marked reduction in unintended recalls, compared to single word searching (Warner, 2010). A revealing example of this would be searching for the phrase, 'so mechanical or routine' which characterizes the opposite or antithesis to creativity in the *Feist* decision (Feist, 1991, p.361; Warner, 2013). Classically, such forms of searching were only enabled for certain culturally central texts and required a substantial amount of human labor in description. A further enhancement of human capacities for recall can then be obtained.

A remaining, and thereby possibly the highest, level of creativity required for searching, is the conjunction of phrases to detect or recall combinations of desired concepts. A further development from the previous example would be to search for the combination 'so mechanical or routine' AND 'mechanical procedure', to determine whether the antithesis to creativity has been connected with the idea of a mechanical procedure. Historically, such conjunctions may have been possible by carefully exploiting technical facilities in description and searching. The particular conjunction would once have recalled no documents, implying the absence of a publicly made connection (alternatives to mechanical procedure, such as algorithm and Turing (machine), were also used, in the actual searches). Conjunction of 'so mechanical or routine' AND 'mechanical procedure' now exists and could be recalled by a *Google* search. However, its existence is almost exclusively a consequence of a previously not made connection having been published in the primary literature (Warner, 2013). A generalizable implication of the development of this example over time is that we can search for unprecedented connections, whereas human description or indexing was limited to connections which were known to exist in advance. Classically, index languages were understood to lag behind the language of the primary literature and we could further extend a limitation of human indexing to connections *well* known in advance.

We can review the contrasts made between forms of selection power offered by current searching and the historically inherited, and continuing, forms of power given by description. Classically, the now everyday capacity to recall documents by individual words and their combinations existed only for certain culturally central texts. There has been an enhancement and privileging of the human capacity for selection by the revolution in the mechanization of mental labor. Searching for a word or phrase can enable selection of highly specific meanings. What may be less immediately obtainable from searching is connections between the specific meanings isolated. A strong instance of such connection would be between expressions which are related in meaning (such as *slave* and *labor*) but which have no special, computationally detectable, similarity as expressions. Within such relations, we could distinguish those knowable in advance (such as oranges and apples as forms of fruit) from those not known in advance. For those

relations knowable in advance, we can, and do, label them in description, often resulting in relations which can be semantically characterized as genus to species (consider labelling oranges and apples as fruit). Humanly labelling relations in this way in description may reduce labour expended in searching. Those relations not known in advance cannot, almost tautologically, not be labelled in description, corresponding to the established understanding that index languages were likely to lag behind the primary literature.

As an overall implication, the contrasts strongly and overwhelmingly indicate that the continuing and current value of human indexing is for generic categorization. As such, it can function as a complement to the specificity immediately and directly obtainable from modern forms searching. In addition, a strong degree of congruence between generic categorization in indexing and in navigational structures is revealed. A significant further implication is that categorizations have to be known in advance.

A gradation within selection power has then been obtained, which provides a rationale for when human indexing may be strongly justified. To summarise, such indexing is justified when it provides selection power which would otherwise require prolonged and reiterated human labor in searching. The particular form of selection power provided may often involve making connections between information objects, with the connections understandable as generic descriptions.

The analysis can be further deepened, by recognizing that selection power is a *use* value whose costs of production reside primarily in the *exchange* value of the human intellectual labor in its production and exploitation. The exchange values connected with the direct human labor in description and searching are likely far to exceed the exchange values connected with the computational operations for description and searching (computational operations are themselves understood as a product of human labor). Human labor could be understood as immediately exchanged between description and searching, or, more fully satisfactory theoretically, as indirectly exchanged through a medium of exchange. We can then arrive at a concluding statement of the justification for human indexing.

> **Human indexing is strongly justified, when the exchange value involved in producing its use value (likely to be realized as generic power) are commensurate with the exchange value it can command.**

We can simultaneously recognize the enhancement of human capacities for selection offered by widely diffused modern search facilities.

Theoretical extension and derivation
We can further suggest that the justification established for human indexing may be a final or teleological formulation, by indicating its possibility of extension to non-written documents, theoretical derivation, reminiscence of theory for searching prior to the current technological revolution, and congruence with practice.

The exposition has been restricted to searching written documents but can be expanded. Selection of words or phrases across documents corresponds to selection by pattern, or *syntax* in a strict sense, of sequences of characters regarded as equivalent to one another. The sequences are characteristically selected for their potential semantic significance, with semantic understood as direct human engagement with meaning, complementing syntax in its strict sense. The argument can then also apply to non-written documents, although, for such documents there may be different, less computationally tractable, relations between pattern or syntax and potential semantic significance.

The conclusion has emerged from, although it is not dependent on, a labor theoretic approach to information retrieval. The central proposition, that selection power is produced by selection labor, or in terms of formal logic, Selection power -> Selection labor (which can be read as, IF Selection power THEN Selection labor) (Warner, 2010, pp.54-55) is retained. The further development of the model can then be re-ordered, to accord with classic and ordinary discourse understandings of categorization, which demand that the first division be made by the characteristic with the most significant discriminatory power (as in the game of twenty questions (see Shannon, 1968, p.215)). Semantic human labor is counterposed to syntactic machine process (rather than first dividing by description and searching). Semantic labor and syntactic machine process can then both be divided by description and searching. In the reformulation of the model, semantic labor in searching and description are brought into closer proximity with each other and possibility of interchange or exchange between them becomes more directly observable and understandable (see Figure 3). Although the argument has emerged from a modification of the labor theoretic approach, it has been conducted independently of that approach, and is thereby not dependent upon the approach.

Selection power   ⟶   Selection labor

Semantic labor   Syntactic machine process

Searching   Description   Searching   Description

Possibility of
Interchange
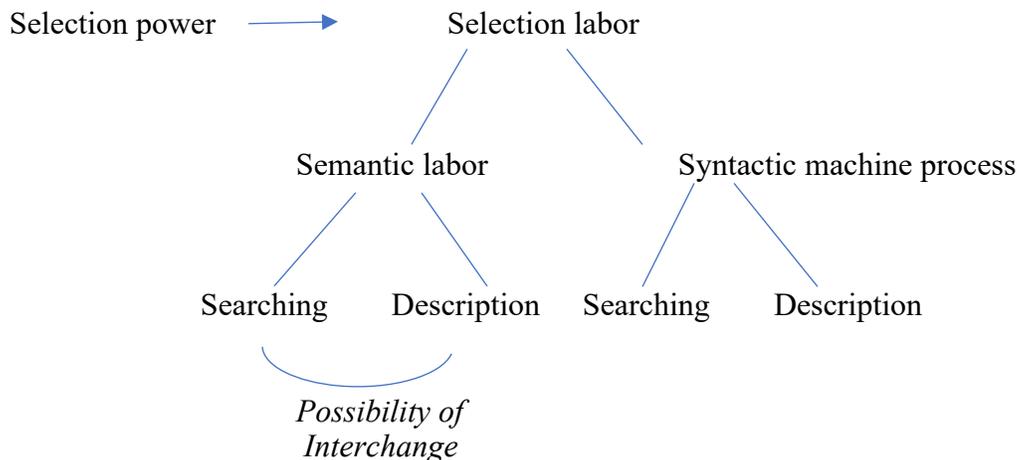
Figure 3. A labor theoretic approach to information retrieval reformulated.

The argument is also reminiscent of classic and widely accepted theories. Gardin's (1973) analysis and description of indexing languages is retained, although the *value* of human indexing and searching has been transformed, bringing the overall analysis into accord with emerging modern practice. The absence of binary opposition between description and searching and the recognition of the possibility of interchange between them is reminiscent of a classic component of information science, of Bradford's (1948/1971, pp.103-164) conception of literature scattering. The distinction between relations established in advance and those which can be newly created corresponds to the value attached to discovery of knowledge, in rather neglected library and information science literature of the 1970s (see, for instance, Watson, Gammage, Grayshon, Hockey, Jones, and Oldman, 1973). The conclusion of the argument also has a correspondence to the recent elucidation of a minimal degree of creativity for the ownership of intellectual property as residing in human labor directly engaged with meaning (admittedly emerging from a common basis in the distinction of semantic from syntactic labor) (Warner, 2013). The re-emergence of human indexing as the residue of what is difficult to accomplish purely computationally is consistent with other domains of human activity connected with computation. We have begun to address Norbert Wiener's injunction, "Render unto man the things which are man's and unto the computer the things which are the computer's." (Wiener, 1964/1966, p.73). Reminiscences and correspondences can be understood as triangulation effects supporting the argument and are also indicative of its significance.

Generic indexing, as deductively predicted from theory, is also strongly congruent with what can be inductively gathered from practice, where forms of indexing which can be understood as generic labelling predominate. The market for indexing has acted as a selection mechanism, in both bibliographic (Swanson, 1980, p.128), and, by extension, broader domains. An explanation for congruence could be found in the possibility of reading the justification given for human indexing as a summary statement of the conditions for survival of services using such indexing, in a competitive market, where costs must be covered by returns.

The conclusion of the argument has then been supported by reminiscence of classic theory and its strong correspondence to current practice.

Conclusion

To return to our opening considerations, taking an approach informed by a long historical perspective, distinguishing different information technological eras, of primary orality, written literacy, and the computational mode, enabled formulation of a novel, radically simple and analytically tractable, consideration, what forms of selection power can *not* be readily obtained from modern searching facilities (and may therefore need to be provided by indexing or description). Our answer could be progressively formulated, was finally simple, admitted of extension to non-written media, recalled classic theories, and is congruent with globally emerging practice.

We have significantly validated and extended the perspective indicated by Wilson (2001, p.204), by revealing a humanly assigned metalanguage to be a development of written literacy, extending into the computational mode with significant residual value. In contrast to widely diffused indexing practices under written literacy, our selection power, particularly for syntactically defined forms of potential semantic significance, has been greatly enhanced. Selection power, or the capacity and responsibility to choose, may be fundamental to our human being outside of Eden, and by acknowledging and developing this we can become deliberately and consciously more fully human.

References

Bradford, S.C. (1948/1971). *Documentation*. First published 1948. 2nd edition. High Wycombe, Bucks., University Microfilms for the College of Librarianship Wales.

Braithwaite, R.B. (1931). Editor's introduction. In F.P. Ramsey. *The foundations of mathematics and other logical essays* (pp.ix-xiv). Edited by R.B. Braithwaite. London: Routledge & Kegan Paul.

Childe, V.G. (1944). Archæological ages as technological stages. *Journal of the Royal Anthropological Institute. 74 (1-2)*, 7-24. doi 10.2307/2844291.

Collingwood, W.G. (2019). *Althing in Session* (painting). From Wikipedia. Retrieved from https://en.wikipedia.org/wiki/File:Law_speaker.jpg.

Feist. (1991). *Feist Publications, Inc. v. Rural Tel. Service Co., Inc.* 499 U.S. 340 (1991).

Gardin, J-C. (1973). Document analysis and linguistic theory. *Journal of Documentation. 29 (2)*, 137-168. doi 10.1108/eb026553.

Hayes, R.M. (2000). Assessing the value of a database company. In B. Cronin and H.B. Atkins (Eds). *The web of knowledge: a festschrift in honor of Eugene Garfield* (pp.73-84). Information Today, Inc.: Medford, NJ.

Marx, K. (1852/1973). The eighteenth Brumaire of Louis Bonaparte. In K. Mark. *Surveys from exile: political writings* (Volume 2, pp.143-249). Edited by D. Fernbach. Harmondsworth: Penguin Books in association with New Left Review.

Marx, K. (1867/1976). *Capital: a critique of political economy, volume one*. Introduced by E. Mandel and translated by B. Fowkes. London and New York: Penguin Books in association with New Left Review, 1976.

Montoya, R. D. (2019). Trace documents and impressions of knowledge. Paper given at Document Academy 2019, 12th June 2019.

Murra, K. O. (1951). History of Some Attempts to Organize Bibliography Internationally. *In* J. H. Shera and M. E. Egan (Eds). *Bibliographic organization: papers presented before the fifteenth annual conference of the Graduate Library School July 24–29, 1950* (pp.24–53). Chicago: University of Chicago Press.

Njal (1280/1960). *Njal's saga*. Translated with an introduction by M. Magnusson and H. Pálsson. Harmondsworth: Penguin Books, 1960.

Shannon, C.E. (1968/1993). Information theory. In C.E. Shannon. *Collected papers* (pp.212-220). Edited by N.J.A. Sloane and Aaron D. Wyner. Piscataway, NJ: IEEE Press.

Short, W. R. (2008). Norse Laws and Legal Procedures. Retrieved from http://www.hurstwic.org/history/articles/society/text/laws.htm.

UNESCO/Library of Congress Bibliographic Survey (1950). *Bibliographical services, their present state and possibilities of improvement. Report*

prepared as a working paper for an international conference on bibliography. Washington, DC.

Warner, J. (2010). *Human information retrieval* (History and Foundations of Information Science Series). Cambridge, MA: MIT Press.

Warner, J. (2018). Feeding two wolves: the human and the computational in document analysis. In W. M. El Hadi(Ed.). *Fondements épistémologiques et théoriques de la science de l'information-documentation* (Actes du 11e colloque ISKO France 2017) (pp.349-357). London: ISTE Editions.

Watson, L.E., Gammage, P., Grayshon, M.C., Hockey, S., Jones, R.K., & Oldman, D. (1973). Sociology and Information Science. *Journal of librarianship. 5 (4)*, 270-283. doi 10.1177/096100067300500405.

Wen-tsun, W. (2002). Some reflections on the mechanization of mental labor in the computer age. MM Research Preprints, No. 21, December 2002, pp.1–5. At https://pdfs.semanticscholar.org/cc7c/7772a2ddd3ff659dbc6cff329da781afb419.pdf.

White, H.D. (2019). Patrick Wilson. In *ISKO encyclopedia of knowledge Organization*. Edited by B. Hjørland and C. Gnoli. At https://www.isko.org/cyclo/wilson.

Wiener, N. (1964/1966). *God and golem, inc. a comment on certain points where cybernetics impinges on religion*. Cambridge, Massachusetts: MIT Press.

Wilson, P. (2001). Review of E. Svenonius. The intellectual foundations of information organization. *College and Research Libraries,* Vol. 62 (2001), pp. 203–204. doi 10.5860/622203.