SISA—Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules

Isidoro Gil-Leiva

University of Murcia, Faculty of Commnunication and Informacion Science,
Campus de Espinardo, Murcia, Spain 30100, <isgil@um.es>

Isidoro Gil-Leiva holds a PhD in library and information science and is Associate Professor in the Information Science Studies department at the University of Murcia. His main research interests are indexing, automatic indexing, controlled vocabularies and knowledge organization.

Gil-Leiva, Isidoro. 2017. "SISA—Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules." *Knowledge Organization* 44(3): 139-162. 123 references.

Abstract: Indexing is contextualized and a brief description is provided of some of the most used automatic indexing systems. We describe SISA, a system which uses location heuristics rules, statistical rules like term frequency (TF) or TF-IDF to obtain automatic or semi-automatic indexing, depending on the user's preference. The aim of this research is to ascertain which rules (location heuristics rules or TF-IDF rules) provide the best indexing terms. SISA is used to obtain the automatic indexing of 200 scientific articles on fruit growing written in Portuguese. It uses, on the one hand, location heuristics rules founded on the value of certain parts of the articles for indexing such as titles, abstracts, keywords, headings, first paragraph, conclusions and references and, on the other, TF-IDF rules. The indexing is then evaluated to ascertain retrieval performance through recall, precision and f-measure. Automatic indexing of the articles with location heuristics rules provided the best results with the evaluation measures.

Received: 14 November 2016; Revised: 24 February 2017; Accepted: 7 March 2017

Keywords: automatic indexing, SISA, scientific articles, terms, controlled vocabulary, information

1.0 Introduction

Cognitive process is the term used to refer to the mental processes performed by rational beings for the selective reception of information, its symbolic coding, storing and retrieval. Cognitive psychology studies cognitive processes like sensorial perception of information, learning (language, reading and writing), memory or reasoning capacity. In Gil-Leiva (2008, 17-54) we indicate that a simultaneous interactive succession of mental processes unfolds in the production of keywords production, indexing terms or subject headings for a document or for an information need during indexing, and that these have to do with the following:

– Perception: the information to be analyzed can arrive at the indexer via three routes—sight, hearing and touch.
– Communication organization: a) textual discourse (with aspects of interest to us such as the text, textuality criteria, structure of the text or types of text); b) oral discourse; and, c) visual discourse
– Memory: sensory memory, short-term memory and long-term memory.
– Comprehension: capturing and connecting ideas, constructing the idea hierarchy, recognizing the pattern of relations between the ideas produced by the textual structure.

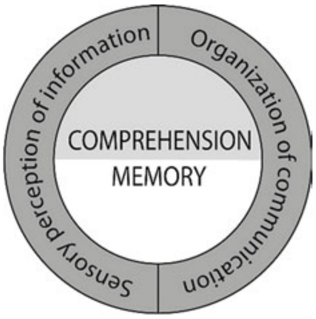

Sensory perception of information
Organization of communication
COMPREHENSION
MEMORY

Figure 1. Cognitive elements present in indexing.

Indexing has been widely studied and there are some valuable contributions on both its theory and practice, among them Frohmann (1990), Lancaster (1991), Farrow (1991), Fugmann, (1993), Hjørland (1997), Anderson and Perez-Carballo (2001), and Mai (2000). ISO norm 5963-1985 defines indexing as "The act of describing or identifying a document in terms of its subject content." To this one can add that, on occasion, concepts are normalized and controlled by controlled vocabulary, as otherwise it would be natural language indexing, and likewise that indexing is carried out—be it consciously or unconsciously—according to the users' information needs in order to convert these (in natural or controlled language) into a search query. Hence, it is an essential process for storing documents and may also be so in the retrieval of information if the result of the indexing (keywords, descriptors, subjects indexing) is used later for retrieval.

Since the end of the 1950s, efforts have been in progress to automate indexing and there has been a substantial amount of research into this, as is borne out in Pulgarín and Gil-Leiva (2004), who, with no claim to being exhaustive, were already handling 839 bibliographical references in a bibliometric analysis of automatic indexing, and the literature has continued to grow noticeably since then, with studies on automatic keyphrase or keyword extraction, for classifying information, data mining, text summarization or information retrieval.

The terminology used in the literature to refer to the process of making indexing automatic is varied; we find names like "automated assisted indexing," "automated indexing," "automated supported indexing," "automatic support to indexing," "computer aided indexing," "computer assisted indexing," among others, amounting to a score or so in total, although the most used is "automatic indexing." The definition of automatic indexing can derive from three perspectives (Gil-Leiva, 2008, 320): a) computer programs that assist in the process of storing indexing terms, once obtained intellectually (computer aided indexing during storage); b) systems that analyze documents automatically, but the indexing terms proposed are validated and published—if necessary—by a professional (semi-automatic indexing); and, c) programs without any further validation programs, i.e., the proposed terms are stored directly as descriptors of that document (automatic indexing).

The methodologies used to automate indexing through the decades have changed. In the early days, indexing documents was done almost exclusively from statistics based on term frequency, but from the 1980s on, techniques like natural language processing to get the roots of words (stems), morphological taggers and syntactic parsers began to be incorporated, along with others. It is, though, usual for the proposals or prototypes submitted by researchers to include a combination of both approaches, i.e., calculating the frequency and more or less complex tools for the automatic processing of texts.

Since the end of the 1950s, the amount of scientific information available has grown tremendously in almost all areas of knowledge, but especially in the experimental sciences. More operational information systems were called for as the amount of research into the treatment of information grew in order to attend to scientists' informational needs more quickly and more efficiently. The idea of the personal computer being a highly useful tool for text processing, especially indexing, spread, since the computer was seen as being objective in repeated operations. The aim was to avoid a center's indexing the same document in different ways at different times or that two indexers might represent the same document in different terms. Thus, for these reasons and because of the greater availability of machines capable of alphanumeric digital processing, the automatic analysis of texts was to become an area of research that continues to this day.

Steven (1965), Sparck Jones (1974) and Liebesny (1974) wrote reports on the state of the art of automatic indexing that are of interest concerning the research situation in the late 1960s and early 1970s. In Gil-Leiva and Rodríguez Muñoz (1996), we systematized the main tasks performed in automatic indexing since the late 1950s, structuring them basically according to statistical and language systems. The first proposals for statistical models were founded on Zipf's Law (Zipf 1949), which states that in most languages a small number of words are used with high frequency while a great number of words are used rarely. So, starting from the term frequency (TF), research was undertaken on the basis of statistical computations for the determination of indexing terms (Luhn 1957a; Luhn 1957b; Damerau 1956); probability computations to determine the terms most appropriate for the representation of a document (Maron and Kuhns 1960; Rosenberg 1971; Bookstein and Sweanson 1974), which gradually lost out to inverse document frequency (IDF) (Sparck-Jones 1972); term discrimination models (Salton, G. and Yang, C. S. 1973; Salton 1974; Salton, Yang and Yu 1975; or Salton, Wu and Yu 1981); or vector space models (Salton, Wong and Yang 1975). From 1961, Gerard Salton, initially at the Harvard University, and later at the Cornell University continued to experiment with all the above in the development, implementation and evaluation of SMART (System for the Mechanical Analysis and Retrieval of Text).

Deerwester et al. (1988), with their Latent Semantic Indexing (LSI) method for extracting and representing the contextual-usage meaning of words by statistical computations, also opened up new strands of research, like Probabilistic Latent Semantic Analysis (PLSA), which is an alternative to LSI proposed by Hofmann (1999) and based on the fact that documents generate a particular distribution of aspects (topics) and that aspects generate a particular distribution of word usage. Later, Blei, Ng and Jordan (2003, 996) took the aforementioned proposals as their reference and sought to enhance them in their Latent Dirichlet Allocation (LDA), a generative probabilistic model of a corpus where the "documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words."

The SMART project by Salton (1980, 1991) was one of the first to incorporate advances produced by the automatic processing of natural language. It brings in tools to extract word roots, thesauri, morphological or syntactic analyzers. These contributions gradually made their way into the proposals of other researchers, where they were used along with frequency computations to select the best noun phrases or the indexing terms themselves. Examples are Trubkin (1979), Maeda (1980), Dillon and Gray (1983) or Faraj et al. (1996), among others. So, for example, CLARIT (Computational-Linguistic Approaches to Indexing and Retrieval of Text), is a system that uses a lexicon for general English which consists of approximately 100,000 root forms and hyphenated phrases, tagged for syntactic category and irregular morphological variation, a morphological analyzer, a lexical disambiguator, a multistage parser, a noun phrase grammar and various indexing algorithms such as the ranking indexing terms. It measures terms using statistical parameters of frequency, distribution and linguistic distinctiveness (Evans et al. 1990, 1991a, 1991b). In the SIMPR (Structured Information Management: Processing and Retrieval) project, output from the morphological analyzer is used to provide the input to indexing software from which some indexing terms are finally obtained and validated manually (Karetnyk, Karlsson and Smart 1991).

In the early 1990s, there appeared some automatic indexing proposals that were grounded in expert systems, e.g., Martinez, Lucey and Linder (1987), Driscoll et al. (1991) or Schuegraf and Bommel (1993). Similarly, Faraj and colleagues (1996) used syntactic document analysis (processing of editing marks, lemmatization, lexico-syntactic labeling to apply frequency computations). Lahtinen (2000) combined linguistic techniques with a variant of term frequency-inverse document frequency (TF-IDF). Elsewhere, the identification of noun phrases via morphological and syntactic tools rather than direct identification of keywords, and the selection of the best noun phrases as indexing terms using statistical methods with IDF or Okapi BM25 has also attracted the attention of researchers (Souza 2005 ; Souza and Raghavan 2006; or Souza and Raghavan 2014). Finally, Joorabchi and Mahdi (2013, 2014), in their automatic subject indexing proposal, seek alternatives to controlled vocabulary use through Wikipedia.

Below, we offer a brief review of a selective rather than exhaustive nature, of the systems or prototypes of automatic indexing, developed, to a greater or lesser extent, in the large information and documentation centers such as the prototypes of the International STN in Karlsruhe (Germany) for chemistry and physics documents (Biebricher et al. 1988), the SAPHIRE (Semantic and Probabilistic Heuristic Information Retrieval Environment) project in the field of biomedicine and the Medline database (Hersh and Greenes 1990; Hersh et al 1991), the automatic indexing programs and projects of the National Library of Medicine "Indexing Initiative," continued with Medical Text Indexer (MTI), known for a wealth of publications such as Humphrey and Miller (1987), Humphrey (1999), Aronson et al. (2000), Humphrey (2006) or Mork et al. (2014) to cite but a handful. It is worth noting here that Hodge (1992) carried out an interesting study on the state of the issue of automating indexing in public and private institutions, and Moens (2000), in turn, made a review of the techniques and methodologies used to create automatic indexing systems.

To continue our review with the development of prototypes from certain centers of information, we would mention here Machine-Aided Indexing (MAI) developed at the NASA Center for Aerospace Information (Klingbiel 1973; or Silvester, Genuardi and Klingbiel 1994), while there are three initiatives of note from the National Agricultural Library: 1) CAIT (Computer-Assisted Indexing Tutor), a program which sought to enhance the quality of indexing and the training of new indexers (Irving 1997); 2) the acquisition of Luxid through TEMIS in 2011, an automated indexing software (Prada, et al. 2011), and more recently the AgNIC (Agriculture Network Information Collaborative) initiative by Salisbury and Smith (2014); 3) the HEP indexer system for indexing high energy physics documents at the European Laboratory for Particle Physics, CERN, Geneva (Montejo Ráez 2001); 4) the *Catalogue et index des sites médicaux de langue* (CISMeF), a French system implemented in the automatic indexing medical information resources (Chebil et al. 2012); and, 5) El-Haj et al. (2013) used the KEA (Keyphrase Extraction Algorithm) system in the UK Data Archive to set the bases for a future automatic indexing system.

Finally, with respect to the problem statement of this paper, it is worth noting that, first, there has been a lot of research into automatic indexing, as was made clear above, as well as by the review included herein. Secondly, in the design and development of automatic indexing systems for scientific papers there is less use of location heuristics rules than of rules or statistical algorithms, in particular TF-IDF. In order to bridge this gap, our research question is: If SISA can use location heuristics rules and statistical rules, which of these techniques offers better terms for indexing a specific collection of documents? Answering this question is important to ascertain the scope of these techniques. Similarly, it enables us to detect their strengths and weaknesses and possibly to extend the methodological approaches in automatic indexing.

2.0 SISA

Most of the early prototypes of automatic indexing worked with scientific articles on the basis of their titles (Scheele 1983), abstracts (Salton 1972; Seo 1993; Wan et al. 1997; Hmeidi, Kanaan and Evens 1997; or Bordoni and Pazienza 1997) and both titles and abstracts (Klingbiel and Rinker 1976; Meulen and Janssen 1977; Barnes, Castantini and Perschke 1978; Roberts and Souter 2000), although they soon began to use the whole text (Andreewsky and Ruas 1982; Haller 1983).

The conceptual development of SISA (Sistema para la Indización Semi-Automática or Automatic Indexing System for Scientific Articles) began in the mid 1990s (Gil-Leiva 1997; Gil-Leiva 1999) with the indexing of information science articles. To determine which information to capture and use we undertook some research (Gil-Leiva and Rodríguez Muñoz 1997) into the values of the titles and abstracts of scientific articles as sources of indexing terms in librarianship and documentation, medicine, chemistry, biology, psychology and physics. We analyzed 450 articles and a total of 2,077 descriptors from the Spanish databases ISOC, IME e ICYT, of the Consejo Superior de Investigaciones Científicas. Of the 2,077 descriptors assigned to the 450 registers, 792 (38.1&) appeared in the title or the abstract, or in both, so 61.9% were in neither, leading us to conclude that it is necessary to use the whole text in automatic indexing. Later, we moved on (Gil-Leiva and Alonso Arroyo 2005; Gil-Leiva and Alonso Arroyo 2007) and studied the use that had been made of keywords provided by authors in scientific articles to ascertain their role these played in indexing these articles for professional indexers. Our findings showed that keywords provide indexers with valuable information and so keywords were included in the sources of capture and valuation in subsequent versions of SISA. Some more recent pieces of research that have looked into the usefulness of authors' keywords for indexing, retrieval, journal editors or social tags are Ansari (2001), Hartley and Kostoff (2003), Craven (2005), Gbur and Trumbo (1995), Strader (2009), Kipp (2011), Smiraglia (2013), Lu and Kipp (2014), Vrkic (2014) or Tanijiri et al. (2016).

Since its appearance, SISA has been used in teaching automatic indexing in PhD and masters' courses, and we are now carrying out tasks to evaluate the system.

2.1 SISA description

Below we give a description of SISA. Figure 2 offers an overall view of SISA with the main modules and processes carried out during semi-automatic indexing (represented by the date in grey and by the letter s) and automatic indexing (represented by the date in black and the letter a).

- Web platform: SISA is available on the Internet to users with a password.
- Language: SISA can currently index documents in Spanish, English and Portuguese (see Figure 3).
- Formats: The formats admitted are PDF, TXT, HTML or XML.
- Stopwords: SISA uses stopword lists for Spanish, English and Portuguese. Although the literature had already indicated this, when working with SISA with articles written in Spanish, we confirmed that half the words in the texts are empty words (articles, prepositions, locutions, etc.) that have no information load and therefore should be eliminated for automatic indexing tasks (Table 1).

	Kilobytes	Words total	Stopwords total	Stopwords (%)
Text 1	49	7806	3952	50.6 %
Text 2	35	5542	2926	52.7 %
Text 3	28	4512	2357	52.2 %
Text 4	10	1479	765	51.7 %
Text 5	31	4827	2241	46.4 %
Text 6	54	8295	4040	48.7 %
Text 7	50	7532	3884	51.5 %
Text 8	25	3772	1589	42.1 %
Text 9	31	4870	2639	54.1 %
Text 10	41	6320	3228	51.0 %
Total	354	54955	27621	**50.2 %**

Table 1. Proportion of words of the articles and stopwords.

The elimination of empty words was applied by H.P. Luhn (1957a) for automatic indexing and it later became a constant feature of most of the systems designed. The inclusion of a stopword list in a prototype requires a detailed analysis since a word may have various meanings. A word may be considered to be empty (without any subject information) in one field but not in another. The

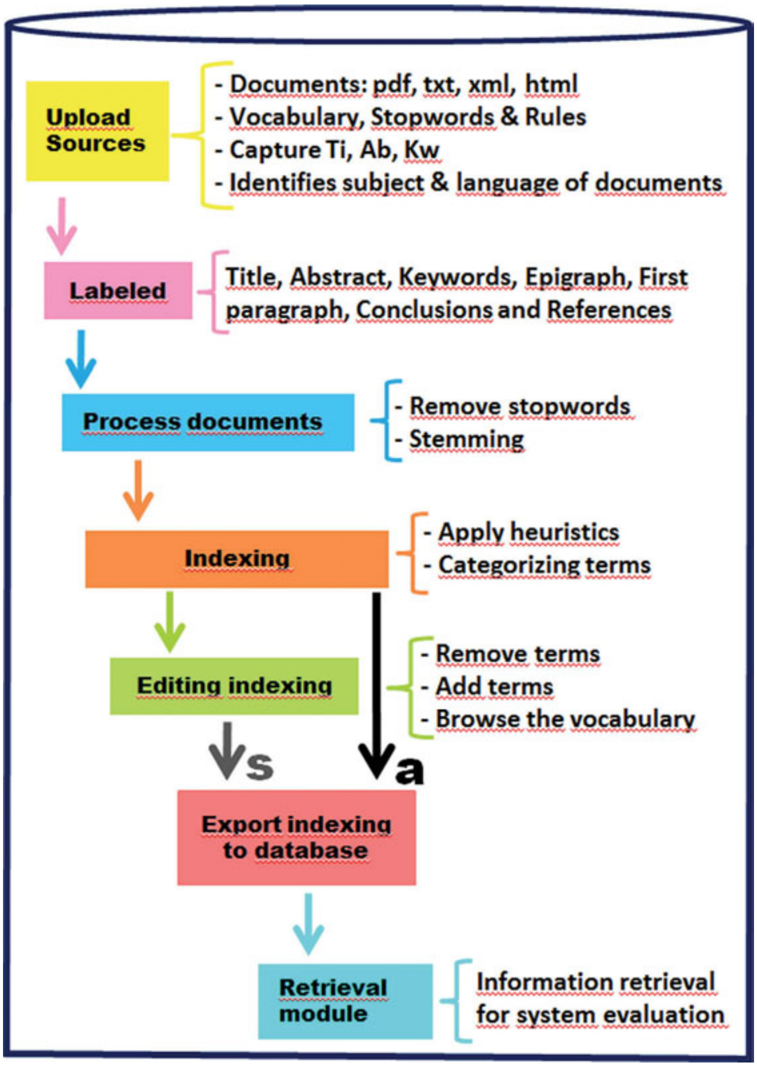

Upload Sources
- Documents: pdf, txt, xml, html
- Vocabulary, Stopwords & Rules
- Capture Ti, Ab, Kw
- Identifies subject & language of documents
Labeled
Title, Abstract, Keywords, Epigraph, First paragraph, Conclusions and References
Process documents
- Remove stopwords
- Stemming
Indexing
- Apply heuristics
- Categorizing terms
Editing indexing
- Remove terms
- Add terms
- Browse the vocabulary
s
a
Export indexing to database
Retrieval module
Information retrieval for system evaluation

Figure 2. Representation of the main modules and processes of SISA.

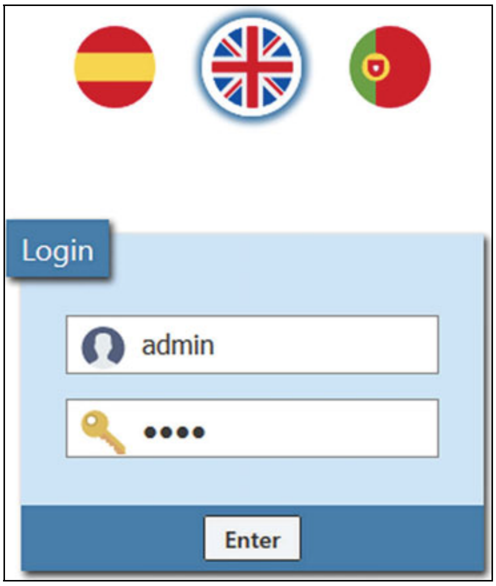


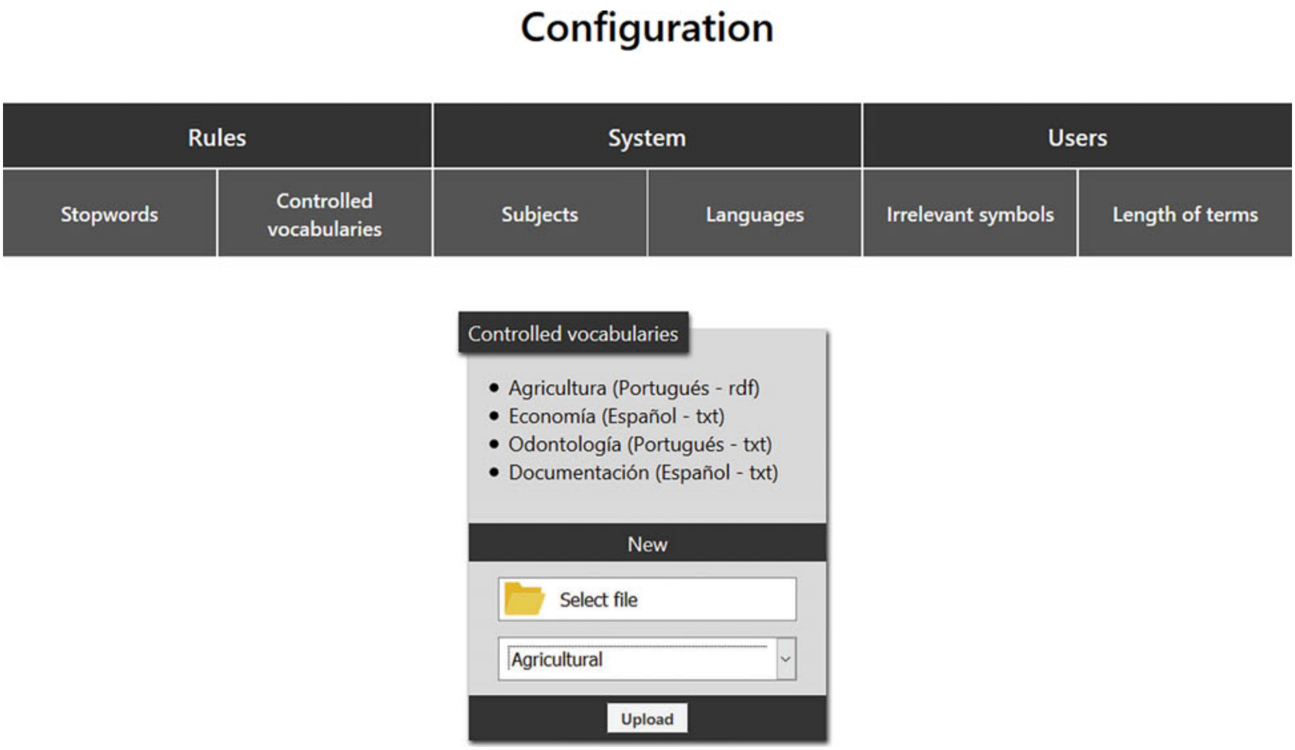
Figure 3. Languages in SISA.

word "account" for example could be found in texts on any subject matter as it is a component of the conjunction "on account of," (e.g., "The match was postponed on 'account' of the bad weather." However, in a text on economics the word "account" could not be considered as a stopword (e.g. "The company lost its most important account.").

– Controlled vocabulary: The use of controlled vocabulary forms part of the original idea of SISA since we considered it to be a useful tool in automatic indexing (Figure 4), even though when we began its design and implementation there was an open debate as to whether the moment had come to forsake thesaurusi with the advent of the Internet and databases which made whole texts available and all the possibilities that this implied for information retrieval. The "thesaurus yes, thesaurus no" debate has become a drawn out affair and there seems to be no end in sight. Indeed, this journal (*Knowledge Organization*, 2016, vol. 43, no. 3) devoted a special issue to the debate and offered different opinions, perspectives and possible uses of thesauri. Dextre Clarke (2016) offers an analysis of this debate. In any case, it is a fact that thesauri or lists of descriptors with only synonym relationships are finding their way into archives and keepers of archives are resorting to them more and more, something which a few decades back would have been unthinkable given the scarce use made of them, especially in our field. Yet today they have even gained a stronghold in e-government for electronic document management.

The use of more or less complex controlled vocabularies in automatic indexing is a fact. Some examples, although by no means all, are: Strode (1977), Valle Bracero and Fernández García (1983), Biebricher, et al. (1988), Silvester, Genuardi and Klingbiel (1994), Gil-Leiva (1997, 1999), Plaunt and Nogard (1998), Aronson et al. (2000), Steinberger, Hagman and Scheer (2000), Lukhashevich and Dobrov (2001), Montejo (2001), Zha and Hou (2002), Beheshti (2003) or Kolar (2005), Medelyanand Witten (2005, 2008), El-Haj, et al. (2013), Willis and Losee (2013), who draw on no fewer than four thesaurusi (AGROVOC, HEP, NALT and MeSH) to evaluate their algorithm, Pickler and Ferneda (2014), Vlachidis and Tudhope (2016) or Dehghani (2015).

Regarding the debate on use of thesauri, we would say, in conclusion, in the future we will continue to need these tools (thesauri, simple or complex lists of descriptors, ontologies with their knowledge inference and others) that help to capture during automatic treatment of the documents the basic conceptual relations like synonymy, hierarchy or proximity (regardless of the type). So, whatever name we might give these tools, their importance lies in being able to use them.



- Agricultura (Portugués - rdf)
- Economía (Español - txt)
- Odontología (Portugués - txt)
- Documentación (Español - txt)

Agricultural

Figure 4. Screen to load the different tools used by SISA.

SISA is able to use a controlled vocabulary in relation to synonymy through USE (Rural workers USE Agricultural laborers) and hierarchy is managed through BT (Violinists BT Instrumentalists) or through a list of descriptors related solely to synonymy.

– Rules: SISA makes use of various rules to evaluate their algorithm (Figure 5). On the one hand are the location heuristics rules, which simulate the intellectual tasks of human indexers, and are of the type:

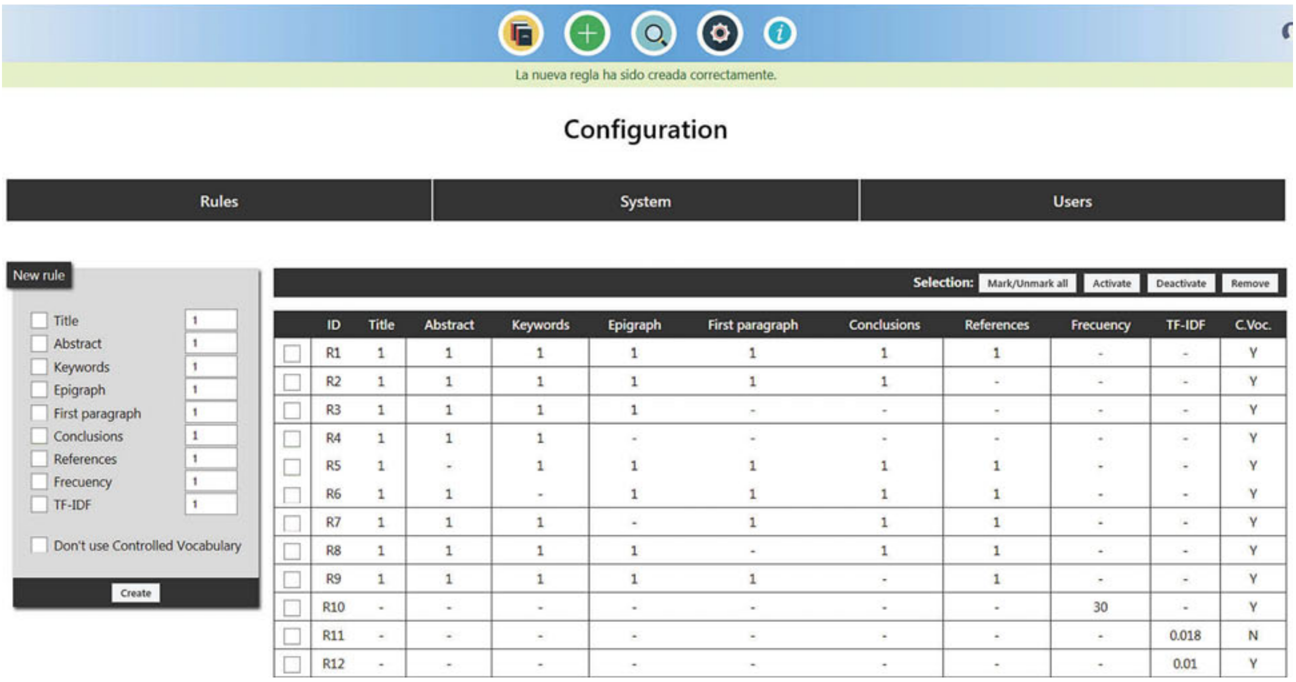
```
IF
term W is in the controlled vocabulary

AND
term W is in the Title, Abstract, Keywords,
Heading, First paragraph, Conclusions and
References

THEN
term W is assigned to the document
```

This rule would be the most demanding of all possible location heuristics rules. From here, there are numerous combinations but avoiding rules like first paragraph and references, since these may not proffer terms representative of the document. But SISA also uses statistical rules, like term frequency (TF), which can fix the minimum number of times a word must appear to be selected, or the TF-IDF which proposes as indexing terms those that exceed an established threshold. Location rules and statistical rules can also be combined. Similarly, location heuristics rules and statistical rules can be combined with controlled vocabulary. These combinations are easy to configure, as Figure 5 shows, where there is a combination of location heuristics rules (R1-R9) and statistical rules like TF with controlled vocabulary (R-10), rule TF-IDF without controlled vocabulary (R-11) and rule TF-IDF with controlled vocabulary (R-12).

– Labeling: SISA uses labels to mark the different parts of a scientific article on the basis that certain parts of articles provide valuable information for indexing. When it was conceived, it used labels for the beginning and end of titles, abstracts, text and paragraphs. Later editions have incorporated keywords, headings, conclusions and references. The labels comprise the initials of words in Spanish for these parts of an article (Table 2).

Positions	**Labels**	
	beginning	end
Title	#ITI#	#FTI#
Abstract	#IRE#	#FRE#
Keyword	#IPC#	#FPC#
Heading	IEP#	#FEP#
First paragraph	#IPP#	#FPP#
Conclusions	#ICO#	#FCO#
References	#IRF#	#FRF#

Table 2. Labels used by SISA.


La nueva regla ha sido creada correctamente.

Configuration

Rules | System | Users

New rule

☐ Title 1
☐ Abstract 1
☐ Keywords 1
☐ Epigraph 1
☐ First paragraph 1
☐ Conclusions 1
☐ References 1
☐ Frecuency 1
☐ TF-IDF 1

☐ Don't use Controlled Vocabulary

Create

Selection: Mark/Unmark all | Activate | Deactivate | Remove

	ID	Title	Abstract	Keywords	Epigraph	First paragraph	Conclusions	References	Frecuency	TF-IDF	C.Voc.
☐	R1	1	1	1	1	1	1	1	-	-	Y
☐	R2	1	1	1	1	1	1	-	-	-	Y
☐	R3	1	1	1	1	-	-	-	-	-	Y
☐	R4	1	1	1	-	-	-	-	-	-	Y
☐	R5	1	-	1	1	1	1	1	-	-	Y
☐	R6	1	1	-	1	1	1	1	-	-	Y
☐	R7	1	1	1	-	1	1	1	-	-	Y
☐	R8	1	1	1	1	-	1	1	-	-	Y
☐	R9	1	1	1	1	1	-	1	-	-	Y
☐	R10	-	-	-	-	-	-	-	30	-	Y
☐	R11	-	-	-	-	-	-	-	-	0.018	N
☐	R12	-	-	-	-	-	-	-	-	0.01	Y

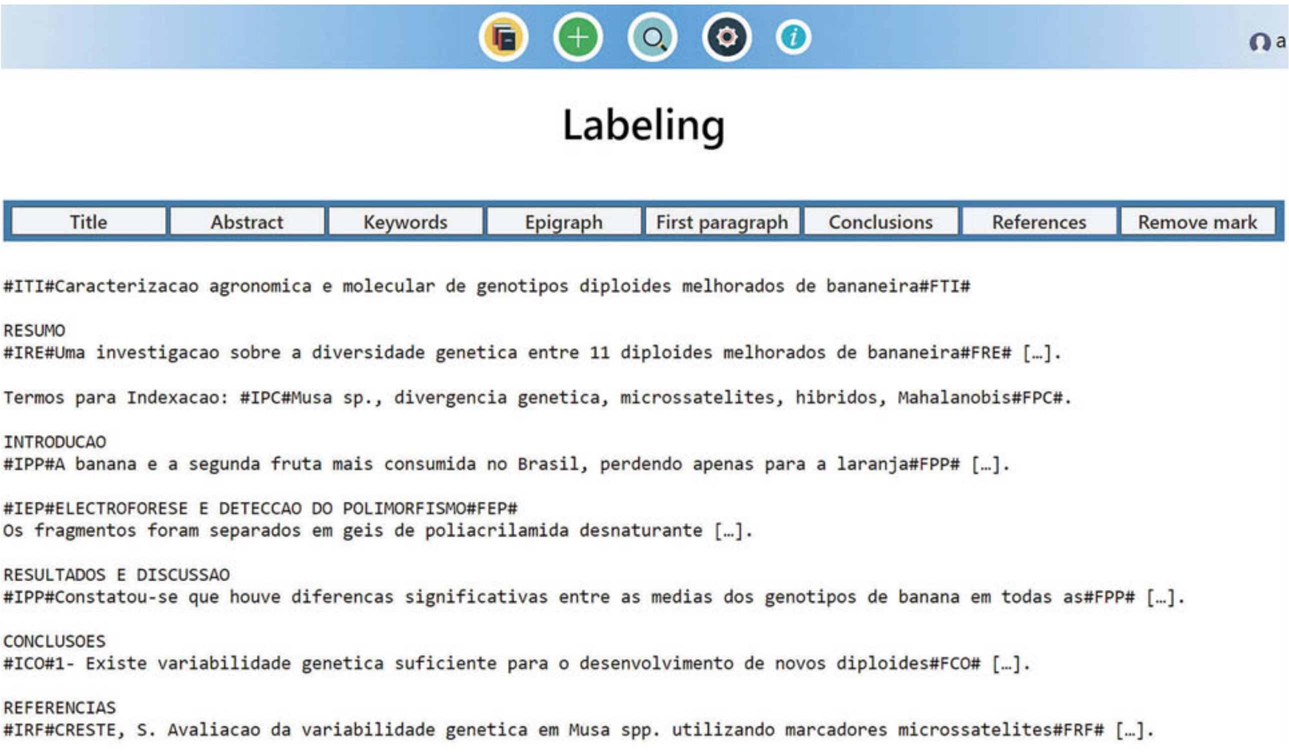
Figure 5. Module to configure rules.

Labeling is performed once the article has been uploaded to SISA or outside the system using a text editor. In the SISA labeling model, the labels are fixed quickly and simply. One only needs to select the information to be labeled and then click on the relevant label (Figure 6). SISA can currently automatically detect different parts of articles published in XML by the *Revista Española de Documentación Científica* and in html for articles in the journal *Information Research: An International Electronic Journal.*

- Automatic recognition of language and subject matter: Once the controlled vocabularies have been uploaded, along with the corresponding stopword lists according to the document languages and several articles in the system (these can be uploaded individually or in batches of several hundred), SISA automatically detects the language and subject matter of the documents, and the indexing process can begin.
- Stemmer: SISA uses the Snowball stemmer algorithm to control for the gender and number of words so that these can be considered as a single item (child and children in English, *menina* and *menino* in Portuguese or *niña* and *niño* in Spanish).
- Database: After indexing the document, the metadata titles, abstract, keywords and the A descriptors (SISA descriptors) are stored in the database. Later the B descriptors (a "gold indexing" by expert indexers or from another automatic system for comparison with SISA) can be inserted in each and any of the stored records.
- Information retrieval: SISA has its own information retrieval module (Figures 7 and 8) for browsing metadata from stored records for evaluation tasks such as that performed for Souza and Gil-Leiva (2016) when comparing SISA with PyPLN or that made here to determine evaluation measures which will be described below.
- SISA (Semi-Automatic Indexing or Automatic Indexing): From the very first version SISA has offered the possibility of editing indexing results by adding terms not proposed or removing erroneous ones (Figures 9, 10 and 11).

Candidate terms were also proposed in the early versions of SISA. These are terms not included in the controlled vocabulary, neither were they stopwords, and they fulfilled requisites like appearing a minimum number of times and in different paragraphs. Thus, the SISA indexing did not depend solely on the presence or absence of a controlled vocabulary term and it was also possible to perform an automatic vocabulary feedback.

A new feature was added to the SISA semi-automatic indexing mode in the second version (2004), with the enhancement of term editing by browsing the controlled vocabulary to assign descriptors not initially proposed by SISA (Figure 10).



```
#ITI#Caracterizacao agronomica e molecular de genotipos diploides melhorados de bananeira#FTI#

RESUMO
#IRE#Uma investigacao sobre a diversidade genetica entre 11 diploides melhorados de bananeira#FRE# […].

Termos para Indexacao: #IPC#Musa sp., divergencia genetica, microssatelites, hibridos, Mahalanobis#FPC#.

INTRODUCAO
#IPP#A banana e a segunda fruta mais consumida no Brasil, perdendo apenas para a laranja#FPP# […].

#IEP#ELECTROFORESE E DETECCAO DO POLIMORFISMO#FEP#
Os fragmentos foram separados em geis de poliacrilamida desnaturante […].

RESULTADOS E DISCUSSAO
#IPP#Constatou-se que houve diferencas significativas entre as medias dos genotipos de banana em todas as#FPP# […].

CONCLUSOES
#ICO#1- Existe variabilidade genetica suficiente para o desenvolvimento de novos diploides#FCO# […].

REFERENCIAS
#IRF#CRESTE, S. Avaliacao da variabilidade genetica em Musa spp. utilizando marcadores microssatelites#FRF# […].
```

Figure 6. Labeled article from the module labeling.

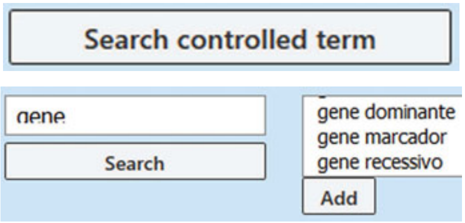
Icon	Legend
P1	Term found in: - Title, abstract, keyword, heading and first paragraph - Title, abstract, keyword, conclusions and references
P2	Term found in: - Title, abstract and keyword - Abstract, keyword and heading - Abstract, keyword and conclusions - Keyword, heading and conclusions
P3	Term found in: - Title and abstract - Abstract, heading and conclusions
✏	Terms are edited (removal of incorrect terms, browsing of the whole text and the controlled vocabulary, and the inclusion of terms put forward by the indexer or those located in the controlled vocabulary. (Semi-automatic indexing).
✖	Removes terms selected incorrectly.
manga	Clicking on the terms browses the whole text to decide on the term's suitability (Figures 8 and 9)
Add new term	Adds a new term.
 Search controlled term gene Search gene dominante gene marcador gene recessivo Add	Browse the controlled vocabulary to add a term, in this case, "gene."

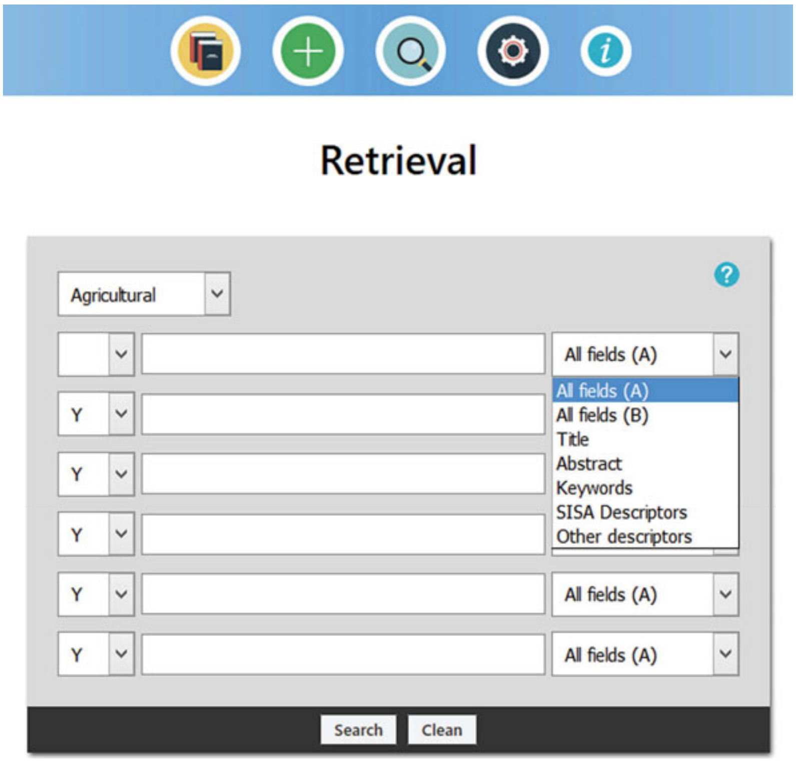
Table 3. The operation and options available in SISA.

Clicking on the candidate terms enables the user to browse the whole text of the article with the term clearly marked in red, making it easy to analyze the context in which it appears (Figure 11).

So, if the user opts for one of these editing possibilities, SISA becomes semi-automatic indexing. If, on the other hand, the terms are stored directly, it is automatic indexing.

– Techniques used for the application development: The initial version of SISA dates from 2002 and was implemented in Java. The second version, in Delphi, appeared in 2004 and the third version came out in 2013, using web technologies like Java (JPA, Servlet and JSP) and JavaScript as programming languages to create interactive pages; Asynchronous JavaScript + XML (AJAX), which is a set technologies like HTML, CSS, DOM, XML, XSLT, JSON, XML HttpRequest and JavaScript, which work together to create interactive applications; Cascading Style Sheets for the design and presentation of SISA; Document Object Model (DOM), application programming interface which means that languages like JavaScript can access the content of an HTML website; Tomcat, as web server with servlets support and JSPs for the deployment of servlet based web applications developed in Java; Web Ontology Language (OWL) for work with the controlled vocabulary; Snowball to extract the root of a word, which includes the Porter algorithm; and MySQL as a database management system. Since 2013, a number of improvements have been added to both the processes and interface of SISA.
– Hardware: SISA is installed on a Proliant server with 32GB RAM ML310E and a CentOS 7.0 operating system.

Following the introduction and description of SISA, it should be noted that within the conception and design of

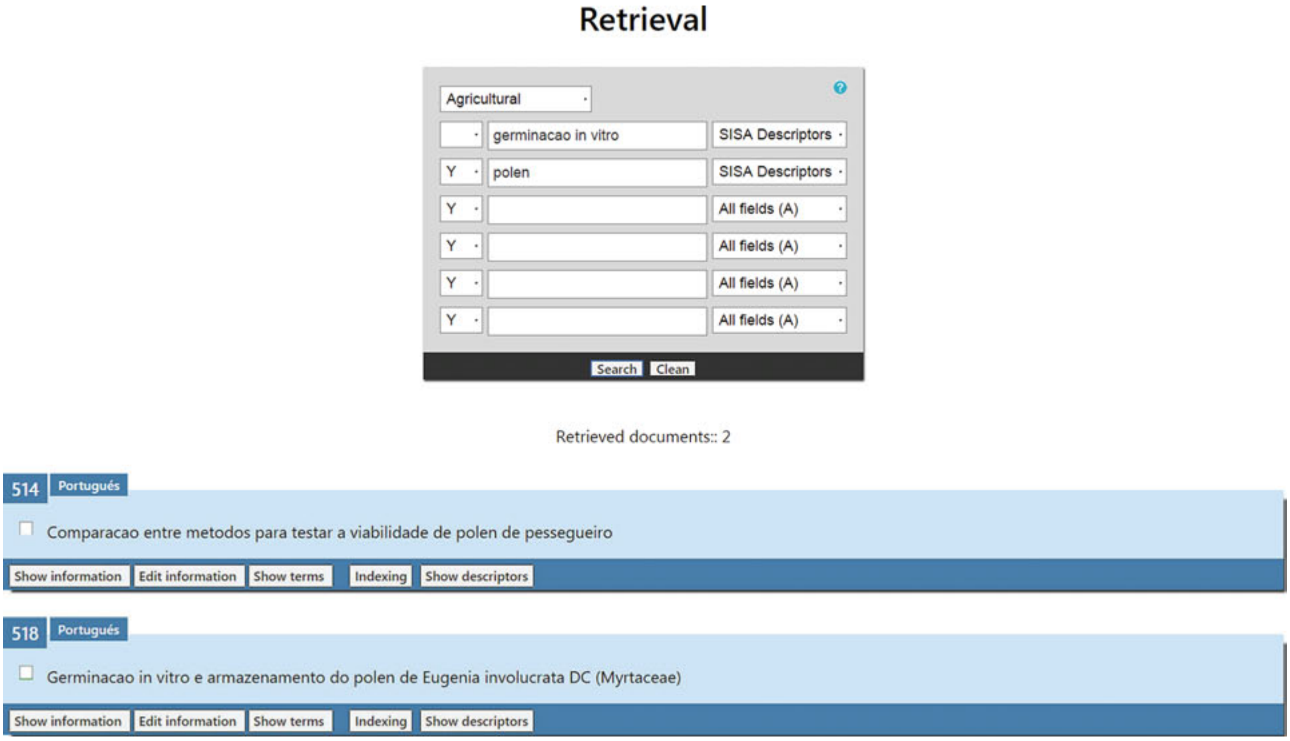


Retrieval

Agricultural

All fields (A)

Figure 7. Search options.



Retrieval

Agricultural

	germinacao in vitro	SISA Descriptors
Y	polen	SISA Descriptors
Y		All fields (A)
Y		All fields (A)
Y		All fields (A)
Y		All fields (A)

Retrieved documents:: 2

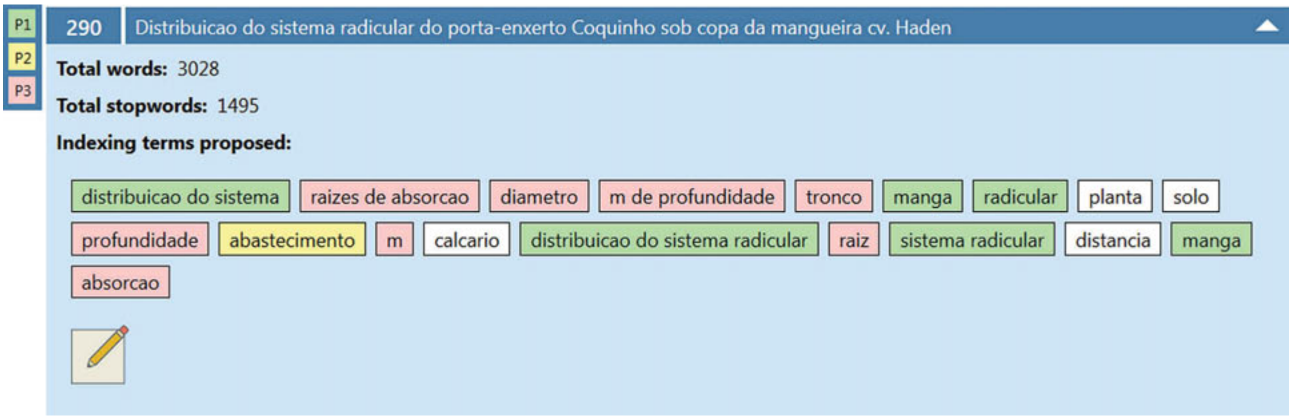
514 Portugués

□ Comparacao entre metodos para testar a viabilidade de polen de pessegueiro

518 Portugués

□ Germinacao in vitro e armazenamento do polen de Eugenia involucrata DC (Myrtaceae)

Figure 8. Search query about "germinação in vitro" AND "polen" in Descriptors field.



290 Distribuicao do sistema radicular do porta-enxerto Coquinho sob copa da mangueira cv. Haden

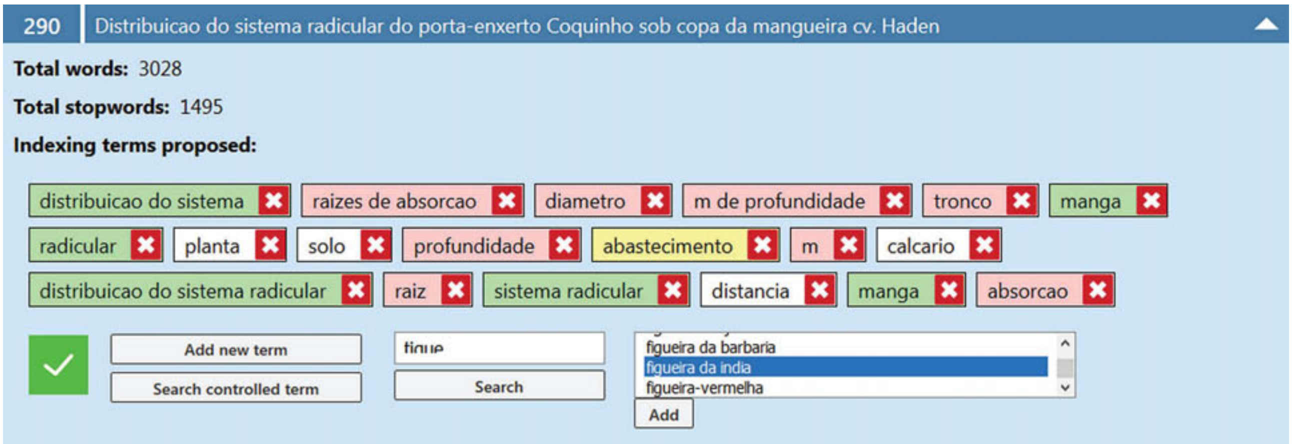
Total words: 3028

Total stopwords: 1495

Indexing terms proposed:

distribuicao do sistema | raizes de absorcao | diametro | m de profundidade | tronco | manga | radicular | planta | solo | profundidade | abastecimento | m | calcario | distribuicao do sistema radicular | raiz | sistema radicular | distancia | manga | absorcao

Figure 9. Indexing editing.



290 Distribuicao do sistema radicular do porta-enxerto Coquinho sob copa da mangueira cv. Haden

Total words: 3028

Total stopwords: 1495

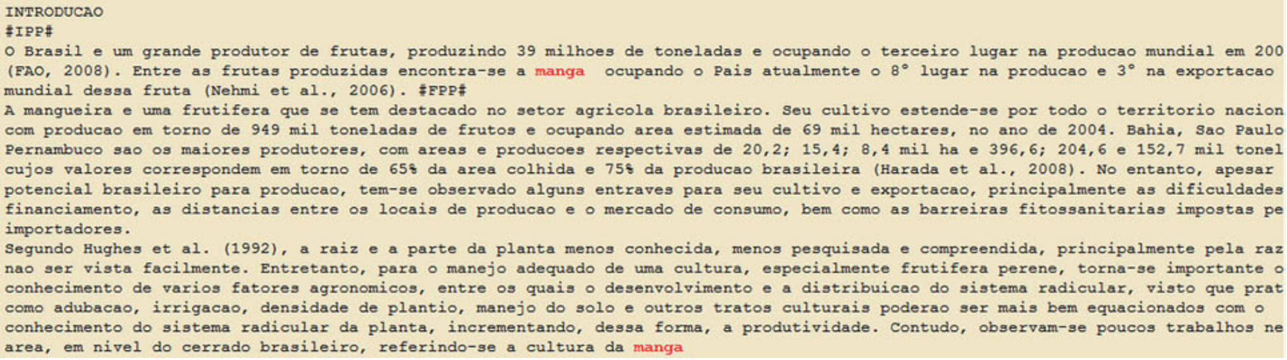
Indexing terms proposed:

distribuicao do sistema | raizes de absorcao | diametro | m de profundidade | tronco | manga | radicular | planta | solo | profundidade | abastecimento | m | calcario | distribuicao do sistema radicular | raiz | sistema radicular | distancia | manga | absorcao

figue

figueira da barbaria
figueira da india
figueira-vermelha

Figure 10. Editing options and browse the controlled vocabulary.



```
INTRODUCAO
#IPP#
O Brasil e um grande produtor de frutas, produzindo 39 milhoes de toneladas e ocupando o terceiro lugar na producao mundial em 200
(FAO, 2008). Entre as frutas produzidas encontra-se a manga  ocupando o Pais atualmente o 8° lugar na producao e 3° na exportacao
mundial dessa fruta (Nehmi et al., 2006). #FPP#
A mangueira e uma frutifera que se tem destacado no setor agricola brasileiro. Seu cultivo estende-se por todo o territorio nacion
com producao em torno de 949 mil toneladas de frutos e ocupando area estimada de 69 mil hectares, no ano de 2004. Bahia, Sao Paulo
Pernambuco sao os maiores produtores, com areas e producoes respectivas de 20,2; 15,4; 8,4 mil ha e 396,6; 204,6 e 152,7 mil tonel
cujos valores correspondem em torno de 65% da area colhida e 75% da producao brasileira (Harada et al., 2008). No entanto, apesar
potencial brasileiro para producao, tem-se observado alguns entraves para seu cultivo e exportacao, principalmente as dificuldades
financiamento, as distancias entre os locais de producao e o mercado de consumo, bem como as barreiras fitossanitarias impostas pe
importadores.
Segundo Hughes et al. (1992), a raiz e a parte da planta menos conhecida, menos pesquisada e compreendida, principalmente pela raz
nao ser vista facilmente. Entretanto, para o manejo adequado de uma cultura, especialmente frutifera perene, torna-se importante o
conhecimento de varios fatores agronomicos, entre os quais o desenvolvimento e a distribuicao do sistema radicular, visto que prat
como adubacao, irrigacao, densidade de plantio, manejo do solo e outros tratos culturais poderao ser mais bem equacionados com o
conhecimento do sistema radicular da planta, incrementando, dessa forma, a produtividade. Contudo, observam-se poucos trabalhos ne
area, em nivel do cerrado brasileiro, referindo-se a cultura da manga
```

Figure 11. Browsing the text.

our tool there are traditional elements that are found in earlier automatic indexing prototypes proceeding from the elimination of stopwords, a stemming algorithm, the use of controlled vocabularies, the calculation of a term's frequency or the inverse document frequency. The significance of our proposal though lies in the configuration of a series of heuristics (rules) based on the position in which the concepts appear in the documents and which can be combined easily amongst themselves as well as with statistical criteria.

3.0 Materials and methods

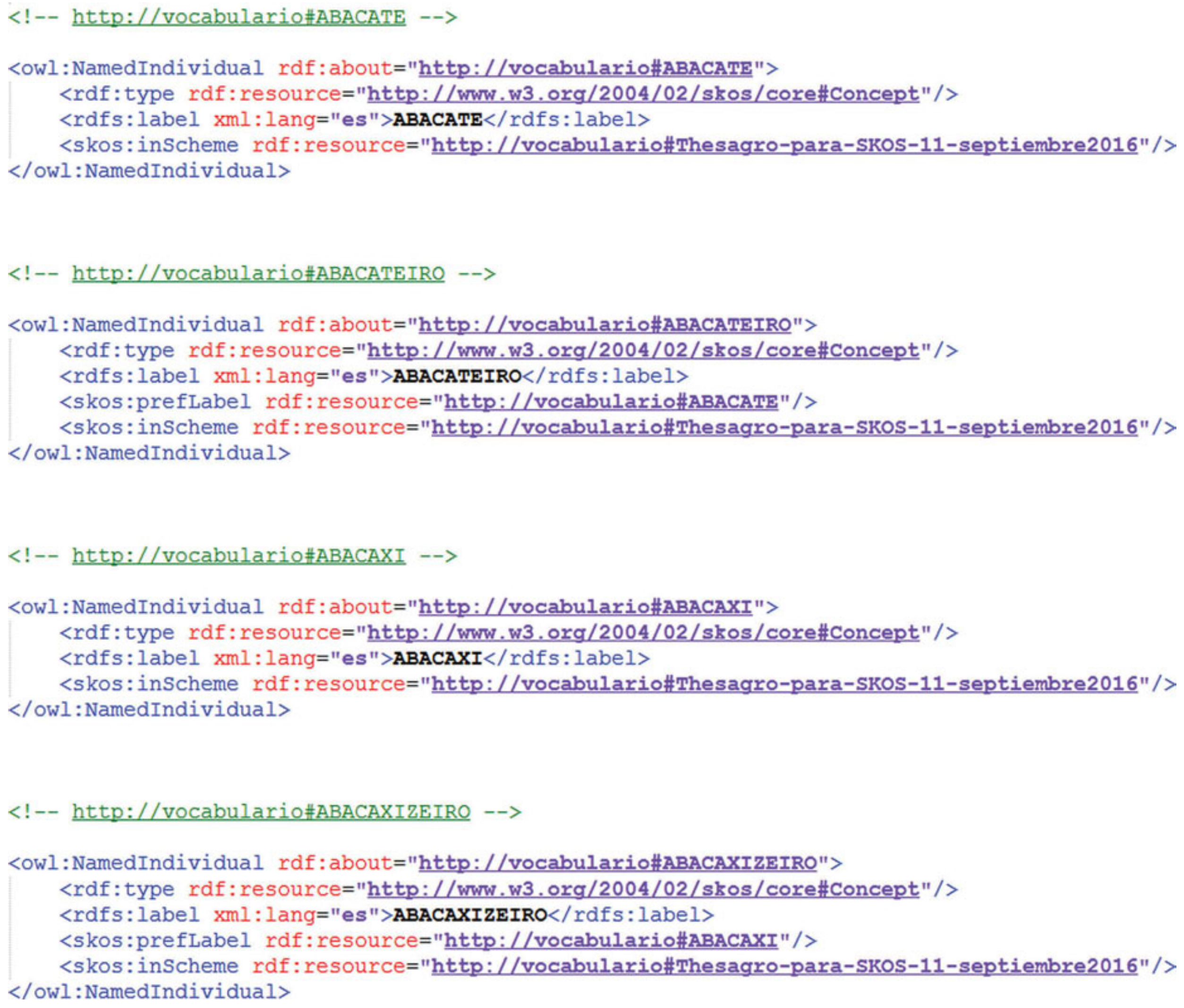
3.1 Test collection, controlled vocabulary and stopwords

We used a test collection comprising 200 scientific agricultural articles published in the *Revista Brasileña de Fruticultura* between 2006 and 2009. We also worked with controlled vocabulary in Portuguese with 9,588 descriptors and 1,122 non-descriptors in SKOS format (Figure 12). This vocabulary only uses synonymy (USE) and comes from Thesagro, a thesaurus prepared by the National Agriculture Library (BINAGRI) of the Brazilian Ministry of Agriculture. Finally, SISA has also used a list of stopwords in Portuguese made up of 586 words.

The sources used in these experiments (test collection, controlled vocabulary and stopwords list) can be found at webs.um.es/isgil.

3.2 Information needs and relevant documents

Fifteen information needs were prepared and were converted into search query equations. The relevant documents for each information need were then identified. Appendix 1 provides an example of all the data relative to information needs 1 and 2.



```
<!-- http://vocabulario#ABACATE -->

<owl:NamedIndividual rdf:about="http://vocabulario#ABACATE">
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <rdfs:label xml:lang="es">ABACATE</rdfs:label>
    <skos:inScheme rdf:resource="http://vocabulario#Thesagro-para-SKOS-11-septiembre2016"/>
</owl:NamedIndividual>



<!-- http://vocabulario#ABACATEIRO -->

<owl:NamedIndividual rdf:about="http://vocabulario#ABACATEIRO">
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <rdfs:label xml:lang="es">ABACATEIRO</rdfs:label>
    <skos:prefLabel rdf:resource="http://vocabulario#ABACATE"/>
    <skos:inScheme rdf:resource="http://vocabulario#Thesagro-para-SKOS-11-septiembre2016"/>
</owl:NamedIndividual>



<!-- http://vocabulario#ABACAXI -->

<owl:NamedIndividual rdf:about="http://vocabulario#ABACAXI">
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <rdfs:label xml:lang="es">ABACAXI</rdfs:label>
    <skos:inScheme rdf:resource="http://vocabulario#Thesagro-para-SKOS-11-septiembre2016"/>
</owl:NamedIndividual>



<!-- http://vocabulario#ABACAXIZEIRO -->

<owl:NamedIndividual rdf:about="http://vocabulario#ABACAXIZEIRO">
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <rdfs:label xml:lang="es">ABACAXIZEIRO</rdfs:label>
    <skos:prefLabel rdf:resource="http://vocabulario#ABACAXI"/>
    <skos:inScheme rdf:resource="http://vocabulario#Thesagro-para-SKOS-11-septiembre2016"/>
</owl:NamedIndividual>
```

Figure 12. Controlled vocabulary terms used by SISA.

3.3 Evaluation measures

In Gil-Leiva (2008, 385-397) we described various ways of evaluating indexing results. We spoke of a qualitative intrinsic evaluation by means of which expert indexers versed in indexing policy and language and the characteristics of the users of the system randomly select a significant number of database records and re-index them to reach a consensus on aspects like thoroughness (i.e., all the characterizing concepts have been extracted from the whole document), specificity (the existence of an exact relation between the conceptual units chosen and the term or terms chosen to represent them) or correctness (no errors in inclusion, no wrongly assigned term; no omissions, the exclusion of a term that should be assigned). Elsewhere we also spoke of quantitative intrinsic evaluation, consisting of re-indexing a set of documents and trying to reproduce as far as possible, the conditions of the original indexing (indexers, indexing policies, indexing language, work conditions, potential users, etc.) in order to get indexes that are consistent with mathematical formulas. One initial formula proposed by Hooper (1965) and a variant of the same incorporated by Rolling (1981) have been used extensively. The closer to 1 the resulting index, the better its consistency (i.e., the higher the coincidence between the two indexations). This quantitative intrinsic evaluation through consistency can be useful for evaluations within the same information unit using periodic intraconsistency tests, i.e., when a professional indexes a document again after some time in order to see if there are any variations with the first indexing.

We also spoke of extrinsic evaluation, using interconsistency, which is the application of one of the formulas mentioned in the previous paragraph to compare the indexing of the same document by two indexers from different institutions. These comparisons are complex because there should be some prevailing homogeneity of the factors involved in an indexing outcome, such as the indexer, the object analyzed and the context in which it is performed. Hence, a comparison between indexers of different documents begins with a study of each of these factors at both institutions and only when there is homogeneity can the formulas be applied to check for consistency. To all the above, one must add that we may find an indexing that is consistently incorrect, either on account of errors of inclusion (both professionals incorrectly assign the same term) or omission (both indexers neglect to assign a certain term).

We described in Gil-Leiva (2001, 69) various works that seek to explain the reasons and factors that cause the similarities or differences in the document analysis and, therefore, the consistency or inconsistency in the indexing. Some of these works were Zunde and Dexter (1969), Tarr and Borko (1974), Markey (1984), McCarthy (1986) or Chan (1989). Other studies analyzed the results being obtained from the many experiments which had been going on since the 1960s in order to ascertain degrees of consistency in indexing, e.g., Leonard (1977) or Markey (1984). As we stated at the time, and still believe, the indexing process is loaded with subjectivity (varying points of view of the same aspect or concept, different ways of converting a keyword in descriptor by means of a controlled vocabulary, the different ways to convert a user's information need into a search query), so "inconsistency is an inherent feature of indexing and not a sporadic anomaly" (Gil-Leiva 2008, 76).

Along with these experiments to evaluate indexing with consistency formulas another evaluation methodology, presented by Lancaster (1968, 1978, 1991) was gaining support using information, and it is still used today. Like the formulas to find consistency, this methodology also allows varied comparisons of indexations (within the same indexer, between two indexers, between an automatic indexing and an intellectual one, or between two automatic indexings). Basically, two databases are queried that contain the same records and identical content, except for the descriptor field, which houses the indexing under evaluation and with beforehand knowledge of the relevant documents for each of the queries to be executed. From the results, one obtains recall, precision and f-measure indices in the retrieval. This evaluation method is rather more laborious in indexing than the consistency method.

This evaluation by retrieval has used different formulae over the decades, but the most common are recall, precision and f-measure:

Recall = Number of relevant items retrieved / Number of relevant items in the collection
Precision = Number of relevant items retrieved / Number of items retrieved.
f-measure = Harmonic mean that combines precision and recall

$$F - measure = 2 \times \frac{recall \times precision}{recall + precision}$$

As is indicated in the following section, where the results of our experiments are analyzed, it is, in general, practically impossible/unviable to compare the results the hundreds of proposals of automatic indexing systems or prototypes mainly because of the heterogeneity of the measures for evaluating. Following the publication of the paper by Golub et al. (2016), let us hope that this situation can change. The paper cited is a valuable piece of research that analyzes and discusses the various methodologies, measures for evaluating and test collections, as well as other aspects, used in recent decades in the evaluation of automatic indexing systems or prototypes. However, it mainly offers an opportune and necessary framework for evaluating automatic indexing based on three complementary and combinable approaches: "1. Evaluating indexing quality directly through assessment by an evaluator or by comparison with a gold standard. 2. Evaluating indexing quality in the context of an indexing workflow. 3. Evaluating indexing quality indirectly through retrieval performance." (Golub et al. 2016, 6).

3.4 Experimental procedure

The main tasks in performing these experiments were:

1. Labeling the 200 Agriculture articles.
2. Uploading the 200 articles to SISA.
3. Establishing the rules. In total 50 location heuristics rules were established, all of them with the intervention of controlled vocabulary and three statistical rules to apply the TF-IDF (0.01, 0.015 and 0.02), i.e., only terms exceeding these thresholds are proposed as indexing terms. Controlled vocabulary was not used in the statistical rules.
4. Preparating 15 information needs.
5. Deciding on the relevant documents for the test collection of each of the information needs.
6. Converting the 15 information needs into search query equations.
7. Indexing the 200 articles by location heuristics rules.
8. Indexing the 200 articles by TF-IDF0.01.
9. Indexing the 200 articles by TF-IDF 0.015.
10. Indexing the 200 articles by TF-IDF0.02.
11. Applying search query to the database with the location heuristics rules indexing.
12. Applying search query to the database with the TF-IDF0.01 indexing.
13. Applying search query to the database with the TF-IDF 0.015 indexing.
14. Applying search query to the database with the TF-IDF0.02 indexing.
15. Calculating the evaluation measures (recall and precision) with the system's answers.

An example in Appendix 1 shows how data were gathered for searches 1 and 2 (information needs, documents in the database relevant to those needs, search query equations and the documents retrieved. Search query equations were used to query the database using the all fields and descriptors field. With all fields, information is sought from the whole register—title, abstract, keywords proposed by the authors of the articles and the SISA automatic indexing terms (with no manual edition). The same search functions are executed in the descriptor field, but here the information only takes in the terms obtained automatically by SISA with each of the rules mentioned in the previous paragraph.

5.0 Results and analysis

As stated in the methodology section, the evaluation measures used were recall, precision and f-measure. For the calculation of the precision measure, it has been established that if the denominator, i.e., the number of documents retrieved is 0, then the precision is 0.

Table 4 shows the duration of the processes. The location heuristics and statistical rules consumed the same time to indexing the 200 articles. As the table shows, labeling the articles consumes almost the whole process time.

Documents number	200
Megabytes	3,78
Labeled	400
Upload vocabulary	0.7
Upload stopwords	0.1
Procces	5:45
Indexing	4:30
Export to database	1:20
Total minutes	411:43
Average per document	2:05

Table 4. Process times in minutes.

In order to have more data available to draw conclusions regarding the functioning of the statistical rules in SISA, the test collection indexing was performed with three thresholds: TF-IDF, 0.01, 0.015 and 0.02. Hence, the system only proposes terms for a document if it surpasses one or more of these thresholds. Table 5 shows how the number of terms proposed by the statistical rules for the test collection is substantially higher than the terms proffered with location heuristics rules.

Terms selected by SISA for 200 articles		
	Terms for collection	Average number per document
Exp. 1 Heurísticas	1651	8.2
Exp. 2 tf-idf 0.01	8167	40.8
Exp. 3 tf-idf 0.015	3965	19.8
Exp. 4 tf-idf 0.020	2310	11.5

Table 5. Indexing terms number proposed by SISA.

Appendix 2 offers, as an example, all the indexing terms selected by SISA for an article and for each of the experiments.

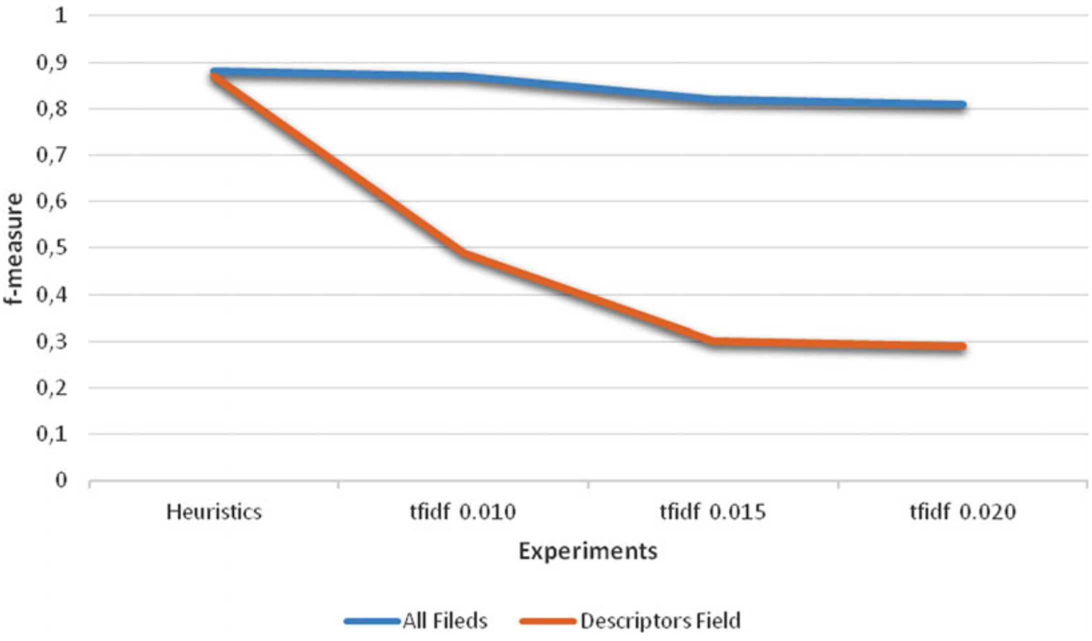


Figure 13. f-measure average.

Average recall		
	All fields	Descriptors field
Experiment 1: Heurísticas	0.85	0.62
Experiment 2: tf-idf 0.01	0.84	0.44
Experiment 3: tf-idf 0.015	0.78	0.26
Experiment 4: tf-idf 0.020	0.76	0.24
Average precision		
	All fields	Descriptors field
Experiment 1: Heurísticas	0.96	0.80
Experiment 2: tf-idf 0.01	0.91	0.66
Experiment 3: tf-idf 0.015	0.91	0.39
Experiment 4: tf-idf 0.020	0.90	0.38
Average f-measure		
	All fields	Descriptors field
Experiment 1: Heurísticas	0.88	0.87
Experiment 2: tf-idf 0.01	0.87	0.49
Experiment 3: tf-idf 0.015	0.82	0.30
Experiment 4: tf-idf 0.020	0.81	0.29

Table 6. Average recall, precision and f-measure for each experiment.

It is known that the assignation of a high number indexing terms to a document, as in experiment 2 (TF-IDF0.01), with a mean of 40.8 terms per document, is an obstacle to information retrieval since it can lead to the retrieval of non-relevant documents.

Table 6 shows that the location heuristics rules for recall and precision have achieved noticeably better results than the experiments with the TF-IDF. Appendix 3 gathers all the data from the experiments. The relative uniformity of the results obtained with the queries to all fields is because as well as using the indexing produced by SISA and stored in the descriptors fields, other fields of the database like title, abstract and authors' keywords were also made use of (see the column "All fields" in Table 6). Hence, the f-measure for the four experiments run on all fields of the database is relatively similar, ranging from 0.88 to 0.81, although improving slightly on the location heuristics rules.

In order to ascertain the role of the information in all fields during retrieval compared to the information in the descriptors field, the same search equations were run in the descriptors field (see column "Descriptors field" in Table 6). There are appreciable differences in the data returned by the location heuristics rules with respect to the different TF-IDF thresholds (Descriptors field column). In terms of recall, experiment 1 with location heuristics achieved the best results with 0.18, which was above the highest result obtained by the TF-IDF in experiment 2 (TF-IDF 0.01). As for the precision, experiment 1 scored 0.14 higher than the best data returned by statistical rules (experiment 2). Finally, in the case of the f-measure of all fields, level results were obtained which ranged from 0.88 in experiment 1 (location heuristics rules) to 0.81 with TF-IDF0.02; while the f-measure of the descriptors field gave notable results in experiment 1 (location heuristics) with 0.38 more than the best data recorded for a statistical heuristic, TF-IDF0.01, and 0.58 better than in experiment 4 (TF-IDF0.02).

Many automatic indexing projects or systems developed in recent decades and grounded on various approaches and methodologies have been mentioned. Most use test collections, controlled vocabularies and different stopword lists. Moreover, the evaluation measures used have likewise been varied. So, it is barely viable to make comparisons between data obtained by SISA and those by other prototypes. In Souza and Gil-Leiva (2016), we made a comparative study between SISA and PyPLN (Distributed Platform for Natural Language Processing) (Table 7). The PyPLN platform, coordinated by Souza, is an ongoing research project in the School of Applied Mathematics, Fundação Getulio Vargas, Brazil (Codeço, Souza, Juste, Amieiro and Mello 2013). In the SISA versus PyPLN experiment, we used the same test collection (100 articles from *Revista Brasileira de Fruticultura*); each system performed the indexing of the documents, and we then stored the resulting indexing in SISA. Search queries were later run on the retrieval module and the results obtained were used to get the evaluation measures. In this experiment SISA obtained better results than PyPLN.

6.0 Conclusions, limitations and future works

In this paper we contextualize indexing and offer an overview of some of the important methodologies and approaches used in automatic indexing asystems, and we also give a description of SISA. The aim of this research is to ascertain what rules (location heuristics or statistical rules) provide the best terms for indexing a test collection. The main contributions of this study can be summarized by saying that indexing of scientific articles with location heuristic rules performs better than with statistical rules when retrieving covering information. Despite the excellent results returned by the location heuristics, there is an important drawback in applying them since documents have to be labelled, which is a time-consuming process that requires around two minutes per article—practically the whole duration of the process.

There are a number of aspects worth extending in the future. First, new experiments with SISA are needed in order to explore all the possible permutations of the rules (location heuristics, statistical with and without controlled vocabulary) and the results of these must be evaluated. Second, there should be further research into procedures for automatic labelling of the various important parts of scientific articles so that the automatic indexing of SISA will take just a few seconds as indicated in Table 4. Third, to combine the processing, indexing and export tasks in a single click when one requires SISA to carry out automatic indexing (without edition). Four, to bestow the retrieval module with SISA options to automatically calculate the recall, precision and f-measure and so speed up the evaluation tasks. Finally, future work that may be of general interest and, therefore, more ambitious, would be to contribute to centralizing the access to tools and sources so that they can be used in the evaluation of automatic indexing systems as well as working on getting evaluation structures of automatic indexing up and running based on the proposals of Golub et al. (2016). We are aware that there are many researchers working on automatic indexing or in very similar, even overlapping, fields like automatic keyphrase extraction or automatic keyword for automatic classification, data mining, text summarization and information retrieval and that their scientific output is high. However, we consider that the findings are not getting transferred to the information systems and documental units in order to complete, enhance and facilitate the work of the professionals.

References

Anderson, James D. and José Perez-Carballo. 2001. "The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part I: Research and the Nature of Human Indexing." *Information Processing & Management* 37: 231-54.

Andreewsky, Alexandre and Vitoriano Ruas. 1982. *Indexaçâo automática baseada em métodos lingüísticos e estatísticos e su aplicabilidade à lingua portuguesa*. Rio de Janeiro: PUC-DI.

Aronson, Alan R., Olivier Bodenreider, H. Florence Chang, Susanne M. Humphrey, James G. Mork, Stuart J. Nelson, Thomas C. Rindflesch and W. John Wilbur. 2000. "The NLM Indexing Initiative." In *Proceedings of AMIA 2000 Annual Symposium: Converging Information Technology and Health Care: The Annual Symposium of the American Medical Informatics Association, November 4-8, 2000, Los Angeles, CA.,* edited J. Marc Overhage. Philadelphia: Hanley & Belfus, 17-21.

Ansari, Mariam. 2001. "Descriptors and Title Keywords: Matching in Medical PhD Dissertations." *Quarterly Journal of the National Library of the Islamic Republic of Iran* 12, no. 2: 23-33.

Barnes, C. I., L. Costantini and S. Perschke. 1978. "Automatic Indexing Using the SLC II System." *Information Processing & Management* 14: 107-19.

Beheshti, Moluksadat. 2003. "Terminology and Word Selection in Automated Indexing an Information Retrieval." *Journal of Information Sciences* 18, nos. 3/4: 31-44.

Biebricher, Peter, Norbert Fuhr, Gerhard Lustig, Michael Schwantner and Gerhard Knorz. 1988. "The Automatic Indexing System Air/Phys - From Research to Application." In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval*, edited by Y. Chiaramella. New York, NY: ACM, 333-42.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993-1022.

Bookstein, Abraham, Don R. Swanson. 1974. “Probabilistic Models for Automatic Indexing.” *Journal of the American Society for Information Science* 25: 312-18.

Bordoni, L. and Maria T. Pazienza. 1997. “Documents Automatic Indexing in an Environmental Domain.” *Internacional Forum on Information and Documentation* 22: 17-28.

Chan, Lois M. 1989. “Inter-indexer Consistency in Subject Cataloging.” *Information Technology and Libraries* 8: 349-57.

Chebil, Wiem, Lina F. Soualmia, Badisse Dahamna and Stéfan Darmoni. 2012. “Indexation automatique de documents en santé : évaluation et analyse de sources d’erreurs.” *IRBM* 33: 316-29.

Codeço, Flávio, Renato Rocha Souza, Álvaro Justen, Flávio Amieiro and Heliana Mello. 2013. “PyPLN: A Distributed Platform for Natural Language Processing.” https://arxiv.org/pdf/1301.7738v2.pdf

Craven, Timothy C. 2005. “Web Authoring Tools and Meta Tagging of Page Descriptions and Keywords.” *Online Information Review* 29: 129-38.

Damerau, Fred J. 1965. “An Experiment in Automatic Indexing.” *American Documentation* 16: 283-9.

Dehghani, Mostafa, Hosein Azarbonyad, Maarten Marx and Jaap Kamps. 2015 “Sources of Evidence for Automatic Indexing of Political Texts.” *Lecture Notes in Computer Science*, 9022: 568-73.

Deerwester, Scoot, Susan T. Dumais, Thomas K. Landauer, George W. Furnas and Louis Beck. 1988. “Improving Information Retrieval with Latent Semantic Indexing.” In *Proceedings of the 51st Annual Meeting of the American Society for Information Science* 25: 36-40.

Dextre, Clarke Stella G. 2016. “Origins and Trajectory of the Long Thesaurus Debate.” *Knowledge Organization* 43: 138-44.

Dillon, Martin and Ann S. Gray. 1983. “FASIT: A Fully Automatic Syntactical Y based Indexing System.” *Journal of the American Society for Information Science* 34: 99-108.

Driscoll, James R., David A. Rajala, William H. Shaffer and Donald W. Thomas. 1991. “The Operation and Performance of an Artificially Intelligent Keywording System.” *Information Processing & Management* 227: 43-54.

El-Haj, Mahmoud, Lorna Balkan, Suzanne Barbalet, Lucy Bell and John Shepherdson. 2013. “An Experiment in Automatic Indexing Using the HASSET Thesaurus.” In *5th Computer Science and Electronic Engineering Conference, CEEC 2013, Colchester; United Kingdom, 17-18 September 2013*. Piscataway, NJ: IEEE, 13-8.

Evans, David A. 1990. “Concept Management in Text via Natural-Language Processing: The CLARIT Approach.” In *Working Notes of the 1990 AAAI Symposium on “Text-Based Intelligent Systems” 9, Stanford University, March, 27-29*, 1990. [Menlo Park, Calif.]: [AAAI], 93-5.

Evans, David A., William R. Hersh, Ira A. Monarch, Robert G. Lefferts and Steven K. Handerson. 1991a. “Automatic Indexing of Abstracts via Natural-Language Processing Using a Simple Thesaurus.” *Medical Decision Making* 11, no. 4: 108-15.

Evans, David A, Steve K. Handerson, Robert G. Lefferts and Ira A. Monarch. 1991b. “A Summary of the CLARIT project. November 1991, Report No. CMU-LCL-91-2.” http://repository.cmu.edu/philosophy/index.6.html

Faraj, Najib, Robert Godin, Rokia Missaouit, Sophie David and Pierre Plante. 1996. “Analyse d’une méthode d’indexation automatique basée sur une analyse syntaxique de texte.” *Canadian Journal of Information and Library Science* 21: 1-21.

Farrow, John F. 1991. “A Cognitive Process Model of Document Indexing.” *Journal of Documentation* 47: 149-66

Frohmann, Bernd. 1990. “Rules of Indexing: A Critique of Mentalism in Information Retrieval Theory.” *Journal of Documentation* 46: 81-101.

Fugmann, Robert. 1993. *Subject Analysis and Indexing: Theoretical Foundation and Practical Advice*. Frankfurt/Main: Indeks Verlag.

Gbur, Edward E. and Bruce E. Trumbo. 1995. “Key Words and Phrases-The Key to Scholarly Visibility and Efficiency in an Information Explosion.” *The American Statistician* 49: 29-33.

Gil-Leiva, Isidoro y Rodríguez Muñoz, José Vicente. 1996. “Tendencias en los sistemas de indización automática. Estudio evolutivo.” *Revista Española de Documentación Científica* 19: 273-92.

Gil-Leiva, Isidoro and José Vicente Rodríguez Muñoz. 1997. “Análisis de los descriptores de diferentes áreas de conocimiento indizadas en bases de datos del CSIC. Aplicación a la indización automática.” *Revista Española de Documentación Científica* 20: 150-60.

Gil-Leiva, Isidoro. 1999. *La automatización de la indización*. Gijón: Trea.

Gil-Leiva, Isidoro. 2001. “Consistencia en la asignación de materias en Bibliotecas Públicas del Estado.” *Boletín de la Asociación Andaluza de Bibliotecarios* 63: 69-86.

Gil-leiva, Isidoro and Alonso Arroyo, Adolfo. 2005. “La relación entre las palabras clave aportadas por autores de artículos de revista y su indización en las Bases de datos ISOC, IME e ICYT.” *Revista Española de Documentación Científica* 28: 62-79.

Gil-leiva, Isidoro and Alonso Arroyo, Adolfo. 2007. “The Presence of the Keywords Given by Authors of Scientific Articles in Databases Descriptors.” *Journal of the American Society for Information Science and Technology* 58: 1175-87.

Gil-Leiva, Isidoro. 2008. *Manual de indización. Teoría y práctica*. Gijón: Trea.

Golub, Koraljka, Dagobert Soergel, George Buchanan, Douglas Tudhope, Marianne Likke and Debra Hiom. 2016. "A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval." *Journal of the Association for Information Science and Technology* 67: 3-16.

Haller, Johan. 1983. "Análise automática de textos em sistemas de informaçâo." *Revista de Biblioteconomia de Brasília* 11: 105-13.

Hartley, J. and R. N. Kostoff. 2003. "How Useful Are 'Key Words' in Scientific Journals?." *Journal of Information Science* 29: 433-8.

Hersh William R. and Robert A. Greenes. 1990. "SAPHIRE: An Information Retrieval Environment Featuring Conceptmatching, Automatic Indexing, and Probabilistic Retrieval." *Computers and Biomedical Research* 123: 410-25.

Hersh William R., David H. Hickam, R. Brian Haynes, K. Ann McKibbon. 1991. "Evaluation of SAPHIRE: An Automated Approach to Indexing and Retrieving Medical Literature." In *Proceedings: The Annual Symposium on Computer Applications in Medical Care*. New York: IEEE, 808-12.

Hersh, William R. and Robert A. Greenes. 1990. "SAPHIRE, An Information Retrieval System Featuring Concept Matching Automatic Indexing, Probabilistic Retrieval, and Hierarchical Relationships." *Computers and Biomedical Research* 23: 410-25.

Hjorland, Biger. 1997. *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Westport, CT: Greenwood Press.

Hmeidi, Ismail, Ghassan Kanaan and Martha Evens. 1997. "Design and Implementation Of Automatic Indexing For Information Retrieval With Arabia Documents." *Journal of the American Society for Information Science* 48: 867-81.

Hodge, Gail M. 1992. *Automated Support to Indexing*. Philadelphia: National Federation of Abstracting and Information Services, NFAIS Report Series 3.

Hofmann, Thomas. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, Berkeley, CA, USA, August 15-19, 1999,* ed Fredric Gey, Marti Hearst and Richard Tong. New York, NY: ACM, 50-7.

Hooper, Robert S. 1965. *Indexer Consistency Tests: Origin, Measurement, Results, and Utilization*. Bethesda: IBM Corporation.

Humphrey, Susanne M. 1999. "Automatic Indexing of Documents from Journal Descriptors: A Preliminary Investigation." *Journal of the American Society for Information Science* 50: 661-74.

Humphrey, Susanne M., Willie, J. Rogers, Halil Kilicoglu, Dina Demner-Fushman and Thomas C. Rindflesch. 2006. "Word Sense Disambiguation by Selecting the Best Semantic Type Based on Journal Descriptor Indexing: Preliminary Experiment." *Journal of the American Society for Information Science and Technology* 57: 96-113.

Humphrey, Susanne M. and Nancy E. Miller. 1987. "Knowledge-Based Indexing of the Medical Literature: The Indexing Aid Project." *Journal of the American Society for Information Science* 38: 84-196.

Irving, Holly Berry. 1997. "Computer-Assisted Indexing Training and Electronic Text Conversion at NAL" *Knowledge Organization* 24: 4-7.

ISO. 1985. *ISO 5963-1985: Documentation -- Methods for Examining Documents, Determining Their Subjects, and Selecting Indexing Terms*. Geneva: ISO.

Joorabchi, Arash and Abdulhussain E. Mahdi. 2014 "Towards Linking Libraries and Wikipedia: Automatic Subject Indexing of Library Records with Wikipedia Concepts." *Journal of Information Science* 40: 211-21.

Joorabchi, Arash and Abdulhussain E. Mahdi. 2013. "Automatic Keyphrase Annotation of Scientific Documents Using Wikipedia and Genetic Algorithms." *Journal of Information Science* 39: 410-26.

Karetnyk, David, Fred Karlsson and Godfrey Smart. 1991. "Knowledge-based Indexing of Morpho-Syntactically Analysed Language*." Expert Systems for Information Management* 4: 1-29.

Kipp, Margaret E. I. 2011. "Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing." *Knowledge Organization* 38: 245-61.

Klingbiel, Paul H. 1973. "Machine-aided Indexing of Technical Literature." *Information Storage and Retrieval* 9: 79-84.

Klingbiel, Paul H. and Catherine C. Rinker. 1976. "Evaluation of Machine-Aided Indexing." *Information Processing & Management* 12: 351-66.

Kolar, Mladen, Igor Vukmirović, Bojana Dalbelo Bašić and Jan Šnajder. 2005. "Computer-aided Document Indexing System." *Journal of Computing & Information Technology* 13: 299-305.

Lahtinen, Timo. 2000. "Automatic Indexing: An Approach Using an Index Term Corpus and Combining Linguistic and Statistical Methods." PhD thesis, Department of General Linguistics, University of Helsinki. https://helda.helsinki.fi/bitstream/handle/10138/19292/automati.pdf?sequence=2

Lancaster, Frederick W. 1968. *Evaluation of the MEDLARS Demand Search Service*. Bethesda: National Library of Medicine.

Lancaster, Frederick W. 1978. "Precision and Recall." In *Encyclopedia of Library and Information Science: Volume 23, Poland to Printers and Printing*, edited by Allen Kent, Harold Lancour and Jay E. Daily. New York: Dekker, 170-80.

Lancaster, Frederick W. 1991. *Indexing and Abstracting in Theory and Practice*. Champaign: University of Illinois.

Leonard, Lawrence E. (ed.). 1977. *Inter-indexer Consistency Studies, 1954-1975: A Review of the Literature and Summary of the Study Results*. University of Illinois.

Liebesny, Felix. 1974. *A State-of-art Survey on Automatic Indexing*. Paris. Unesco.

Lu, Kun and Margaret E. I. Kipp. 2014. "Understanding the Retrieval Effectiveness of Collaborative Tags and Author Keywords in Different Retrieval Environments: An Experimental Study on Medical Collections." *Journal of the Association for Information Science and Technology* 65: 483-500.

Luhn, Hans Peter. 1957a. "A Statistical Approach to Mechanized Enconding and Searching of Literary Information." *IBM Journal of Research and Development* 1: 309-17.

Luhn, Hans Peter. 1957b. "The Automatic Creation of Literature Abstracts." *IBM Journal of Research and Development* 2: 159-65.

Lukashevich, Natalia V. and Boris V. Dobrov. 2001. "Modifieres of Conceptual Relationships in a Thesaurus for Automated Indexing." *Nauchno- Tekhnicheskaya Informatsiya*. Series 2: 21-8.

Maron, Melvin Earl and John Lary Kuhns. 1960. "On Relevance, Probabilistic Indexing and Information Retrieval." *Journal of Association for Computing Machinery* 25: 216-44.

Maeda, Tkashi, Yoshio Momouchi and Hajime Sawamura. 1980. "An Automatic Method for Extracting Significant Phrases in Scientific or Technical Documents." *Information Processing and Management* 16: 119-27.

Mai, Jens-Erik. 2000. "Deconstructing the Indexing Process." *Advances in Librarianship*, 23: 269-98.

Markey, Karen. 1984. "Interindexer Consistency Tests: A Literature Review and Report of a Test of Consistency in Indexing Visual Materials." *Library & Information Science Research* 6: 155-77.

Martínez, Clara, John Lucey and Elliot Linder. 1987. "An Expert System for Machine-Aided-Indexing*." Journal of Chemical Information and Computer Sciences* 27: 158-62.

Mccarthy, C. 1986. "The Reliability Factor in Subject Access." *College and Research Libraries* 47: 48-56.

Medelyan, Olena and Ian H. Witte. 2008. "Domain-Independent Automatic Keyphrase Indexing with Small Training Sets." *Journal of the American Society for Information Science & Technology* 59: 1026-40.

Medelyan, Olena and Ian H. Witten. 2005. "Thesaurus-based Index Term Extraction for Agricultural Documents." In *Proceedings of the 6th Agricultural Ontology Service (AOS) workshop at EFITA/WCCA* 2005, Vila Real, Portugal.

Meulen, W. A. and P. J. Janssen. 1977. "Automatic Versus Manual Indexing." *Information Processing & Management* 13: 13-21.

Moens, Marie Francine. 2000. *Automatic Indexing and Abstracting of Document Texts*. Boston: Kluwer.

Montejo Ráez, Arturo. 2001. "Proyecto de indexado automático para documentos en el campo de la física de altas energías." *Boletín de Sociedad Española para el Procesamiento del Lenguaje Natural* 27: 295-96.

Mork James G., Dina Demner-Fushman, Susan C. Schmidt and Alan R. Aronson. 2014. "Recent Enhancements to the NLM Medical Text Indexer." In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18*, edited by Linda Cappellato, Nicola Ferro, Martin Halvey and Wessel Kraaij. [S. l.]: CEUR-WS, 1328–36. http://ceur-ws.org/Vol-1180/CLEF2014wn-QA-MorkEt2014.pdf

Pickler, Maria Elisa Valentim and Edberto Ferneda. 2014. "Um Método para a Utilização de Ontologias na Indexação Automática." *Informação & Tecnologia* 1, no. 2: 13-33.

Plaunt, Christian and Barbara A. Nogard. 1998. "An Association-Based Method for Automatic Indexing with a Controlled Vocabulary." *Journal of the American Society for Information Science* 49: 888-902.

Prada, Laura, Javier García, J. García and Jesús Carretero. 2011. "Agricultural Library Speeds Indexing." *KM World* 20, no. 11 6.

Pulgarín, Antonio and Gil-Leiva, Isidoro. 2004. "Bibliometric analysis of the automatic indexing literature: 1950-2000." *Information Processing & Management* 40: 365-77.

Roberts, David and Clive Souter. 2000. "The Automation of Controlled Vocabulary Subject Indexing of Medical Journal Articles." *Aslib Proceedings* 52: 384-401.

Rolling, Loll N. 1981. "Indexing Consistency, Quality and Efficiency." *Information Processing & Management* 17: 69-76.

Rosenberg, Victor. 1971. "A Study of Statistical Measures for Predicting Terms Used to Index Documents." *Journal of the American Society for Information Science* 22: 41-50.

Salisbury, Lutishoor and Jemery J. Smith. 2014. "Building the AgNIC Resource Database Using Semi-Automatic Indexing of Material." *Journal of Agricultural & Food Information* 15, no. 3: 159-79.

Salton, Gerard. 1972. "A New Comparison Between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART)." *Journal of the American Society for Information Science* 23: 75-84.

Salton, Gerad and Chung-Shu Yang. 1973. "On the Specification of Term Values In Automatic Indexing." *Journal of Documentation* 29: 351-72.

Salton, Gerard, Chung-Shu Yang and Clement T. Yu. 1974. "Contribution to the Theory of Indexing." *Information Processing* 74: 584-90.

Salton, Gerard, Chung-Shu Yang and Clement T. Yu. 1975. "A Theory of Term Importance in Automatic Text Analysis." *Journal of the American Society for Information Science* 26: 33-44.

Salton, Gerard, Andrew Wong and Chung-Shu Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the Association for Computing Machinery* 18: 613-20.

Salton, Gerard, Harry Wu and Clement T. Yu. 1981. "The Measurement of Term Importance in Automatic Indexing." *Journal of the American Society for Information Science* 32: 175-86.

Salton, Gerard. 1980. "The SMART System 1961-1976: Experiments in Dynamic Document Processing." *Encyclopedia of Library and Information Science*, ed. Allen Kent, Harold Lancour and Jay E. Daily. New York: Dekker, 28: 1-28.

Salton, Gerard. 1991. "The Smart Document Retrieval Project." In *Proceeding SIGIR '91 Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *Chicago, Il, October 13-16, 1991,* ed. Edward Fox. New York: ACM, 356-58.

Scheele, Martin. 1983. "Automatic Indexing of Titles and Keywords on the Bases of a Model for an Overall Thesaurus of Knowledge." *International Classification* 10: 135-7.

Schuegraf, Ernst J. and Martin F. Bommel. 1993. "An Automatic Document Indexing System based on Cooperating Expert Systems: Design and Development." *Canadian Journal of Information and Library Science* 18: 32-50.

Seo, Eun-Gyoun. 1993. *An Experiment in Automatic Indexing with Korean Texts: A Comparison of Syntactic-Statistical and Manual Methods.* Urbana-Champaign: University of Illinois.

Silvester, June P., Michael T. Genuardi and Paul H. Klingbiel. 1994. "Machine-aided Indexing at NASA." *Information Processing & Management* 30: 631-45.

Smiraglia, Richard P. 2013 'Keywords, Indexing, Text Analysis: An Editorial.' *Knowledge Organization* 40: 155-9.

Souza, Renato Rocha. 2005. "Uma proposta de metodología para escolha automática de descritores utilizando sintagmas nominais." 214 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência de Informação, Universidade Federal de Minas Gerais, Belo Horizonte.

Souza, Renato Rocha and Koti S. Raghavan. 2006. "A Methodology for Noun Phrase-Based Automatic Indexing." *Knowledge Organization* 33: 45-56.

Souza, Renato Rocha and Koti S. Raghavan. 2014. "Extraction of Keywords from Texts: An Exploratory Study Using Noun Phrases." *Informação & Tecnologia* 1: 5-16.

Souza, Renato Rocha and Isidoro Gil-Leiva. 2016. "Automatic Indexing of Scientific Texts: A Methodological Comparison." In *Knowledge Organization for a Sustainable World: Challenges and Perspectives for Cultural, Scientific, and Technological Sharing in a Connected Society: Proceedings of the Fourteenth International ISKO Conference* 27-29 *September* 2016, *Rio de Janeiro, Brazil,* ed. José Augusto Chaves Guimarães, Suellen Oliveira Milani and Vera Dodebei. Advances in Knowledge Organization 15. Würzburg: Ergon Verlag, 243-50.

Sparck Jones, Karen. 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." *Journal of Documentation* 28: 11-21.

Sparck Jones, Karen. 1974. *A State of the Art Review.* Cambridge. University of Cambridge.

Steinberger, Raff, Johan Hagman and Stefan Scheer. 2000. "Using Thesauri for Automatic Indexing Multilingual Document Collections" In *OntoLex'200 –Workshop on Ontologies and Lexical Knowledge Bases, Sozopol, Bulgaria, 8-10 September 2000,* https://pdfs.semanticscholar.org/5536/9539136c2c80634ef4f0af8b0f71a986e913.pdf

Stevens, Mary Elizabeth. 1965. *Automatic Indexing: A State of the Art Report, Monograph 91.* Washington, D.C.: National Bureau of Standards.

Strader, C. Rockelle. 2009. "Author-Assigned Keywords versus Library of Congress Subject Headings Implications for the Cataloging of Electronic Theses and Dissertations." *Library Resources & Technical* Services 53: 243-50.

Strode, Margaret S. 1977. "Automatic Indexing Using a Thesaurus." Master's thesis, Department of Computer Science, University of North Carolina at Chapel.

Tanijiri, Junki, Manabu Ohta, Atsuhiro Ohta and Jun Adachi. 2016. "Important Word Organization for Support of Browsing Scholarly Papers Using Author Keywords." In *Proceedings of the 2016 ACM Symposium on Document Engineering, 13 September 2016*, *Vienna, Austria,* ed. Robert Sablatnig. New York: ACM, 135-8.

Tarr, Daniel and Harold Borko. 1974. "Factors Influencing Inter-Indexer Consistency." In *Information Utilities: Proceedings of the 37th ASIS Annual Meeting, Atlanta, Georgia, October 13-17, 1974,* ed. Pranas Zunde and Tamir Hassan. Washington D.C.: American Society for Information Science, 50-5.

Trubkin, Loene. 1979. "Auto-indexing of the 1971-77 AB1/INFORM Database." *Database* 2, no. 2: 56-61.

Valle Bracero, Antonio and Justo A. Fernández García. 1983. "Automatización de la indización y coordinación de descriptores." *Revista Española de Documentación Científica* 6: 9-16.

Vlachidis, Andreas and Douglas Tudhope. 2016. "A Knowledge-Based Approach to Information Extraction for Semantic Interoperability in the Archaeology Domain." *Journal of the Association for Information Science & Technology* 67: 1138-52.

Vrkic, Dina. 2014. "Are They a Perfect Match? Analysis of Usage of Author Suggested Keywords, IEEE Terms and Social Tags." In *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 26-30, 2014, Opatija, Croatia,* ed. Petar Biljanovic et al. Rijeka, Croatia: Croatian Society for Information and Communication Technology, Electronics and Microelectronics, 732-7.

Wan, Tian-Long, Martha Evans, Yuen-Wen Wan and Yuen-Yuan Pao. 1997. "Experiments with Automatic Indexing and a Relational Thesaurus in a Chinese Information Retrieval System." *Journal of the American Society for Information Science and Technology* 48: 1086-96.

Willis, Craig and Robert M. Losee. 2013. "*A Random Walk on an Ontology: Using Thesaurus Structure for Automatic Subject Indexing.*" *Journal of the American Society for Information Science & Technology* 64: 1330-44.

Zha, Guiting and Hanqing Hou. 2002. "Automatic Indexing Based on Multi-Vocabularies." *Journal of the China Society for Scientific and Technical Information* 21: 273-7.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort.* Cambridge, Mass.: Addison-Wesley Press.

Zunde, Pranas and Margaret E. Dexter. 1969. "Indexing Consistency and Quality." *American Documentation* 20: 259-67.

Appendix 1:
Search 1 and 2.

Search	Information needs	Relelevant documents	Retrieved documents by heuristics rules	Retrieved documents by statistical rules
1	Enraizamento de estacas	19, 22, 59, 61, 77, 95, 97, 134, 173, 195	All fields: 22, 59, 61, 77, 97, 134, 171, 195 ******************* Descriptors field: 22, 59, 61, 77, 134, 195	All fields: **0.01 =** 19, 22, 59, 61, 77, 97, 134, 136, 171, 173, 195 **0.015 =** 19, 59, 61, 77, 97, 134, 136, 171, 173, 195 **0.020 =** 19, 59, 61, 77, 97, 134, 136, 171, 195 ****************************** Descriptors field: **0.01 =** 19, 22, 59, 61, 77, 97, 134, 136, 173, 195 **0.015 =** 19, 61, 77, 97, 134, 136, 173, 195 **0.020 =** 19, 61, 77, 134, 136, 195
2	Tratamentos para a conservação de frutas	14, 41, 43, 82, 104, 105, 107, 108, 109, 110, 111, 141, 142, 143	All fields: 14, 41, 43, 104, 105, 108, 141, 200 ******************* Descriptors field: 0 documents	All fields: **0.01 =** 14, 41, 43, 104, 105, 108, 141, 200 **0.015 =** 14, 41, 43, 104, 105, 108, 141, 200 **0.020 =** 14, 41, 43, 104, 105, 108, 141, 200 ****************************** Descriptors field: **0.01 =** 0 documentos **0.015 =** 0 documentos **0.020 =** 0 documentos

Appendix 2:

Example of indexing terms proposed by SISA for an article

Article: **Relações filogenéticas e diversidade de isolados de Guignardia spp oriundos de diferentes hospedeiros nas regiões ITS1-5,8S-ITS2.** ***Revista Brasileira de Fruticultura,*** **2009, 31 (2): 360-380, ISSN 0100-2945.**

Heurísticas	**tf-idf 0.01**	**tf-idf 0.015**	**tf-idf 0.020**
analise	aveia-agar	banco de dados	citricarpa
especie	banco	citricarpa	endofiticos
fruta citrica	banco de dados	distancias geneticas	especies
fruto	citricarpa	dna	guignardia
fungo	depositarias	endofiticos	hospedeiro
goiaba	diferentes hospedeiros	especies	id
hospedeiro	distancias	fungo	identificar
isolamento	distancias geneticas	goiabeira	isolado
manga	dna	grupos de isolados	isolados obtidos
niveladora	endofiticos	guignardia	mangiferae
planta	especies	hospedeiro	
podridão	especies g	id	
	fungo	identificar	
	genetico	isolado	
	goiabeira	isolados obtidos	
	grupo	jabuticabeira	
	grupos de isolados	mangiferae	
	guignardia	mangueira	
	halo	sequencia	
	hospedeiro	sequencia de dna	
	id		
	identificados como g		
	identificar		
	isolado		
	isolados obtidos		
	jabuticabeira		
	laranja-'azeda'		
	laranja-'pera'		
	lima-acida		
	lima-acida 'tahiti'		
	mangiferae		
	mangueira		
	meio aveia-agar		
	phyllosticta		
	pitangueira		
	psidii		
	sequencia		
	sequencia de dna		
	similarmente		
	'tahiti'		

Appendix 3: Experiment data

Experiment 1:
Heuristics rules

Heuristics						
	Recall		**Precision**		**f-measure**	
Searchs	All Fields	Descriptors Field	All Fields	Descriptors Field	All Fields	Descriptors Field
Search 1	0.7	0.6	0.87	1	0.78	0.75
Search 2	0.5	0	0.87	0	0.64	0.00
Search 3	0.5	0	1	0	0.67	0.00
Search 4	1	1	1	1	1.00	1.00
Search 5	1	1	1	1	1.00	1.00
Search 6	1	1	1	1	1.00	1.00
Search 7	0.88	0.55	1	1	0.94	0.71
Search 8	1	0.66	1	1	1.00	0.80
Search 9	1	0.66	1	1	1.00	0.80
Search 10	1	1	1	1	1.00	1.00
Search 11	1	0.8	0.71	1	0.83	0.89
Search 12	0.22	0	1	0	0.36	0.00
Search 13	1	0.66	1	1	1.00	0.80
Search 14	1	1	1	1	1.00	1.00
Search 15	1	1	1	1	1.00	1.00
Average	**0.85**	**0.62**	**0.96**	**0.8**	**0.88**	**0.87**

Experiment 2:
TF-IDF 0.01

TF-IDF 0.01						
	Recall		**Precision**		**f-measure**	
Searchs	All Fields	Descriptors Field	All Fields	Descriptors Field	All Fields	Descriptors Field
Search 1	0.90	0.90	0.81	0.9	0.85	0.90
Search 2	0.5	0	0.87	0	0.64	0.00
Search 3	0	0	0	0	0.00	0.00
Search 4	1	1	1	1	1.00	1.00
Search 5	1	0	1	0	1.00	0.00
Search 6	1	1	1	1	1.00	1.00
Search 7	1	0.88	1	1	1.00	0.94
Search 8	1	0.66	1	1	1.00	0.80
Search 9	1	0.16	1	1	1.00	0.28
Search 10	1	1	1	1	1.00	1.00
Search 11	1	0.2	1	1	1.00	0.33
Search 12	0.33	0	1	0	0.50	0.00
Search 13	1	0.66	1	1	1.00	0.80
Search 14	1	0.2	1	1	1.00	0.33
Search 15	1	0	1	0	1.00	0.00
Average	**0.84**	**0.44**	**0.91**	**0.66**	**0.87**	**0.49**

Experiment 3:
TF-IDF 0.015

TF-IDF 0.015						
	Recall		**Precision**		**f-measure**	
Searchs	All Fields	Descriptors Field	All Fields	Descriptors Field	All Fields	Descriptors Field
Search 1	0.8	0.7	0.8	0.87	0.80	0.78
Search 2	0.5	0	0.87	0	0.64	0.00
Search 3	0	0	0	0	0.00	0.00
Search 4	1	0	1	0	1.00	0.00
Search 5	1	0	1	0	1.00	0.00
Search 6	1	1	1	1	1.00	1.00
Search 7	0.88	0.55	1	1	0.94	0.71
Search 8	1	0.33	1	1	1.00	0.50
Search 9	0.66	0	1	0	0.80	0.00
Search 10	1	1	1	1	1.00	1.00
Search 11	1	0	1	0	1.00	0.00
Search 12	0.22	0	1	0	0.36	0.00
Search 13	0.66	0.33	1	1	0.80	0.50
Search 14	1	0	1	0	1.00	0.00
Search 15	1	0	1	0	1.00	0.00
Average	**0.78**	**0.26**	**0.91**	**0.39**	**0.82**	**0.30**

Experiment 4:
TF-IDF 0.02

TF-IDF 0.02						
	Recall		**Precision**		**f-measure**	
Searchs	All Fields	Descriptors Field	All Fields	Descriptors Field	All Fields	Descriptors Field
Search 1	0.7	0.5	0.77	0.83	0.73	0.62
Search 2	0.5	0	0.87	0	0.64	0.00
Search 3	0	0	0	0	0.00	0.00
Search 4	1	0	1	0	1.00	0.00
Search 5	1	0	1	0	1.00	0.00
Search 6	1	1	1	1	1.00	1.00
Search 7	0.88	0.55	1	1	0.94	0.71
Search 8	1	0.33	1	1	1.00	0.50
Search 9	0.5	0	1	0	0.67	0.00
Search 10	1	1	1	1	1.00	1.00
Search 11	1	0	1	0	1.00	0.00
Search 12	0.22	0	1	0	0.36	0.00
Search 13	0.66	0.33	1	1	0.80	0.50
Search 14	1	0	1	0	1.00	0.00
Search 15	1	0	1	0	1.00	0.00
Average	**0.76**	**0.24**	**0.90**	**0.38**	**0.81**	**0.29**