TENDÊNCIAS ATUAIS E PERSPETIVAS FUTURAS EM ORGANIZAÇÃO DO CONHECIMENTO

ATAS DO III CONGRESSO ISKO ESPANHA-PORTUGAL
XIII CONGRESSO ISKO ESPANHA

Universidade de Coimbra, 23 e 24 de novembro de 2017

Com a coordenação de

Maria da Graça Simões, Maria Manuel Borges

TÍTULO
Tendências Atuais e Perspetivas Futuras em Organização do Conhecimento: atas do III Congresso ISKO Espanha e Portugal - XIII Congresso ISKO Espanha

COORDENADORES
Maria da Graça Simões
Maria Manuel Borges

EDIÇÃO
Universidade de Coimbra. Centro de Estudos Interdisciplinares do Século XX - CEIS20

ISBN
978-972-8627-75-1

ACESSO
https://purl.org/sci/atas/isko2017

OBRA PUBLICADA COM O APOIO DE

PROJETO UID/HIS/00460/2013

LA INDIZACIÓN DE ARTÍCULOS CIENTÍFICOS CON EL SISTEMA DE INDIZACIÓN AUTOMÁTICA SISA COMPARADA CON LA INDIZACIÓN EN LAS BASES DE DATOS AGRICOLA, WOS Y SCOPUS

Isidoro Gil-Leiva

Facultad de Comunicación y Documentación, Universidad de Murcia, isgil@um.es

RESUMEN Desde hace unos años la generación de documentos digitales es enorme así como su incorporación masiva a los sistemas de información y ambas realidades parecen imparables. Del mismo modo, no hay duda de que la indización es uno de los procesos fundamentales ejecutados en las unidades documentales. Aunque las primeras investigaciones en automatización de la indización se iniciaron hace décadas este asunto sigue suscitando interés. Desde entonces diferentes propuestas y metodologías han sido planteadas. SISA es un sistema de indización automática multilingue para artículos científicos fundamentado en principios heurísticos y estadísticos regido mediante reglas basadas en dichos principios. **Objetivo**. En este contexto descrito de incremento digital constante, se persigue conocer las capacidades de SISA en la indización automática de artículos en relación a cómo lo hacen en las bases de datos Agricola, WOS y SCOPUS. **Material y método**. Se seleccionaron al azar cien artículos publicados en diferentes años por la revista de Agricultura *Agronomy for sustainable development*, se localizó la indización asignada a los artículos en las mencionadas bases de datos, se indizaron los documentos con SISA, se compararon las diferentes indizaciones y se calcularon los índices de consistencia entre Agricola y SISA. **Conclusiones.** Las capacidades de indización de SISA en relación a las bases de datos de referencia han sido satisfactorias, si bien se precisan algunos ajustes. SISA ha producido un número medio de descriptores por documento similares a Agricola y Scopus, si bien, los descriptores compuestos de SISA es menor que en estas dos bases de datos. Asimismo, el 21,61% de consistencia conseguido entre SISA y Agricola se encuentra dentro de los porcentajes en este tipo de estudios. Por último, la propuesta de una fórmula integral para la evaluación de la indización automática denominada Evaluación Robusta de la Indización (ERI) permitiría estimar de una manera sólida la viabilidad de un sistema de indización automática.

PALAVRAS-CHAVE *Indización automática, Evaluación, SISA, Bases de datos, Agricola, WoS, Scopus, ERI, Evaluación Robusta de la Indización.*

ABSTRACT Since some years the generation of digital documents is enormous as well as its massive incorporation to the information systems and both realities seem unstoppable. Likewise, there is no doubt that indexing is one of the fundamental processes executed in documentary units. Although the first investigations in automatic indexing began decades ago this subject continues to raise interest. Since then different proposals and methodologies have been presented. SISA is a multilingual automatic indexing system for scientific articles based on heuristic and statistical principles governed by rules based on these principles. **Objective**. In this described context of constant digital increase, it is sought to know the SISA capabilities in the automatic indexing of articles in relation to how they do it in the Agricola, WOS and SCOPUS databases. **Material and method**. One hundred articles published in different years by the journal Agronomy for sustainable development were randomly selected, the indexing assigned to the articles in the mentioned

databases was located, the documents were indexed with SISA, the different indexing were compared and they were calculated the consistency between Agricola and SISA. **Conclusions.** The indexing capabilities of SISA in relation to the reference databases have been satisfactory, although some adjustments are needed. SISA has produced a mean number of descriptors per document similar to Agricola and Scopus, although the composite descriptors of SISA are smaller than in these two databases. Also, the 21.61% consistency achieved between SISA and Agricola is within the percentages in this type of studies. Finally, the proposal of a comprehensive formula for the evaluation of the automatic indexing called Robust Indexing Evaluation (RIE) would allow a solid estimation of the viability of an automatic indexing system.

KEYWORDS *Automatic indexing, Evaluation, SISA, databases, Agricola, WoS, Scopus, RIE, Robust Indexing Evaluation.*

1. INTRODUCCIÓN

La indización ha sido extensamente estudiada. Algunos trabajos valiosos sobre la teoría y práctica de la indización son Frohmann (1990), Lancaster (1991), Farrow (1991), Fugmann, (1993), Hjorland, B. (1997), Anderson y Perez-Carballo (2001) o Mai (2000), entre otros. La norma ISO 5963-1985 define la indización como "la acción de describir o identificar un documento en relación con su contenido." Esto se puede completar señalando que en ocasiones los conceptos resultantes se normalizan y controlan por medio de un vocabulario controlado, de lo contrario sería un indización en lenguaje natural, y por otro lado, cabe añadir que la indización también se ejecuta –consciente o incoscientemente- sobre las necesidades de información de los usuarios para convertirla mediante lenguaje natural o controlado en una ecuación de búsqueda, de ahí que sea un proceso esencial para el almacenamiento de los documentos y puede serlo en la recuperación de información si su resultado (palabras clave o descriptores) es usado posteriormente en la recuperación. Las primeras propuestas de indización automática se fundamentaron en la Ley de Zipf a las que prosiguieron otras basadas en cálculos estadísticos para conseguir términos de indización como la frecuencia inversa del documento (Sparck Jones, 1972), las aportaciones de Gerard Salton con su modelo de discriminación del término o modelo espacio vectorial (Salton and Yang, 1973; Salton, Wong and Yang, 1975) o entre otros, Deerwester et al. (1988) con su indización semántica latente. Asimismo, prototipos de indización automática se fueron desarrollando en menor o mayor medida en grandes centros de información y documentación, como por ejemplo, en el *STN Internacional* de Karlsrube, de Alemania; en la Biblioteca Nacional de Medicina de los Estados Unidos, en el Centro de Información Aeroespacial de la NASA; en la Biblioteca Nacional de Agricultura también de los Estados Unidos; o más recientemente en el *Data Archive* del Reino Unido. (Gil-Leiva, 2017, p. 140).

Como se ha señalado, desde el inicio de estas investigaciones sobre la automatización de la indización a finales de la década de 1950 se han realizado numerosas y variadas propuestas para acometer el proceso intelectual que supone la indización. La terminología utilizada en la literatura para referirse al proceso de la automatización de la indización es variada, pudiendo encontrar estas denominaciones, entre otras: «Automated assisted indexing», «Automated indexing», «Automated support to indexing», «Automatic support to indexing», «Computer aided indexing», «Computer assistance in indexing», si bien la más utilizada es «Automatic indexing». La definición de la automatización de la indización se debe acometer desde una triple perspectiva: a) Programas informáticos que asisten en el proceso de almacenamiento de los términos de indización, una vez obtenidos de modo intelectual (Indización

Asistida por Ordenador Durante el Almacenamiento); b) Sistemas que analizan los documentos de modo automático, pero los términos de indización propuestos los valida y edita -si es necesario- un profesional (Indización Semiautomática); y c) Programas sin ningún tipo de validación, es decir, los términos propuestos se almacenan directamente como descriptores de dicho documento (Indización Automática), (Gil Leiva, 2008).

Desde hace unos años la generación de documentos digitales es enorme así como su incorporación masiva a determinadas unidades documentales como por ejemplo, los libros electrónicos a las bibliotecas académicas, o los artículos científicos a las bases de datos (ver Tabla 1 como ejemplo ilustrativo). De ahí que más frecuentemente se recurra a sistemas de indización automática o semiautomática que ayuden a ejecutar esta tarea. En este contexto de crecimiento digital imparable y la necesidad de disponer de sistemas automáticos o semiautomáticos para llevar a cabo procesos técnicos documentales o que asistan en ellos, se enclava este trabajo. Concretamente se persigue conocer las capacidades de SISA en la indización automática de artículos en relación a cómo lo hacen en las bases de datos Agricola, Web Of Science (en adelante WoS) y SCOPUS, bases de datos de referencia y prestigio de ámbito internacional. Así pues, para cada artículo indizado por SISA se dispondrá de indizaciones previas producidas por estas unidades, lo que nos permitirá responder a cuestiones como ¿cuál es el número medio de términos asignados a cada documento en cada sistema de información? ¿Cuáles son las características de los términos? o ¿Cuál es la semajanza de la indización de SISA con respecto a la indización en Agricola?, entre otros aspectos.

2. MATERIAL Y MÉTODO

2.1 MATERIAL

SISA es un sistema de indización automática que ha sido desarrollado en JAVA, maneja diferentes librerías para extraer la información de los documentos en PDF, txt ó XML y también puede emplear un vocabulario controlado en formato txt o SKOS. SISA está diseñado para la indización de artículos de revista e implementado en plataforma web. Procesa documentos en español, portugués e inglés usando para ello, listas de palabras vacías (artículos, preposiciones, etc.) y vocabularios controlados en estos idiomas. Hace uso del stemming para contabilizar la aparición de raíces y no contar como diferentes los términos *economía*, *económico*, *economías* o *económicamente*, por ejemplo. Para la asignación de los descriptores usa un conjunto de reglas fundamentas en métodos heurísticos (posicionamiento) y estadísticos (frecuencia). En el Anexo 1 se muestran algunas de estas reglas. Las tareas sucesivas para la indización de un artículo con SISA son las siguientes: etiquetar los artículos, procesarlos (aplicar el stemming, calcular la frecuencia de aparición de los términos en los documentos y en la colección, calcular el TFIDF y registrar el lugar en el que aparecen palabras y frases y seguidamente se indizan de acuerdo a las reglas establecidas. Interrelacionado con las utilidades de recuperación disponibles, en la actualidad se está incorporando a SISA el módulo de evaluación mediante la recuperación para hallar índices de exhaustividad, precisión y f-measure. Por otro lado, se ha usado Cascading Style Sheets para el diseño de la aplicación y MySQL como base de datos para guardar las fuentes, los documentos y los resultados de la indización. Finalmente, señalar que SISA está instalado en un servidor Proliant ML310E con 32GB RAM y con un sistema operativo CentOS 7.0. (Gil Leiva, 2008; Rocha Souza y Gil-Leiva, 2016 y Gil-Leiva, 2017).

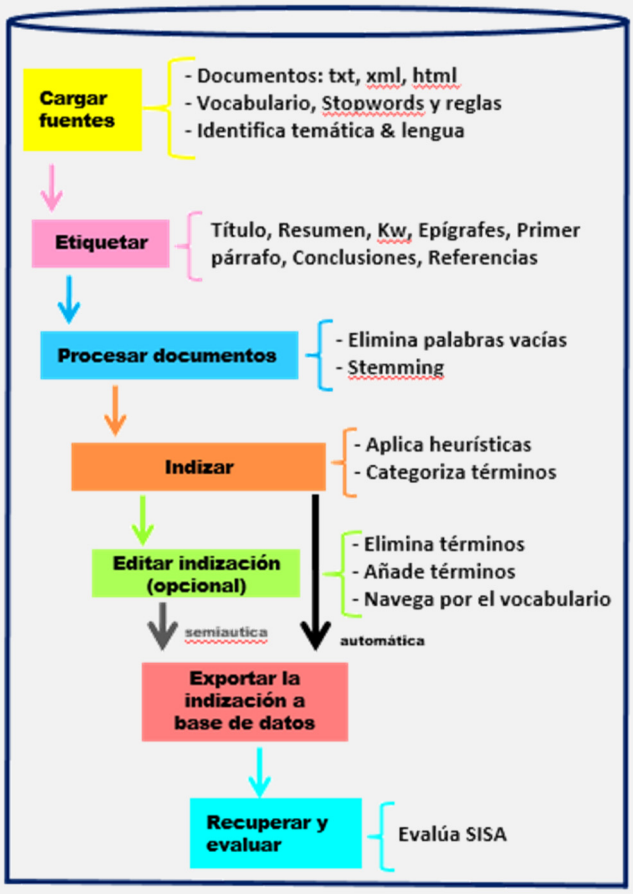


Figura 1. Esquema de los procesos en SISA.

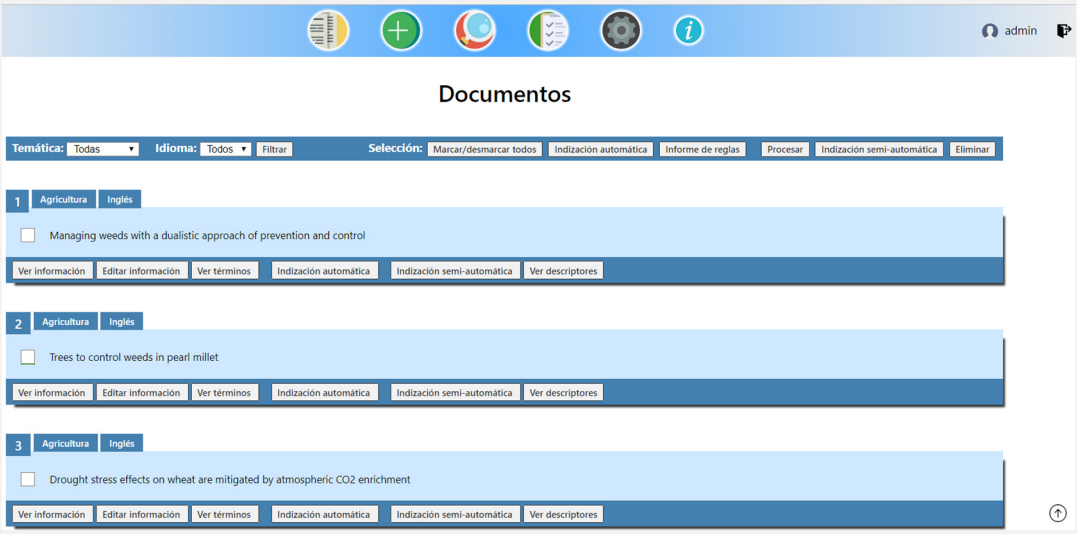


Figura 2. Pantalla principal de SISA.

En la siguiente figura aparecen las utilidades de los iconos que aparecen en cada momento en la parte superior de la interfaz de SISA.

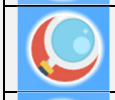
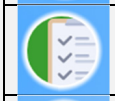
	Retorno a la página principal
	Carga documentos
	Búsqueda en la base de datos una vez indizados los documentos
	Evaluación del sistema
	Configuración (sistema, reglas y usuarios)

Figura 3. Utilidades de los iconos de SISA

Agricola, la base de datos de la Biblioteca Nacional de Agricultura de los Estados Unidos contiene más de 5,2 millones de registros entre artículos, capítulos de libro, tesis, etc. Scopus más de 67 millones de artículos con cerca de 8 millones de "conference proceedings", además de libros, revisiones o patentes. Y WoS supera los 100 millones de artículos procedentes de treinta y tres mil revistas, y cuenta además con más de 7,4 millones de "conference proceedings".

Tabla 1. Artículos incorporados a las bases de datos por años

	2013	2014	2015	2016
Agricola*	128.804	185.006	332.680	430.126
Scopus**	1.831.731	1.915.250	1.922.662	1.877.243
WoS**	1.396.330	1.435.393	1.478.959	1.511.814

* Proporcionados por la responsable de indización de la base de datos
** Obtenidos por medio de sendas búsquedas en estas bases de datos

Como se observa en la Tabla 1 solo el número de artículos incorporados cada año a estas bases de datos es enorme. En Scopus la media de registros añadidos anualmente en el período 2013-2016 ha sido de 2.800.000 documentos y de 2.400.000 en WoS. Los artículos suponen aproximadamente el 65% del total de registros incorporados cada año en Scopus, mientras que en WoS representa el 60%. Los documentos procesados cada año en estas bases de datos hace casi inviable una indización manual de los mismos. Tanto Scopus como WoS intentan en un plazo máximo de cuatro o cinco semanas convertir los artículos publicados en registros accesibles de sus bases de datos obtenidos previamente en su mayoría vía electrónica en formatos XML, PDF o descargados de los sitios webs de las propias revistas. Desde 2012 la Biblioteca Nacional de Agricultura ha automatizado la indización en Agricola por medio de un software de indización semiautomática, denominado Luxid de la empresa Temis, de ahí el fuerte incremento en el número de artículos añadidos desde 2014 en adelante.

2.2 METODOLOGÍA

Para llevar a cabo este experimento se ha manejado SISA para obtener la indización automática de un corpus de cien artículos de Agricultura en inglés publicados en la revista *Agronomy for sustainable development* entre 2007 y 2016. SISA ha usado un vocabulario controlado compuesto por casi 128.000 términos, de los cuales más de 67.000 son descriptores y 60.000 son no descriptores. Estos términos

han sido extraídos del Tesauro de la Biblioteca Nacional de Agricultura disponible en la web de la institución. Asimismo SISA ha empledoa un fichero de palabras vacías en inglés con casi 600 palabras.

Las principales tareas para la realización de este ensayo han sido las siguientes:

- - Extraer los descriptores y no descripotores del Tesauro de la Biblioteca Nacional de Medicina, edición 2017 y adaptarlo al formato de SISA.
- - Seleccionar al azar cien artículos de la revista *Agronomy for sustainable development* pubicados en los años 2007, 2008, 2009, 2014 y 2015.
- - Localizar la indización asignada a los cien artículos en Agricola, WoS y SCOPUS
- - Indizar los 100 artículos con SISA
- - Analizar y comparar las tres indizaciones
- - Hallar los índices de consistencia entre la indización de Agricola y la indización de SISA, ya que en los dos sistemas de indización se ha usado el mismo vocabulario controlado.

Para valorar o evaluar la indización o la indización automática se viene recurriendo de manera mayoritaria al cálculo de los índices de exhaustividad y precisión en la recuperación, como hicimos por ejemplo en Rocha Souza y Gil-Leiva (2016) o Gil-Leiva (2017); o bien, al índice de consistencia entre indizaciones, como llevamos a cabo en (Gil-Leiva, 2001; 2002). Aquí de nuevo recurrimos a la consistencia para comparar la indización de Agricola con la de SISA, ya que se ha configurado y nos hemos servido del mismo vocabulario controlado empleado en Agricola. No calculamos los índices de consistencia entre SISA y WoS o Scopus porque no disponemos de los vocabularios controlados de estas bases de datos. Confrontar la indización de dos bases de datos que usan vocabularios controlados diferentes es desaconsejable para un inexperto en la materia, en este caso, en Agricultura, ya que un mismo concepto puede estar representado por descriptores diferentes en cada base de datos. Nos podemos encontrar que SISA proponga a un documento los descriptores *no-tillage* y Scopus *zero tillage*, o SISA *soil temperature* y Scopus *surface temperature*, pero en estos casos parece que estemos ante los mismos conceptos, por tanto, hay que tener un conocimiento profundo del ámbito terminológico para llevar a cabo comparaciones de este tipo.

Hooper (1965) introdujo una fórmula para hallar la consistencia entre dos indizaciones, y posteriormente, Rolling (1981) introdujo una variante de aquella. Estas dos fórmulas han sido extensamente usadas en numeros experimentos desde entonces. Nosotros venimos sirviéndonos de la siguiente variante de la fórmula de Hooper:

$$C_i = \frac{T_{co}}{(A+B)-T_{co}}$$

en donde,

T_{co} = Número de términos comunes en las dos indizaciones
A= Número de términos usados en la indización A
B= Número de términos empleados en la indización B

La apliación de esta fórmula se puede realizar de manera “rígida” o “relajada”. El Anexo 2 proporciona la indización de Agricola y SISA para el documento 83. La coincidencia del descriptor *biodiversity*

sumaría 1, y la coincidencia del término *food*, sumaría en cambio 0,5, ya que solamente existe coincidencia en una parte del descriptor. A esto llamamos una aplicación "relajada" y si se aplicara una comparación "rígida", en este caso de *food*, sumaría 0. En el experimento que nos ocupa, hemos aplicado la fórmula de manera "relajada".

3. RESULTADOS Y DISCUSIÓN

3.1 LA INDIZACIÓN DE AGRICOLA, SCOPUS, WOS Y SISA

Tabla 2. Total de descriptores asignados a los 100 artículos

	nº total de descriptores asignados	nº medio de deescriptores por documento	nº de descriptores simples	%	nº de descriptores compuestos	%
Agricola	1569	15,6	744	47,4	825	52,5
Scopus	1416	14,2	608	42,9	808	57,0
SISA	1446	14,4	891	61,6	555	38,3
WoS	801	8,01	450	56,1	351	43,8

El tiempo medio empleado por SISA para la indización de un documento es de diecisiete segundos aproximadamente, si bien en el preprocesamiento previo (localización del artículo en PDF, conversión a txt, etiquetado del texto que en esta ocasión se ha realizado de manera integral) se consumen de media unos siete minutos aproximadamente.

De acuerdo a la Tabla 2 en Agricola, Scopus y SISA parece existir cierta homegeneidad en el número de descriptores asignados por documento. En WoS el número es sensiblemente menor (Anexo 3), y aunque el corpus de este ensayo es pequeño parece observarse que no suelen asignar como descriptores conceptos ya incluidos en las palabras clave aportadas por los autores de los artículos, por tanto, en WoS parecen realizar una indización a partir de las palabras clave (Anexo 4), de ahí quizás el menor número de descriptores asignados, ya que WoS (al igual que Scopus) disponen de los campos "Palabras clave de autor". En cambio, en Agricola y en Scopus se percibe que las palabras clave de los autores sí que terminan convirtiéndose en descriptores de dichos artículos. Las heurísticas de posicionamiento SISA conceden un peso destacado a las palabras clave de los autores, de ahí que SISA proponga descriptores que también aparecen como palabras clave.

En cuanto a los descriptores parece existir cierta similitud entre Agricola y Scopus porque tienen un número semejante de descritpores simples y compuestos; mientras que parece alinearse SISA con WoS, si bien, SISA presenta claramente el menor número de descriptores compuestos.

En la indización de los cien artículos por parte de Agricola hemos encontrado dos documentos con cuatro y cinco descriptores respectivamente, en cambio con SISA el número menor de descriptores asignados ha sido de siete, concretamente a dos documentos. En cambio, tanto Agricola como SISA a la mayor parte de los documentos se han asignado entre diez y diecinueve descriptores; y curioamente al documento 36 es al que más descriptores han asignados ambos sistemas, Agricola cuarenta y tres y SISA ha asignado cuarenta y cinco. El artículo 36 cuenta con 28 páginas, casi más del doble que la media de páginas de los artículos del corpus, y aunque solamente se trata de un caso coincidente en Agricola y SISA, (Anexo 5), ¿será que el tamaño de los documentos y sus características interviene

directamente en el resultado de la indización automática o semiautomática?, porque en la indización manual, no parece existir una correlación directa entre el tamaño y el número de descriptores asignados Gil-Leiva y Rodríguez Muñoz (1997, p. 162).

3.2 LA CONSISTENCIA EN LA INDIZACIÓN ENTRE AGRICOLA Y SISA

La consistencia media entre Agricola y SISA ha sido del 21,61%. El índice de consistencia más bajo obtenido ha sido en el artículo 83 con el 6,98%, mientras que el más alto se ha dado en el artículo 2 con un 44,44% (Anexo 2). En este tipo de ensayos la consistencia media oscila entre el 20 y 60%, por tanto, aunque en la parte más baja, se encuentra dentro de esta horquilla.

A la hora de la evaluación de la indización, un elemento a considerar es la corrección, que implica la a la vez la ausencia de errores de inclusión (no asignar un descriptor incorrecto) y de omisión (no dejar de asignar un descriptor que corresponda). En la indización de SISA en este ensayo se han detectado tanto errores de omisión como de inclusión. Errores de inclusión como por ejemplo proponer en un mismo documento “grain yield”, “grains” y “yields”; o bien en otro documento “habitat preferences” y “habitats”. De igual modo, se ha detectado un error al adaptar el tesauro de la Biblioteca Nacional de Agricultura al formato de SISA que ha provocado que más de mil quinientos descriptores erróneos aparezcan en el vocabulario controlado configurado y empleado, lo que ha podido interferir en los resultados.

Desde que estamos desarrollando SISA, ya se ha señalado que hemos tenido oportunidad de aplicar formas para valorar o evaluar su indización. En Gil-Leiva (2017, p. 150) tomando como base el oportuno e interesante trabajo de Golub et al. (2016) señalamos la necesidad de trabajar en la búsqueda de fórmulas robustas para evaluar la indización automática. Ya hemos dicho que para la evaluación se recurre a la exhaustividad y precisión en la recuperación y a la consistencia entre indizaciones, pero su aplicación de manera independiente se nos antoja insuficiente, lo que nos lleva a presentar la siguiente manera integral para evaluar la indización automática. Aunque será expuesta con más detalle y espacio en un futuro próximo, cabe decir ahora que se fundamentada en tres elementos: la f-measure (que es la media harmónica que combina la exhaustividad y precisión), la consistencia (aplicación de la fórmula de la consistencia a dos indizaciones) y la valoración de expertos (juicio de un experto sobre una indización). Esta propuesta la hemos denominado Evaluación Robusta de la Indización (ERI) y se formula de la siguiente manera:

$$ERI = (\text{f-measure} + \text{consistencia} + \text{valoración de expertos}) / 3$$

en donde,

f-measure = media entre los índices exhaustividad y precisión
consistencia = índice de consistencia medio
valoración de expertos = índice medio de la valoración de los expertos

Para la aplicación de ERI se precisa un corpus documental (compuesto por n documentos), la indización perfecta o ideal del corpus (gold indexing), un experto indizador en la temática a evaluar y un conjunto de n necesidades de información para calcular la f-measure. Cada elemento que conforma ERI aporta un valor numérico de 0 a 1, por tanto, cuanto más se acerque ERI a 1 mejor funciona u opera el sistema evaluado.

4. CONCLUSIONES Y TRABAJOS FUTUROS

Los resultados logrados nos han permitido aproximarnos a las capacidades de SISA frente a otros sistemas de indización relevantes. A modo de resumen, y a falta de análisis más profundos de los aquí mostrados y próximas mejoras en la herramienta, parece que la indización de SISA puede llegar a ser tan válida como la de Agricola, WoS y Scopus. La configuración de SISA para este ensayo ha producido de media un número de descriptores por documento similares a Agricola y Scopus, si bien quizás habría que ajustar su reglas para intentar alcanzar una media de 8-12 descriptores por documento. Asimismo, el número de descriptores compuestos de SISA es sensiblemente menor que en estas dos bases de datos, aspecto que requiere un estudio detallado. El 21,61% de consistencia conseguido entre SISA y Agricola se encuentra dentro de los porcentajes en este tipo de estudios, si bien en la parte baja de la horquilla que oscila entre el 20 y 60% aproximadamente. Por último, se ha propuesta una fórmula integral para la evaluación de la indización automática denominada ERI, Evaluación Robusta de la Indización que podrá servir para estimar de una manera sólida la viabilidad de un sistema de indización automática.

Esta investigación ha abierto líneas de trabajo para el futuro como la revisión del algoritmo de SISA para su ajuste, un análisis pormenorizado descriptor a descriptor en relación al contenido de los artículos procesados para seguir mejorando nuestra herramienta, así como ampliar este estudio a otras bases de datos internacionales. Y por último, profundizar en la concepción teórica y práctica de ERI, que aquí solamente ha sido apuntada.

REFERENCIAS BIBLIOGRÁFICAS

Anderson, James D. & Perez-Carballo, José. (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(2), 231-54.

Farrow, John F. (1991), A cognitive process model of document indexing. *Journal of Documentation*, 47(2), 149-166.

Frohmann, Bernd. (1990). Rules of indexing: a critique of mentalism in information retrieval theory. *Journal of Documentation*, 46(2), 81-101.

Fugmann, Robert. (1993). *Subject analysis and indexing: Theoretical foundation and practical advice*. Frankfurt/Main: Indeks Verlag.

Gil-Leiva, Isidoro & Rodríguez Muñoz, José Vicente (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. (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. (2002). Consistencia en la indización de documentos entre indizadores noveles. *Anales de Documentación*, 5, 99-111.

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

Gil-Leiva, Isidoro (2017). SISA: Automatic indexing system for scientific articles. Experiments with location heuristics rules versus TF-IDF rules. *Knowledge Organization*, 43(3), 139-162.

Golub, Koraljka, Soergel, Dagobert, Buchanan, George, Tudhope, Douglas, Likke, Marianne and Hiom, Debra. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology,* 67(1), 3-16.

Hjorland, Biger. (1997). *Information seeking and subject representation: An activity-theoretical approach to information science*. Westport, CT: Greenwood Press.

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

ISO 5963:1985 : *Documentation -- Methods for examining documents, determining their subjects, and selecting indexing terms*. Geneva: ISO.

Lancaster, Frederick W. (1991). *Indexing and abstracting in theory and practice*. Champaign: University of Illinois.

Mai, Jens-Erik. (2000). Deconstructing the Indexing Process. *Advances in Librarianship*, 23, 269-298.

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

Souza, Renato Rocha & Gil-Leiva, Isidoro. (2016). Automatic Indexing of Scientific Texts: A Methodological Comparison. In Chaves Guimarães, José Augusto, Oliveira Milani, Suelen & Dodebei, Vera. *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*, Advances in Knowledge Organization, 2016 (pp. 243-250). Würzburg: Ergon Verlag.

ANEXO 1: EJEMPLO DE ALGUNAS REGLAS USADAS POR SISA EN EL ENSAYO.

ID	Título	Resumen	Palabras clave	Epígrafe	Primer párrafo	Conclusiones	Referencias	Frecuencia documento (DF)	TF-IDF	Voc. Controlado
R1	1	1	1	1	1	1	1	-	-	S
R2	1	1	1	1	1	1	-	-	-	S
R3	1	1	1	1	1	-	1	-	-	S
R16	1	1	-	1	1	-	1	-	-	S
R78	1	-	-	-	-	1	-	-	-	S
R79	-	-	1	-	-	-	-	-	-	S
R80	1	-	-	-	-	-	-	-	0.025	S

ANEXO 2 : ARTÍCULOS CON EL ÍNDICE DE CONSISTENCIA MENOR Y MAYOR.

	Descriptores en Agricola	Descriptores en SISA	Consistencia
Artículo 83: Título: Using our agrobiodiversity: plant-based solutions to feed the world	1. agricultural land 2. arable soils 3. biodiversity 4. food availability 5. gene pool 6. people 7. politics 8. population growth 9. risk 10. sustainable agriculture	1. andean crops 2. biodiversity 3. breeding 4. climate change 5. feeds 6. food production 7. foods 8. nutrition 9. planting 10. solutes 11. utilities 12. yields 13. zoning	6,98%
Artículo 2: Título: Trees to control weeds in pearl millet	1. agroforestry 2. canopy 3. crop yield 4. Faidherbia albida 5. millets 6. parasitic plants 7. Pennisetum glaucum 8. plant growth 9. semiarid zones 10. Striga hermonthica 11. weed control 12. Nigeria	1. canopy 2. controllers 3. Faidherbia albida 4. inflorescences 5. millets 6. pearls 7. Pennisetum glaucum 8. plant growth 9. planting 10. Striga hermonthica 11. surveys 12. trees 13. weed control 14. Nigeria	44,44%

ANEXO 3 : EJEMPLOS DE INDIZACIÓN EN WOS VERSUS SCOPUS Y SISA.

Documento	Descriptores WoS	Descriptores Scopus	Descriptores SISA
25	1. EUTROPHICATION 2. GROWTH 3. STRATEGIES	1. concentration (composition) 2. crop yield 3. growing season 4. irrigation 5. leaching 6. life cycle analysis 7. Lycopersicon esculentum 8. Mediterranean environment 9. nitrate 10. sustainability 11. water use efficiency	1. climate 2. concentrates 3. controllers 4. decrease in nitrate leaching 5. drainage 6. environmental impact 7. eutrophication 8. fertigation 9. greenhouses 10. hydroponics 11. nitrification 12. nutrient solutions 13. physics 14. reduction 15. solanum lycopersicum var.lycopersicum 16. solutes 17. tomatoes 18. water-use efficiency 19. yields
41	1. GROWTH 2. YIELD	1. aboveground biomass 2. agronomy 3. alternative agriculture 4. Citrus 5. crop yield 6. environmental impact 7. fertilizer application 8. growth rate 9. industrial waste 10. leaf area index 11. Mediterranean environment 12. nutrient cycling 13. soil fertility 14. soil organic matter 15. Triticum aestivum 16. Triticum turgidum subsp. durum 17. waste disposal 18. wheat	1. crops 2. durum wheat 3. industrialization 4. mineral fertilizers 5. mineralization 6. oranges 7. organic fertilizers 8. organic soil fertility 9. organic soils 10. organisms 11. reduction 12. wastes 13. yields
78	1. FRANCE 2. SYSTEMS	1. agricultural intensification 2. agricultural modeling 3. agricultural practice 4. agronomy 5. alternative agriculture 6. automation 7. biophysics 8. climate change 9. environmental factor 10. farm 11. farming system 12. grass 13. grassland 14. grazing 15. management practice 16. model test 17. prediction 18. remote sensing 19. surface area 20. France	1. administrative management 2. climate change 3. design 4. farming systems 5. grasses 6. grasslands 7. models 8. range management 9. remote sensing

ANEXO 4 : DESCRIPTORES PROPUESTOS EN LAS BASES DE DATOS PARA EL DOCUMENTO 14.

Título: High efficacy of extracts of Cameroon plants against tomato late blight disease.

Agricola	WoS	Scopus	SISA	Palabras clave del autor
1. agar 2. biodegradability 3. biopesticides 4. Chrysopogon zizanioides 5. Cupressus 6. disease control 7. disease severity 8. ecosystems 9. food contamination 10. fungi 11. fungicide resistance 12. fungicides 13. germination 14. greenhouse experimentation 15. greenhouses 16. indigenous species 17. pathogens 18. pests 19. Phytophthora infestans 20. plant diseases and disorders 21. plant protection 22. sporangia 23. tomatoes 24. toxic substances 25. toxicity 26. Cameroon	1. field 2. pathogenicity 3. phytophthora-infestans 4. potato 5. resistance	1. disease control 2. disease severity 3. experimental study 4. fungal disease 5. fungicide 6. fungus 7. germination 8. inhibition 9. plant extract 10. toxic substance 11. Africa 12. Cameroon 13. Sub-Saharan Africa 14. West Africa 15. Cupressus 16. Cupressus benthamii 17. Fungi 18. Lycopersicon esculentum 19. Phytophthora infestans 20. Vetiveria 21. Vetiveria zizanioides	1. biopesticides 2. controllers 3. disease control 4. extraction 5. fungicides 6. greenhouses 7. late blight disease 8. lates 9. pathogenicity 10. phytophthora infestans 11. plant extracts 12. planting 13. sporangial germination 14. tomatoes 15. Cameroon	1. antigungal activity 2. biopesticide 3. disease suppression 4. late blight 5. Phytophthora infestans 6. plant extracts 7. tomato

Anexo 5: Relación entre el nº de descriptores y el nº de documentos con dicho nº de descriptores.

	Agricola	SISA
Con 4 descriptores	1	0
Con 5 descriptores	1	0
6	1	0
7	1	2
8	4	2
9	3	4
10	6	8
11	5	7
12	11	8
13	4	15
14	8	8
15	5	11
16	10	10
17	9	7
18	6	5
19	5	5
20	4	3
21	3	0
22	2	1
23	1	0
24	5	2
25	3	1
26	1	0
43	1	0
45	0	1
Total	100	100